MobilePoser: Real-Time Full-Body Pose Estimation and 3D Human Translation from IMUs in Mobile Consumer Devices

Vasco Xu
University of Chicago
Chicago, USA
vascoxu@uchicago.edu

Chenfeng Gao
Northwestern University
Evanston, USA
chenfenggao2029@u.northwestern.edu

Henry Hoffmann
University of Chicago
Chicago, USA
hankhoffmann@cs.uchicago.edu

Karan Ahuja
Northwestern University
Evanston, USA
kahuja@northwestern.edu

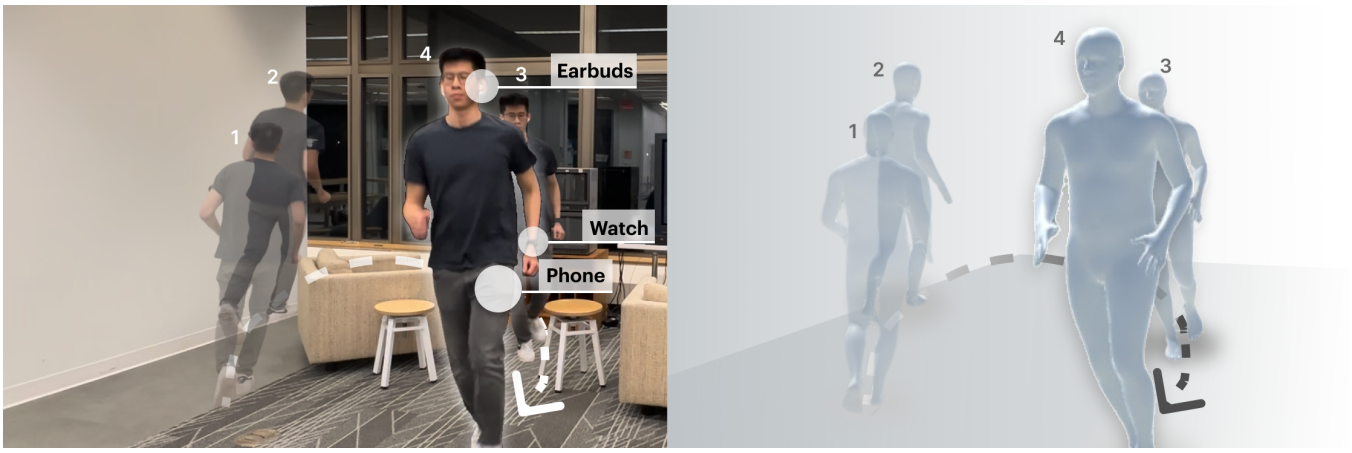

Figure 1: MobilePoser uses any subset of consumer mobile devices (phones, watches, earbuds) available to estimate full-body pose and global translation.

ABSTRACT

There has been a continued trend towards minimizing instrumentation for full-body motion capture, going from specialized rooms and equipment, to arrays of worn sensors and recently sparse inertial pose capture methods. However, as these techniques migrate towards lower-fidelity IMUs on ubiquitous commodity devices, like phones, watches, and earbuds, challenges arise including compromised online performance, temporal consistency, and loss of global translation due to sensor noise and drift. Addressing these challenges, we introduce MobilePoser, a real-time system for full-body pose and global translation estimation using any available subset of IMUs already present in these consumer devices. MobilePoser employs a multi-stage deep neural network for kinematic pose estimation followed by a physics-based motion optimizer, achieving state-of-the-art accuracy while remaining lightweight. We conclude with a series of demonstrative applications to illustrate the unique potential of MobilePoser across a variety of fields, such as health and wellness, gaming, and indoor navigation to name a few.

UIST '24, October 13–16, 2024, Pittsburgh, PA, USA

ACM ISBN 979-8-4007-0628-8/24/10
https://doi.org/10.1145/3654777.3676461

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing**.

KEYWORDS

Motion capture, sensors, inertial measurement units, mobile devices

ACM Reference Format:
Vasco Xu, Chenfeng Gao, Henry Hoffmann, and Karan Ahuja. 2024. MobilePoser: Real-Time Full-Body Pose Estimation and 3D Human Translation from IMUs in Mobile Consumer Devices. In *The 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24), October 13–16, 2024, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3654777.3676461

1 INTRODUCTION

Full-body motion capture has numerous applications in gaming, fitness, and virtual and augmented reality (VR/AR), enabling immersive experiences and context-aware interactions. While vision-based approaches for 3D human pose estimation have shown great

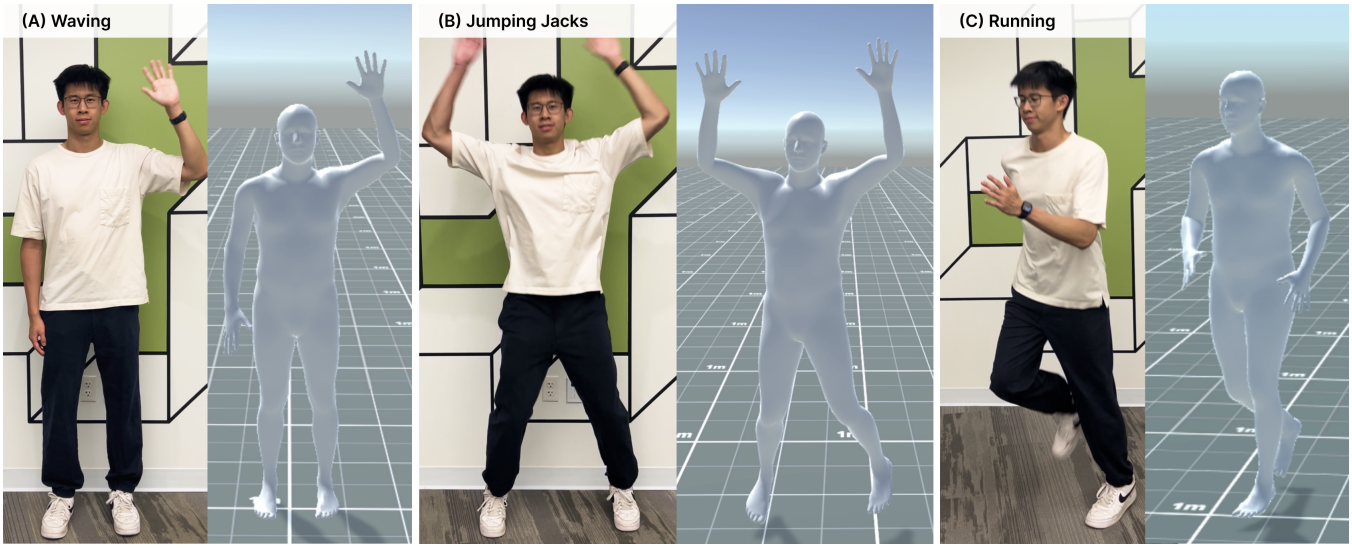

Figure 2: Real-time global pose estimation powered by MobilePoser: (A) Person with smartwatch (left wrist) waving their hands. (B) Person with smartwatch (left wrist) performing jumping jacks. (C) Person wearing a smartwatch (left wrist) and carrying a phone in their right pocket running.

promise, they require subjects to be within the camera's field of view, limiting their practicability for mobile and on-the-go applications. In contrast, inertial measurement unit (IMU) based techniques offer an attractive alternative, enabling less intrusive and occlusion-free user digitization [3].

Commercial systems such as Xsens [45] use up to 17 special-purpose sensors to provide highly accurate pose estimations. However, such approaches are intrusive, making them undesirable for everyday use. Consequently, there has been a trend towards minimizing instrumentation. Sparse inertial pose capture methods, such as TransPose [49] and DIP [14], use 6 IMUs to achieve a balance between accuracy and practicality. Yet, these methods still require expensive and special-purpose IMUs attached to specific body joints. To enable full-body motion tracking without any external infrastructure, IMUPoser [28] leverages IMUs in devices we already carry around with us, namely smartphones, smartwatches, and earbuds. These commodity devices, however, use lower-fidelity IMUs, which compromises online performance, temporal consistency, and global translation estimation.

