Towards Real-World Wearable Motion Reconstruction

Andrea Boscolo Camiletto[1,2], Rishabh Dabral[1,2], Eduardo Alvarado[1], Thabo Beeler[3], Marc Habermann[1,2], and Christian Theobalt[1,2]

[1] Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
[2] Saarbrücken Research Center for Visual Computing, Interaction and AI, Germany
[3] Google, Switzerland

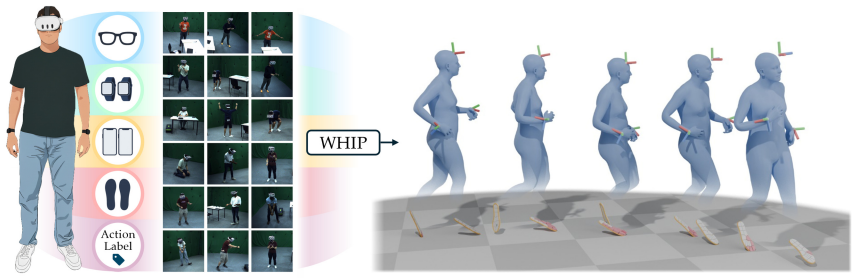


Fig. 1: We present a multimodal wearable sensor suite for motion capture (left) and show full-body motion reconstructed by our model from sparse sensor readings (right).

Abstract. The modern-day surge in popularity of wearable devices poses a fundamentally unique motion capture problem: reconstructing full-body movement from any set of sensing hardware worn at a given moment. Yet, most research efforts assume fixed sensor configurations (e.g., IMU suits or HMD-centric rigs) and cannot generalize across them. In contrast, we argue that motion capture should prioritize unobtrusive and lightweight devices such as smartphones, smartwatches, smart glasses, and smart insoles, and study the interplay between them. To this end, we make three contributions. First, we present a large-scale multimodal dataset synchronizing these consumer-grade sensors with ground-truth 3D motion, spanning 50 diverse activities including everyday tasks, sports, and social interactions. Second, we propose WHIP, a baseline generative model that reconstructs motion from arbitrary subsets of available sensors, robustly handling missing modalities and producing physically plausible motions. Third, we conduct a systematic study of sensor complementarity, quantifying how different modalities complement one another. Code and dataset are available at this URL.

Keywords: Motion capture · Wearable sensors · Multimodal learning · Human pose estimation

1 Introduction

Human Motion Capture (MoCap) is essential across numerous fields, including gaming, sports, medicine, VR/AR, robotics, and filmmaking, enabling the accurate tracking and reproduction of human movements. Traditional MoCap solutions, however, present challenges for real-world deployment: optical marker-based systems [34] achieve excellent precision but require cumbersome suits and controlled environments, whereas markerless vision-based systems require extensive camera infrastructure [5] for high-quality results. This has motivated a shift toward lightweight, accessible sensing: smartphones and smartwatches are now ubiquitous, and smart insoles and head-mounted devices increasingly common. Together they offer complementary cues that no single modality provides, making them a promising basis for untethered motion capture.

However, converting such sparse, commodity signals into reliable full-body tracking remains a challenging and underconstrained problem. Prior work has addressed individual modalities in isolation. IMU-based methods use as few as six sensors [13, 22, 44, 45], yet even these multi-sensor setups require careful calibration and remain impractical for daily use. A parallel line of work leverages VR/AR hardware: headsets and controllers provide native head and hand tracking, often achieving strong reconstruction accuracy [14, 38]. However, inferring leg motion from only upper-body cues is fundamentally ambiguous. Headsets with body-facing cameras partially mitigate this ambiguity by directly observing the body [2, 4, 48], yet occlusions and limited hardware availability still limit their practicality. A complementary direction of increasing interest is plantar pressure sensing: by measuring foot–ground interactions, it provides informative cues on gait, balance, and contact events [1]. This trend is fueled by the growing availability of commercial smart insoles [25, 27–29], enabling practical out-of-lab use. These signals improve motion plausibility when fused with other inputs [39, 46] and can even support standalone capture [40, 41], though datasets remain scarce and many systems still require controlled settings.

In practice, people wear diverse combinations of devices, and wearable motion capture should adapt to whichever subset is available. Existing methods, however, are specialized to fixed hardware configurations. To bridge this gap, we introduce a large-scale dataset