In this work, we present MobilePoser, a real-time user digitization technique that tracks both *poses* and global *movement* (referred to as *translation*) using consumer devices (Figure 1) such as watches, phones and earbuds. To enable on-the-go motion tracking without any external infrastructure, we must address a set of unique challenges. First, the number of instrumented points is dynamically changing and sparse (at most three devices, with as few as one) [1], making the problem highly under-constrained. Second, IMUs do not directly measure positional data, making global translation tracking non-trivial. Additionally, noise and drift from the low-cost IMUs found in commodity devices complicates pose and translation estimation. Finally, such a system should operate directly on-device for real-time use, anywhere, anytime.

MobilePoser tackles these challenges by employing a multi-stage approach. For pose estimation, it utilizes a deep neural network (DNN) to predict full-body pose from the available IMU data, followed by a physics-based optimization step to ensure spatio-temporal consistency and plausible kinematics. This greatly helps resolve ambiguous instrumented joint motion profiles, such as differentiating between waving (Figure 2 A) versus jumping jacks (Figure 2 B) from only a single smartwatch on the wrist. To aid in generalizability, the model is trained on a large dataset of synthesized IMU measurements generated from high-quality motion capture (MoCap) data. For global translation estimation, MobilePoser employs a hybrid approach that fuses predictions from a foot contact-based method and a DNN-based method that directly regresses the root joint velocity. This combination enables accurate and robust translation estimation, even in challenging scenarios where both feet are in motion together (Figure 2 C). Importantly, MobilePoser is optimized to run on-device, achieving real-time performance of 60 frames per second on a smartphone (iPhone 15 Pro), making it suitable for mobile applications.

In summary, MobilePoser makes the following key contributions:

(1) It presents a novel framework for inertial translation estimation using consumer devices, enabling accurate tracking of global movement without specialized hardware.
(2) It achieves state-of-the-art full-body pose estimation across various on-body configurations of commodity IMU devices, demonstrating robust performance with as few as one and up to three wearable devices.
(3) It provides an open-source implementation that runs in real-time on edge devices, making it accessible and practical for widespread use.

[1]Note, we count the left and right earbuds as a unified single IMU stream

System	# Inst. Joints	FPS	Consumer Device	Translation	MPJVE (cm)	Jitter ($10^2 m/s^3$)
Xsens [45]	17	120	×	✓	-	-
SIP [43]	6	60	×	✓	7.7	3.8
DIP [14]	6	29	×	×	8.9	30.13
TransPose [49]	6	90	×	✓	7.1	1.4
PIP [48]	6	60	×	✓	5.9	0.24
IMUPoser [28]	1–3	25	✓	×	12.1	1.9
MobilePoser (our work)	1–3	60	✓	✓	10.6	0.97

Table 1: Comparison with key prior work on the DIP-IMU dataset.

2 RELATED WORK

2.1 User Digitization with External Sensors

Commercial motion capture systems such as OptiTrack [29] and Vicon [41] use specialized hardware, such as multiple calibrated high-speed infrared cameras, to track retroreflective markers attached to a user's body. Such setups are commonly used in games, movies and character animations that require millimeter accuracy and are the gold standard of motion capture. The expensive infrastructure required by commercial systems, makes them impractical for everyday use. Therefore, much research has been devoted to instrumentation-free approaches using monocular cameras. Such approaches generally rely on RGB [9, 13, 36] or depth [27] cameras based computer vision techniques to predict body pose.

There also exists specialized external hardware for pose tracking in Extended Reality (XR). For example, the HTC Vive [2], PlayStation VR [1] and Oculus Rift [32] track the head, handheld controllers and other limb-borne accessories using external sensor base stations for Virtual Reality (VR) applications. The un-sensed joints are estimated with inverse kinematics [15] or learning-based methods [16, 35]. Other non-optical external approaches for pose estimation include capacitive sensing [50], magnetic fields [31, 33], RF [51], and mechanical linkages [39].

2.2 User Digitization with non-IMU Worn Sensors

Wearable sensors provide a portable and flexible alternative to external sensors. For example, MI-Poser [7] uses magnetic tracking in wristbands and AR glasses to estimate upper-body poses. Other works have explored wrist-worn cameras [20, 44], EMG sensors [24], EIT sensors [22], wrist-worn antennas [19] and depth sensor armbands [10]. However, these works focus solely on capturing the motion of specific body parts (e.g., wrist or upper-body).

To capture full-body motion, a popular approach is to use body-mounted cameras coupled with computer vision techniques [5, 38]. Other works have explored different sensor technologies such as ultrasonic sensors [42] and RFID [18]. Nevertheless, these works require users to wear sensors they do not already have. Pose-On-The-Go [4] addresses this by estimating full-body pose via extreme sensor fusion, leveraging a phone's front and rear cameras, thus requiring no special instrumentation. However, its computationally expensive and relies heavily on heuristics to power body poses, often resulting in unnatural motions. MobilePoser differentiates itself by focusing on full-body pose estimation using power-efficient IMUs already found in consumer devices, such as smartphones, smartwatches, and earbuds.

2.3 User Digitization with IMU Worn Sensors

Commercial motion capture systems, such as Xsens [45], use a large number of inertial sensors (typically 17) strapped to the body to provide high-quality motion capture. These setups consist of homogeneous, high-grade IMUs that are calibrated for noise and have known positions on the body, resulting in a less ill-posed problem compared to using sparse, heterogeneous sensors. However, such an approach is highly inconvenient and intrusive for everyday use.

To address this limitation, researchers have explored reconstructing human motions from a reduced number of sensors. Works such as SIP [43], DIP [14], PIP [48], TIP [17], and TransPose [49] have demonstrated the feasibility of using only 6 commercial-grade Xsens IMU sensors for full-body motion capture. Works have further explored integrating other input modalities (e.g. UWB [8] and egocentric images [47]) in addition to the 6 IMUs for increased performance. All these approaches leverage the homogeneity and known calibrated positions of the sensors to achieve accurate pose estimation. However, even 6 sensors can be cumbersome for on-the-go applications, especially those that require passive sensing.

Recent research has investigated even sparser IMU configurations using commodity devices. IMUPoser [28], which is most closely related to our work, performs pose estimation using any combination of smartphone, smartwatch, and earbuds. While IMUPoser tackles the challenges of heterogeneous sensor quality for pose estimation, it lacks global translation due to IMU noise and drift, and contains unrealistic spatio-temporal motion artifacts. Additionally, IMUPoser runs on a laptop at 25Hz, limiting its practicality for real-time mobile applications.

In contrast, MobilePoser addresses these limitations by demonstrating improved pose estimation accuracy on widely used benchmarks *while also estimating global translation* (see Table 1). Furthermore, our system is designed to run fully on-device, achieving real-time performance of 60 fps on edge mobile devices. This enables MobilePoser to provide a more practical and accessible solution for on-the-go motion capture using commodity devices.

3 MOBILEPOSER

Estimating a user's full-body pose from a sparse set of IMU observations is a severely under-constrained problem as it aims to infer a high-dimensional quantity, i.e., the full-body pose, from low-dimensional observations that only capture partial motion at each instrumented point. Moreover, multiple possible solutions could

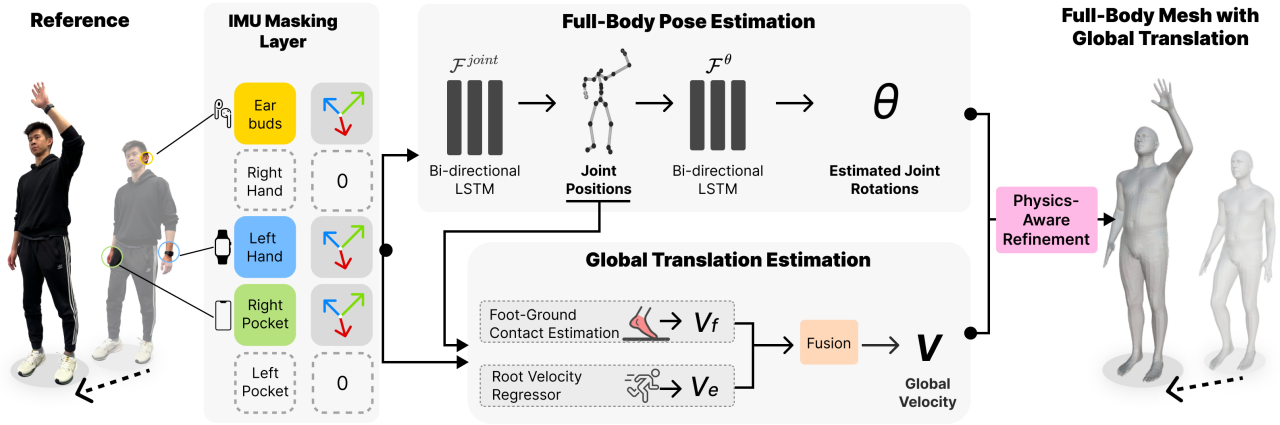

Figure 3: MobilePoser system overview. MobilePoser accepts any available subset of IMU data from the user and masks absent devices by setting their values to zero. The IMU data is then fed into two main modules: (1) Pose Estimation, which first estimates joint positions followed by joint rotations, and (2) Translation Estimation, which combines foot-ground contact probabilities with a direct neural network-based approach to regress global velocity. Finally, a Physics Optimizer refines the predicted joint rotations and global translation to ensure they satisfy physical constraints.

explain the observed data, making it challenging to determine the correct pose. To tackle these challenges, we introduce MobilePoser, a system that leverages data-driven learning and physics-based optimization to estimate accurate and plausible full-body poses and global translations from sparse IMU inputs. Figure 3 provides an overview of our pipeline, which we describe in detail in the following sections.

3.1 System Input

MobilePoser takes as input acceleration and orientation readings from IMUs across any subset of three consumer devices: smartphones, smartwatches, and earbuds. Each of these devices can be placed at different body locations, resulting in various possible combinations. For instance, a smartphone can be stored in the left or right pocket, held in the left or right hand, placed next to the head during a call, or not carried by the user at all. Similarly, smartwatches can be worn on either wrist or not worn at all, while earbuds can be worn, placed in a charging case stored in either pocket, or not carried by the user.

Following IMUPoser [28], we consider 24 plausible device-location combinations across five body locations: right pocket, left pocket, right wrist, left wrist, and head. These combinations cover the various ways users might carry or wear their devices throughout the day. Regardless of the input device combination, our model expects IMU data from the five predefined body locations.

The IMU signal at each location consists of acceleration (3 values) and orientation (a 3×3 rotation matrix), resulting in a total of 12 IMU values per location. Across all five locations, this yields an input vector $\boldsymbol{x} \in \mathbb{R}^{60}$. However, since at any given time only a subset of 1–3 devices may be present, data from absent devices is masked and set to zero. This masking approach allows us to build a unified model that can handle the varying number of available devices and their changing on-body location seamlessly. This further eliminates the need for training separate models for each possible combination, making the system more practical and efficient.

3.2 Full-Body Pose Estimation

To learn a mapping from IMU input to full-body pose, we employ a data-driven, multi-stage neural network approach. Specifically, our pose estimation network consists of two submodules: Joint predictor ($\mathcal{F}^{joint}$) and Rotation predictor ($\mathcal{F}^{\theta}$). More specifically, $\mathcal{F}^{joint}$ estimates joint positions as an intermediate task and $\mathcal{F}^{\theta}$ solves for the joint angle orientations. Both submodules use a bidirectional LSTM (bi-LSTM), to model both spatial and temporal information [14]. We input data into both submodules in a sliding-window fashion with window length N.

3.2.1 Joint Pose Estimation ($\mathcal{F}^{joint}$). This module estimates the joint positions from a sequence of IMU measurements. We explicitly estimate joint positions as an intermediate step, as it helps extract useful information from linear accelerations due to its linear correlation with joint positions [49]. The input to $\mathcal{F}^{joint}$ is $\boldsymbol{x}^{imu}(t) = [\boldsymbol{x}_{t-N}, \ldots, \boldsymbol{x}_t]$, where t is the current time step and N is the time window length. The output are the root (pelvis) relative 3D positions of the 24 SMPL body joints [25] $\boldsymbol{p}(t) = [\boldsymbol{p}_{t-N}, \ldots, \boldsymbol{p}_t] \in \mathbb{R}^{N\times 72}$. The loss function used to train this network is:

$$\mathcal{L}_{joint} = \|\mathbf{p} - \mathbf{p}_{GT}\|_2^2 \tag{1}$$

where the subscript GT denotes the ground truth and p represents the full-body SMPL joint positions.

3.2.2 Joint Rotation and Body Mesh Estimation ($\mathcal{F}^{\theta}$). Here we employ a neural kinematic estimator to regress joint rotations from the previously estimated positions. We concatenate the joint coordinates from $\mathcal{F}^{joint}$ with IMU measurements, which serves as the input to $\mathcal{F}^{\theta}$. Note, while the SMPL body encodes 24 joints, only 18 are relevant from a rotation prediction perspective as the fingers, wrist and toes are independent of the on-body IMUs and

are hence set to identity rotation matrices [49]. The outputs of the network are the 18 root relative joint orientations represented as 6D rotations: $\boldsymbol{\theta}(t) = [\boldsymbol{\theta}_{t-N}, \ldots, \boldsymbol{\theta}_t] \in \mathbb{R}^{N\times 108}$.

Our joint rotation loss consists of three terms: $\mathcal{L}_{ori}, \mathcal{L}_{pos}, \mathcal{L}_{jerk}$. The loss term $\mathcal{L}_{ori}$ is a standard L2 loss from the ground truth joint rotations. The term $\mathcal{L}_{pos}$ penalizes error accumulating along the kinematic chain. Finally, $\mathcal{L}_{jerk}$ promotes temporally smooth predictions, where $jerk(\theta) = \theta_{t-3} + 3\theta_{t-2} - 3\theta_{t-1} + \theta_t$ is a function that computes the jerk of a signal θ at time step t, penalizing the deviation between neighboring frames [49].

Our combined joint rotation loss function can be represented as,

$$\mathcal{L}_\theta = \mathcal{L}_{ori} + \mathcal{L}_{pos} + \lambda\mathcal{L}_{jerk} \tag{2}$$

$$\mathcal{L}_{ori} = \|\theta - \theta_{GT}\|_2^2 \tag{3}$$

$$\mathcal{L}_{pos} = \|\mathrm{FK}(\theta) - \mathbf{p}_{GT}\|_2^2 \tag{4}$$

$$\mathcal{L}_{jerk} = \sum_{t}^{T} jerk\,(\theta) \tag{5}$$

where $FK(\cdot)$ is the forward kinematics function, that computes joint coordinates from joint rotations. Given the joint rotations, the parametric SMPL body model generates a corresponding body mesh with 6890 vertices.

3.3 Global Translation Estimation

Translation estimation from IMUs is challenging as they lack direct distance measurements. Moreover, IMUs are prone to noise and biases, which causes techniques such as double-integration of acceleration to rapidly accumulate errors [46]. Therefore, inspired by prior work [23, 48, 49], we estimate per-frame velocity of the root joint using two submodules: a foot-ground contact (v_f) and a neural network based root velocity estimator (v_e). We fuse the output of the two submodules to obtain a final estimate of global translation.

3.3.1 Foot-Ground Contact based Root Velocity (v_f). Here we estimate the probability of each foot contacting the ground independently using a bi-LSTM network. The input to the model is the concatenated vector of joint positions and IMU measurements. The output of the network is the likelihood that each foot is contacting the ground, denoted as $\boldsymbol{c}_{foot} = \left[\boldsymbol{c}_{lfoot}, \boldsymbol{c}_{rfoot}\right] \in \mathbb{R}^2$. The foot with the higher foot-ground contact probability is defined as the supporting foot, $s = \max\{c_{\text{lfoot}}, c_{\text{rfoot}}\}$. The root velocity, $v_f(t) \in \mathbb{R}^3$, is then computed as the coordinate difference of the supporting foot between consecutive frames. This approach helps capture natural body motions, as movement is significantly influenced by the supporting foot's dynamics [37]. For example, when walking, the body's movement is propelled forward and stabilized by the foot contacting the ground. The network is trained using binary cross-entropy loss.

3.3.2 Neural Network based Root Velocity (v_e). While the supporting foot contact based method yields plausible human movement, it inherently fails when both feet are not contacting the ground (e.g., when running or jumping). To accommodate such cases, we estimate per-frame root velocity directly using a neural network. We again use the predicted joint coordinates and IMU measurements as input. Compared to previous submodules that use a bi-LSTM for prediction, this module uses a unidirectional LSTM due to its capacity to capture longer historical context. The output is per-frame root velocity, denoted as $v_e(t) \in \mathbb{R}^3$. The network is trained using a cumulative L2 loss [49].

3.3.3 Module Fusion. Both modules offer different trade-offs in terms of predicting translation. Supporting foot provides more realistic estimates by leveraging human kinematics but fails when both feet are off the ground. On the other hand, directly estimating root velocity is more general but is highly prone to unnatural movements such as foot sliding [52]. To achieve the benefits of both, we adopt the heuristic-based fusion approach, inspired by TransPose [49]. In summary, when the foot contact c is higher than an upper-threshold $\overline{q}$, we are confident of ground contact by a foot and hence we rely on (v_f) for translation estimation. When the foot contact is below a lower-threshold, $\underline{q}$, we rely on (v_e). For intermediate probabilities, we fuse both velocity estimations using a weighted sum, to output the final global velocity estimate v:

$$v = \frac{q - \overline{q}}{\underline{q} - \overline{q}} v_e + \frac{q - \underline{q}}{\overline{q} - \underline{q}} v_f \tag{6}$$

Following previous work [49], we use $\underline{q} = 0.5$ and $\overline{q} = 0.9$.

3.4 Physics-Aware Refinement

Our pose and translation estimation networks output the user's global pose based on a history of IMU measurements. When trained on sufficiently large amounts of data, the full-body pose estimation and global translation estimation neural networks learn the human motion manifold and produce realistic poses. However, despite the best modeling efforts, the outputs may still contain inter-mesh penetration, temporal artifacts such as jitter, foot-floor penetration and foot skating. To address these issues, we add an off-the-shelf physics motion optimizer [48]. The physics optimizer uses two proportional derivative (PD) controllers to compute the desired acceleration of the simulated character that best reproduces the estimated pose while satisfying physical constraints, such as the equation of motion [12]. The inputs to the physics optimizer are the estimated joint angles θ, the foot-ground contact probabilities c_{foot}, and the neural network based root velocity v_e. The outputs are the optimized joint angles and global translation with reduced jitter and foot-ground penetration (Figure 4). For a detailed overview of the physics optimizer, we refer readers to PIP [48].

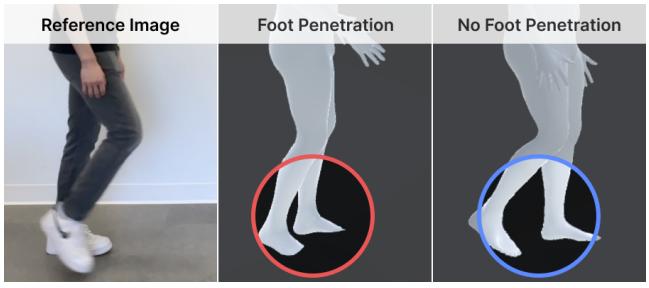

Figure 4: Demonstration of the physics optimizer's ability to reduce foot-ground penetration.

3.5 Real-time Inference

We implement proof-of-concept applications in iOS, using an Apple iPhone 15 Pro, Apple Watch Series 9 and Apple AirPods Pro. The iPhone, Apple Watch and AirPods sample IMU data at 60, 60 and 25 Hz respectively. For uniformity, we convert all the IMU data to 60 Hz by upsampling the AirPods.

We employ the active device selection strategy proposed by IMUPoser [28], wherein the UWB and inertial data is used to track the active devices and their on-body locations. For initial prototyping, the Apple Watch and AirPods communicate over Bluetooth to the iPhone, which streams data to a MacBook Air 2022 via socket. Post connection, a small calibration step is performed to align the IMU measurements with the training data, similar to prior work [14, 28, 49]. Following the setup, data is streamed to the laptop for pre-processing, inference and then relayed to Unity applications for visualization.

To further prototype an on-device edge model, we convert our trained PyTorch model into CoreML with mixed precision quantization and evaluate its performance. On an iPhone 15 Pro, our model incurs ~14ms model inference time running at 60 Hz, capped by input IMU sampling rate.

4 DATA SYNTHESIS AND MODEL TRAINING

Model training requires a large collection of synchronized IMU measurements and corresponding SMPL body poses. We leverage the AMASS [26] MoCap dataset, which provides an extensive collection of such data(~40 hours), including translation.

4.1 Full-Body Pose Estimation

Our models expect IMU measurements as input. We synthesize IMU data following the approach proposed in DIP [14]. In summary, we place *virtual* sensors on the corresponding SMPL mesh vertices (left and right wrists, left and right pockets, and the head) and obtain joint rotations via limb orientations, while acceleration values are computed using finite differences. During training, we scale down the acceleration by a factor of 30 m/s^2, such that its values are on a similar scale to orientations, for better learning. Of note, we do not normalize our IMU measurements to a root joint (e.g., the pelvis), as the number of available devices can vary.

4.2 Global Translation Estimation

The translation estimation networks require (1) binary labels for foot-ground contact states and (2) per-frame root velocity values. To generate foot-ground contact states, we assume that a foot in contact with the ground displays very little movement between frames. Therefore, when the movement of one foot between consecutive frames is less than a threshold u, then we consider it to be contacting the ground. We set $u = 0.008$, following previous work [49]. To train v_e, we require per-frame root velocities. Since the AMASS dataset provides root position data, we can compute root velocities as the coordinate difference of the root position between consecutive frames.

4.3 Training Setup and Procedure

We train our models on a NVIDIA A40 GPU, which takes roughly a day for all modules and device-combinations. In total, our model has ~6.7M trainable parameters. Each module is trained separately using a batch size of 256 and the Adam optimizer [21] with a learning rate of lr = 10^{-3} for 80 epochs. We also apply a gradient clipping with norm of 1, to prevent the gradients from exploding.

During training of $\mathcal{F}^{\theta}$, v_e, and v_f, we add Gaussian noise with $\sigma = 0.04$ to the joint positions to prevent overfitting and deal with prediction errors from $\mathcal{F}^{joint}$. We empirically set $\lambda = 10^{-5}$ when training $\mathcal{F}^{\theta}$, to encourage temporally smooth predictions.

5 EVALUATION

We systematically isolate and analyze the efficacy of MobilePoser across different datasets, evaluation metrics and protocols. We show both qualitative and quantitative results, and also run ablation studies to evaluate our translation estimation design choices.

5.1 Datasets

We evaluate MobilePoser on three real-world, inertial datasets, summarized in Table 2:

- *DIP-IMU [14]* contains data from 10 participants, collected using commercial-grade Xsens [45] IMUs at 60 Hz. It includes a rich variety of activities such as arm raises, stretches, lunges, squats, and punches. However, DIP-IMU does not contain global translation data.
- *TotalCapture [40]* provides real IMU measurements with ground-truth pose and translation, captured using commercial Xsens IMUs at 60 Hz. Following PIP [48], we re-calibrate the acceleration measurements to account for constant bias.
- *IMUPoser [28]* is collected from 10 participants using consumer-grade devices: an iPhone 11 Pro, Apple Watch Series 6, and AirPods, at 25 Hz. It provides ground-truth pose and global translation data.

5.2 Full-Body Pose Estimation

5.2.1 Evaluation Metrics. Like prior work, we use the following evaluation metrics for pose estimation (lower is better for all):

- *Mean Per Joint Rotation Error (MPJRE):* Measure of mean angular error across all root aligned joints in degrees (°).
- *Mean Per Joint Position Error (MPJPE):* Measure of mean Euclidean distance error across all root aligned joints in centimeters (cm).
- *Mean Per Joint Vertex Error (MPJVE):* Measure of mean Euclidean distance error across all root aligned vertices of the SMPL body mesh in centimeters (cm).
- *Mean Per Joint Jitter (Jitter):* Measure of mean jerk across all body joints of the predicted motion in m/s^3.

We use MPJVE as our primary metric of evaluation for ease of comparison with prior work [28].

Dataset	Capture Device	Translation	Data FPS
DIP-IMU	Commercial	×	60 Hz
TotalCapture	Commercial	✓	60 Hz
IMUPoser	Consumer	✓	25 Hz

Table 2: Real-world IMU datasets for MobilePoser Evaluation.

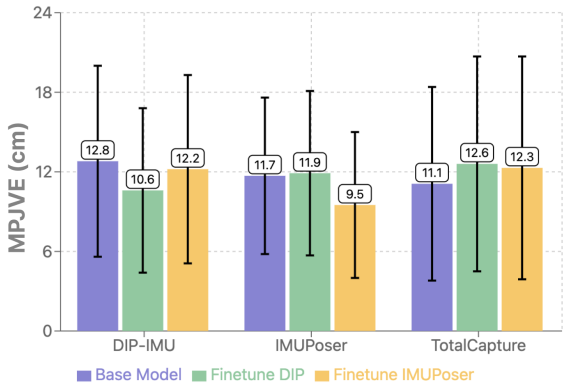

Figure 5: Comparison of MobilePoser's Full-Body Pose Estimation Error across different Evaluation Protocols on the DIP-IMU, IMUPoser and TotalCapture dataset respectively.

5.2.2 Evaluation Protocol. We outline three evaluation protocols for training and fine-tuning to evaluate MobilePoser's efficacy across different data sources and noise profiles.

- **Base Model**: We train our model on the synthetic data generated on the AMASS dataset.
- **Finetune DIP-IMU**: Like prior work, we train on AMASS and then fine-tune on 8 DIP-IMU participants. The 2 holdout participants are used for testing the Finetune DIP-IMU model on the DIP-IMU dataset.
- **Finetune IMUPoser**: We train on AMASS and fine-tune on the first 8 IMUPoser participants. The 2 holdout participants are used for testing the Finetune IMUPoser model on the IMUPoser dataset.

5.2.3 Accuracy across Datasets. Figure 5 shows our full-body pose estimation accuracy for all three protocols across the three datasets listed in Section 5.1. Averaged across all three datasets, the MPJVE for the Base Model, Finetune DIP-IMU and Finetune IMUPoser protocols are 11.89, 11.73 and 11.33 cm respectively. It is interesting to note that the addition of commercial-grade IMU data (Finetune DIP-IMU) only improves accuracy by 1.3% over the base model, while the addition of noisy IMU data from consumer devices (Finetune IMUPoser) results in a bigger improvement of 4.7%.

5.2.4 Accuracy across Activities. We further analyze results on different activities on the IMUPoser dataset, as it provides activity label meta-data. MobilePoser's accuracy generalizes across most everyday activity contexts: the error (MPJVE) for locomotion is 8.2 cm (walking 7.6 cm, jogging 8.8 cm), exercises is 10 cm (kicking: 7.5 cm, jumping jacks: 11.1 cm, boxing: 11.5 cm), sitting is 11.5 cm and freestyle motions such as tennis and basketball are 9.1 cm and 11.7 cm respectively. The accuracy degrades for postures with the user lying/facing down, e.g. push-ups have higher error of 16.1 cm.

5.2.5 Comparison with prior work. To aid in direct comparison with prior work [14, 28, 48, 49], we now make use of the Finetune DIP-IMU evaluation protocol, that is training a base model on the synthetic IMU data from AMASS and fine-tuning it on the 8 participants from DIP-IMU dataset. Tables 1 and 3 offer a quantitative

System	# Inst. Joints	MPJRE	MPJVE	Jitter
DIP	6	17.2°	11.2	3.62
TransPose	6	12.8°	7.4	0.95
PIP	6	12.1°	6.5	0.20
IMUPoser	1–3	25.6°	15.4	1.30
MobilePoser	1–3	23.7°	12.6	0.55

Table 3: Comparison with key prior work on the TotalCapture dataset.

comparison against key prior work, evaluated on the DIP-IMU and TotalCapture, dataset respectively. Given that our system targets a very sparse configuration of IMUs (1-3), it is unsurprising that we perform worse than systems utilizing 6 IMUs, strategically placed around the body. On the DIP-IMU and TotalCapture dataset, compared to IMUPoser, which considers the same device-location combinations, we perform significantly better displaying a 12.4% and 18.2% decrease in vertex error respectively.

On the IMUPoser dataset, Figure 7 (A) provides a detailed breakdown of accuracy for different on-body device locations. Averaging across the 1, 2 and 3 device conditions, MobilePoser outperforms IMUPoser by 24.1%, 14.2% and 8.7% respectively. Furthermore, Figure 7 (B) provides an accuracy breakdown for the instrumented and non-instrumented joints in comparison with IMUPoser. If a limb has an IMU placed on any part, we consider all the joints pertaining to it as instrumented joints, while the rest are marked as non-instrumented. MobilePoser is 18.1% and 17.4% better than IMUPoser for predicting instrumented and non-instrumented joints respectively. This can be seen in Figure 6 which depicts a visual comparison of our pose estimation with IMUPoser.

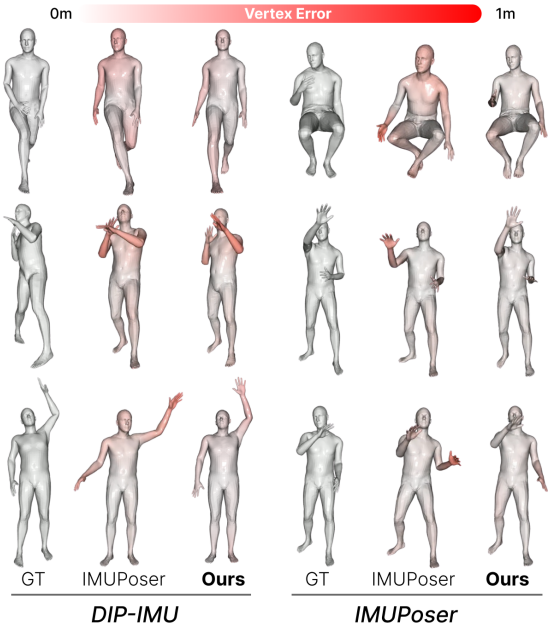

Figure 6: Qualitative comparisons between our method and IMUPoser on the DIP-IMU and IMUPoser dataset.

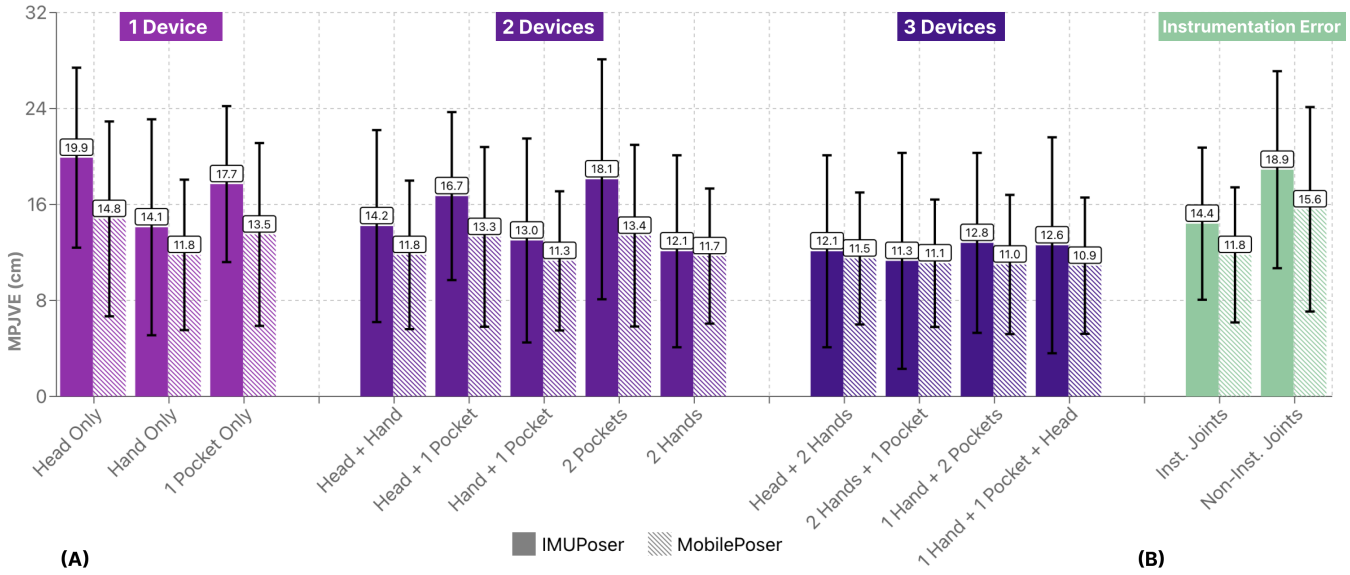

Figure 7: MPJVE comparison between IMUPoser and MobilePoser (our system) on the IMUPoser Dataset for: (A) Different on-body device combinations (B) Instrumented vs Non Instrumented joints.

5.3 Global Translation Estimation

5.3.1 Evaluation Protocol. We evaluate our Global Translation Estimation module on the TotalCapture and IMUPoser datasets, as DIP-IMU lacks translation data. Like prior work [48, 49], we use the Finetune DIP-IMU protocol (Section 5.2.2), that is we train on AMASS and fine-tune on 8 participants of DIP-IMU to track the Root Translation Error (Euclidean norm of the cumulative distance errors within 1 second).

5.3.2 Accuracy across Datasets and Body Regions. On the TotalCapture and IMUPoser dataset, our mean root translation error across all device combinations is 27.55 and 17.63 cm respectively. Interestingly, for both IMUPoser and TotalCapture datasets, we observe only a slight decrease in error when increasing the number of devices from one to two (6.1%) and no significant improvement (4.0%) when increasing from two devices to three. Analysing the error across different body regions for the single device scenario (Figure 8) (A), we see that a device in the pocket has a much lower error (14.8 cm) compared to that on the wrist (25.7 cm) or the head (29.7 cm). This can be attributed to the legs capturing most of the locomotion data during translation, resulting in marginal gains from sensors on the upper-body. Figure 8 (B) shows the the cumulative distance error over time.

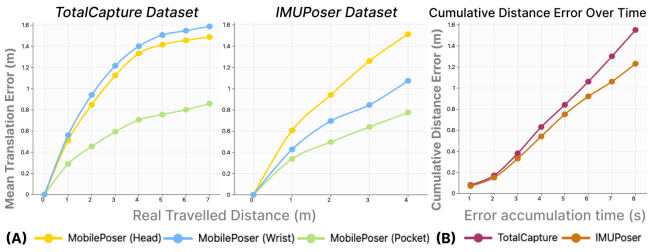

Figure 8: (A) Comparison of cumulative translation error for different instrumented joints on the IMUPoser and TotalCapture dataset. (B) Evaluation of cumulative distance errors with respect to time.

5.3.3 Ablation Study. We perform ablation studies to understand the impact of key components in our system and their effects on performance. At the core of our system lies a subtle yet powerful concept: higher-order digitization (e.g., body pose) improves lower-order digitizations (e.g., steps). To quantify this idea, we run an ablation study of our translation estimation technique using *both* IMU data and the corresponding full-body pose inferred from it versus using only IMU data. Figure 9 summarizes our results. Our IMU-only, direct regression has an error of 21.4 cm across both

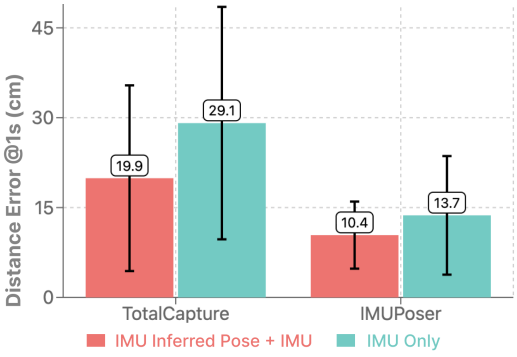

Figure 9: Benefits of using high-order digitization (i.e., IMU inferred poses) for estimating global translation.

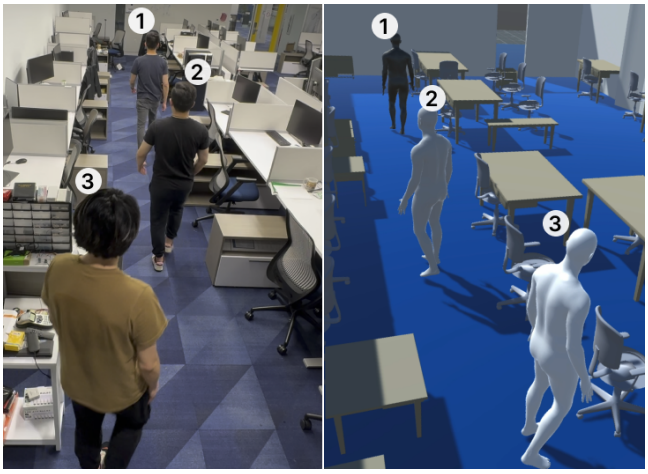

Figure 10: Example indoor navigation application where MobilePoser digitizes multiple users within an office space.

datasets, while our integrated (IMU + IMU inferred pose) approach decreases error by 29.4% to 15.1 cm.

Building on the multi-stage architecture, we further evaluate the impact of two additional components: jerk loss and physics refinement. These elements were designed to enhance motion smoothness and physical plausibility. For the IMUPoser dataset, the jerk loss reduces jitter by 23.9% and translation error by 3.33%, but increases mean pose error by 0.05%. Further, the physics-aware refinement reduces jitter by 29.7% and translation error by 0.4%, but increases the mean pose error by 0.7%. The negligible increase in mean pose error is expected, as it may occasionally over-smooth the motion. This phenomenon is also seen in the PIP [48]. We believe that significant improvements in jitter and translation far outweigh the minimal increase in pose error, resulting in a more realistic motion.

5.3.4 Comparison with prior work. To the best of our knowledge, no other works have explored both *full-body pose* and *translation* from such a sparse set of commodity IMUs. IMUPoser [28], which also targets consumer devices, does not estimate global translation. On the TotalCapture dataset, TransPose (6 IMUs) has a translation error of 12.8 cm while that of MobilePoser is 19.9 cm when a single IMU device is placed in the pocket. Unsurprisingly, a commercial grade, 6 IMU-based system has higher accuracy due to their waist and knee mounted sensors, which capture larger ranges of locomotion compared to devices carried in the pocket.

6 EXAMPLE USES

MobilePoser enables full-body pose estimation with global motion tracking using devices that users already own, opening up a wide range of novel applications. This section showcases three proof-of-concept applications in indoor navigation, gaming, and healthcare to illustrate MobilePoser's unique capabilities and potential impact.

6.1 Indoor Localization and Navigation

To demonstrate MobilePoser's potential in this domain, we scan an office space using the PolyCam [34] LiDAR scanner app with an Apple iPhone 15 Pro. As shown in Figure 10, multiple users walk through the virtual office space, with their interactions and movements seamlessly digitized and represented in real-time. Here, one user has a phone in their pocket and a watch on their wrist, while the other two only have a phone in their pocket. By leveraging the IMUs in these consumer devices, MobilePoser enables accurate indoor navigation and localization without the need for additional infrastructure or specialized hardware. This opens up exciting possibilities for applications such as indoor way finding, context-aware virtual assistants, and immersive virtual tours.

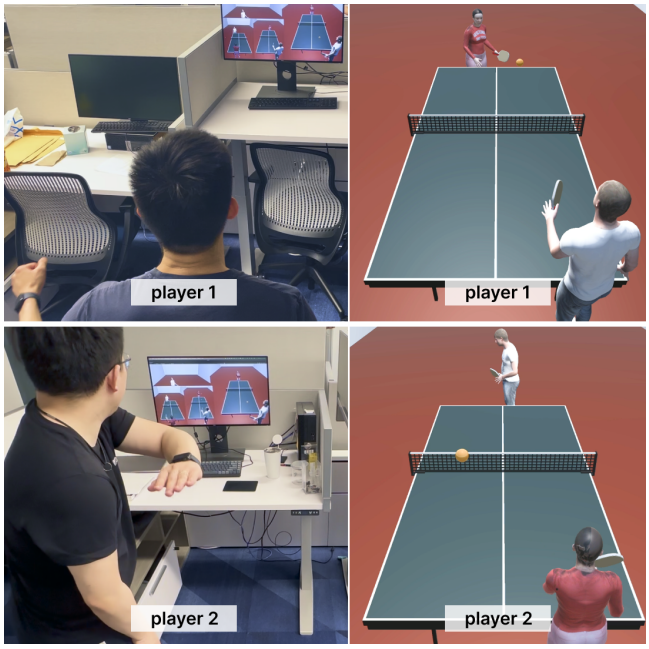

Figure 11: In this table tennis game users can move around the table freely and use their wrist-instrumented hand to control their racket.

6.2 Mobile Gaming Experiences

To showcase this potential, we developed a virtual table tennis game (Figure 11) that allows users to play remotely with others, similar to how Nintendo games are played in front of a TV. Each player has a phone in their pocket and a watch on the dominant (left) hand, which is controlling the racket. Players can freely move within their local space to control their avatars, adding a new level of physical interaction to the gaming experience. MobilePoser's ability to track full-body movements using everyday devices eliminates the need for specialized controllers, making immersive gaming experiences more accessible to a wider audience.

6.3 Fitness and Wellness

MobilePoser has the potential to revolutionize fitness tracking and rehabilitation by providing accurate, real-time feedback on a user's movements and poses without the need for external sensors or camera setups. This enables users to monitor their exercise form, track progress, and receive personalized guidance using the devices

Figure 12: MobilePoser's full-body pose and locomotion can be used to automatically detect and count exercise repetitions, better estimate calories and monitor form.

they already own. In this example (Figure 12), a user performs a workout routine while MobilePoser captures the session using the IMU data from the smartphone in the user's pocket. This not only allows the user to review their performance and track progress over time but also enables remote monitoring by fitness instructors or physical therapists. Moreover, MobilePoser's ability to track full-body movements facilitates interactive rehabilitation regimens [4] and other passive health sensing applications such as gait analysis [30] or hyperactivity detection [6], among others.

7 OPEN SOURCE

To enable other researchers and practitioners to build upon our work, we release our pre-trained models, data pre-processing scripts, and model training code as open-source software at: https://github.com/SPICExLAB/MobilePoser. By making our work fully reproducible and extensible, we hope to accelerate research and development in the field of mobile motion capture using everyday devices.

8 LIMITATIONS AND FUTURE WORK

While MobilePoser demonstrates promising results in estimating full-body pose and translation using minimal instrumentation, there are several limitations and opportunities for future work. First, as a purely inertial-based technique, MobilePoser's translation estimation is still susceptible to drift, particularly when devices deviate from their calibrated positions. This can occur when users wear loose clothing, causing the phone in the pocket to move around and resulting in orientation changes. To address this issue, future work could explore re-calibration techniques based on stationary poses or leverage additional sensory information, such as GPS, UWB or visual odometry, to correct for drift.

Second, akin to prior wor, our evaluation has limitations of being tested on lab collected datasets. All the test datasets (DIP, TotalCapture, IMUPoser) were collected in lab settings due to the need for an accurate external ground truth motion capture system. Although we empirically demonstrate that MobilePoser works in real-world settings (as seen in the accompanying video), we acknowledge the need for future datasets captured in-the-wild.

Another limitation of MobilePoser, much like other prior works [14, 28, 48, 49], is the need for a calibration step. Currently, users first stand in a T-pose, which aligns the IMU data with the training data based on the SMPL kinematic model. While this calibration process is acceptable for some use cases, such as gaming, it may be less desirable for applications that demand seamless interactions, like indoor navigation. Future work could investigate more natural and unobtrusive calibration procedures, such as detecting common poses like standing with arms by the side using UWB, similar to SmartPoser [11].

In conclusion, while MobilePoser presents a significant step forward in enabling full-body pose and translation estimation using everyday devices, there remain several avenues for future research to extend the capabilities of this approach.

9 CONCLUSION

In this paper, we present MobilePoser, a real-time, on-device system for estimating full-body pose and translation using IMUs in consumer mobile devices (phones, watches, earbuds). By leveraging a multi-stage approach that combines data-driven learning and physics-based optimization, MobilePoser achieves state-of-the-art accuracy while remaining lightweight and efficient. Our extensive evaluation on public datasets demonstrates clear improvements over prior work, both in terms of full-body pose estimation accuracy and enabling novel global translation estimation. Furthermore, we showcase the potential of MobilePoser through a series of proof-of-concept applications in gaming, fitness, and indoor navigation, highlighting its ability to enable new and immersive experiences using the devices people already own.

ACKNOWLEDGMENTS

We thank Jianru Ding from the University of Chicago and Zeya Chen from the Institute of Design, Illinois Institute of Technology for helping film the video. Vasco Xu's and Henry Hoffmann's work on this project is supported by NSF (CCF-1823032 and CNS-1956180).

REFERENCES

[1] [n. d.]. PlayStation VR. https://www.playstation.com/en-us/explore/playstation-vr/.
[2] 2023. HTC Vive. https://www.vive.com.
[3] Karan Ahuja. 2024. Practical and Rich User Digitization. arXiv:2403.00153 [cs.HC] https://arxiv.org/abs/2403.00153
[4] Karan Ahuja, Sven Mayer, Mayank Goel, and Chris Harrison. 2021. Pose-on-the-go: Approximating user pose with smartphone sensor fusion and inverse kinematics. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12.
[5] Karan Ahuja, Vivian Shen, Cathy Mengying Fang, Nathan Riopelle, Andy Kong, and Chris Harrison. 2022. Controllerpose: inside-out body capture with VR controller cameras. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–13.
[6] Riku Arakawa, Karan Ahuja, Kristie Mak, Gwendolyn Thompson, Sam Shaaban, Oliver Lindhiem, and Mayank Goel. 2023. LemurDx: Using Unconstrained Passive Sensing for an Objective Measurement of Hyperactivity in Children with no Parent Input. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 2 (2023), 1–23.
[7] Riku Arakawa, Bing Zhou, Gurunandan Krishnan, Mayank Goel, and Shree K Nayar. 2023. MI-Poser: Human Body Pose Tracking Using Magnetic and Inertial Sensor Fusion with Metal Interference Mitigation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (2023), 1–24.

[8] Rayan Armani, Changlin Qian, Jiaxi Jiang, and Christian Holz. 2024. Ultra Inertial Poser: Scalable Motion Capture and Tracking from Sparse Inertial Sensors and Ultra-Wideband Ranging. In *ACM SIGGRAPH 2024 Conference Papers*. 1–11.

[9] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*. Springer, 561–578.

[10] Nathan Devrio and Chris Harrison. 2022. DiscoBand: Multiview Depth-Sensing Smartwatch Strap for Hand, Body and Environment Tracking. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–13.

[11] Nathan DeVrio, Vimal Mollyn, and Chris Harrison. 2023. SmartPoser: Arm Pose Estimation with a Smartphone and Smartwatch Using UWB and IMU Data. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–11.

[12] Roy Featherstone. 2014. *Rigid body dynamics algorithms*. Springer.

[13] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. 2023. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14783–14794.

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

[15] Fan Jiang, Xubo Yang, and Lele Feng. 2016. Real-time full-body motion reconstruction and recognition for off-the-shelf VR devices. In *Proceedings of the 15th ACM SIGGRAPH Conference on Virtual-Reality Continuum and Its Applications in Industry-Volume 1*. 309–318.

[16] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. 2022. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *European Conference on Computer Vision*. Springer, 443–460.

[17] Yifeng Jiang, Yuting Ye, Deepak Gopinath, Jungdam Won, Alexander W Winkler, and C Karen Liu. 2022. Transformer Inertial Poser: Real-time human motion reconstruction from sparse IMUs with simultaneous terrain generation. In *SIGGRAPH Asia 2022 Conference Papers*. 1–9.

[18] Haojian Jin, Zhijian Yang, Swarun Kumar, and Jason I Hong. 2018. Towards wearable everyday body-frame tracking using passive RFIDs. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–23.

[19] Daehwa Kim and Chris Harrison. 2022. Etherpose: Continuous hand pose tracking with wrist-worn antenna impedance characteristic sensing. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–12.

[20] David Kim, Otmar Hilliges, Shahram Izadi, Alex D Butler, Jiawen Chen, Iason Oikonomidis, and Patrick Olivier. 2012. Digits: freehand 3D interactions anywhere using a wrist-worn gloveless sensor. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. 167–176.

[21] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[22] Alexander Kyu, Hongyu Mao, Junyi Zhu, Mayank Goel, and Karan Ahuja. 2024. EITPose: Wearable and Practical Electrical Impedance Tomography for Continuous Hand Pose Estimation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–10.

[23] Jiye Lee and Hanbyul Joo. 2024. Mocap Everyone Everywhere: Lightweight Motion Capture With Smartwatches and a Head-Mounted Camera. *arXiv preprint arXiv:2401.00847* (2024).

[24] Yilin Liu, Shijia Zhang, and Mahanth Gowda. 2021. NeuroPose: 3D hand pose tracking using EMG wearables. In *Proceedings of the Web Conference 2021*. 1471–1482.

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34, 6 (Oct. 2015), 248:1–248:16.

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

[27] Microsoft Corporation. [n. d.]. Microsoft Kinect.

[28] Vimal Mollyn, Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja. 2023. 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*. 1–12.

[29] NaturalPoint, Inc. [n. d.]. OptiTrack. https://www.optitrack.com.

[30] Shu Nishiguchi, Minoru Yamada, Koutatsu Nagai, Shuhei Mori, Yuu Kajiwara, Takuya Sonoda, Kazuya Yoshimura, Hiroyuki Yoshitomi, Hiromu Ito, Kazuya Okamoto, et al. 2012. Reliability and validity of gait analysis by android-based smartphone. *Telemedicine and e-Health* 18, 4 (2012), 292–296.

[31] Northern Digital Inc. 2020. trakSTAR. https://www.ndigital.com/msci/products/drivebay-trakstar.

[32] Mathias Parger, Joerg H Mueller, Dieter Schmalstieg, and Markus Steinberger. 2018. Human upper-body inverse kinematics for increased embodiment in consumer-grade virtual reality. In *Proceedings of the 24th ACM symposium on virtual reality software and technology*. 1–10.

[33] Polhemus. 2020. Polhemus Motion Capture System. https://polhemus.com/.

[34] PolyCam. [n. d.]. PolyCam. https://poly.cam/.

[35] Jose Luis Ponton, Haoran Yun, Andreas Aristidou, Carlos Andujar, and Nuria Pelechano. 2023. SparsePoser: Real-time Full-body Motion Reconstruction from Sparse Data. *ACM Transactions on Graphics* 43, 1 (2023), 1–14.

[36] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. 2021. Tracking people with 3D representations. *arXiv preprint arXiv:2111.07868* (2021).

[37] Nirupam Roy, He Wang, and Romit Roy Choudhury. 2014. I am a smartphone and i can tell my user's walking direction. In *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*. 329–342.

[38] Takaaki Shiratori, Hyun Soo Park, Leonid Sigal, Yaser Sheikh, and Jessica K Hodgins. 2011. Motion capture from body-mounted cameras. In *ACM SIGGRAPH 2011 papers*. 1–10.

[39] Ivan E Sutherland. 1968. A head-mounted three dimensional display. In *Proceedings of the December 9-11, 1968, fall joint computer conference, part I*. 757–764.

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

[41] Vicon Motion Systems Ltd. [n. d.]. Vicon. https://www.vicon.com.

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

[43] Timo Von Marcard, Bodo Rosenhahn, Michael J Black, and Gerard Pons-Moll. 2017. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer graphics forum*, Vol. 36. Wiley Online Library, 349–360.

[44] Erwin Wu, Ye Yuan, Hui-Shyong Yeo, Aaron Quigley, Hideki Koike, and Kris M Kitani. 2020. Back-hand-pose: 3d hand pose estimation for a wrist-worn camera via dorsum deformation network. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 1147–1160.

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

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

[47] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Vladislav Golyanik, Shaohua Pan, Christian Theobalt, and Feng Xu. 2023. EgoLocate: Real-time Motion Capture, Localization, and Mapping with Sparse Body-mounted Sensors. *arXiv preprint arXiv:2305.01599* (2023).

[48] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. 2022. Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13167–13178.

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

[50] Yang Zhang, Chouchang Yang, Scott E Hudson, Chris Harrison, and Alanson Sample. 2018. Wall++ room-scale interactive and context-aware sensing. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–15.

[51] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. 2018. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7356–7365.

[52] Li'an Zhuo, Jian Cao, Qi Wang, Bang Zhang, and Liefeng Bo. 2023. Towards Stable Human Pose Estimation via Cross-View Fusion and Foot Stabilization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 650–659.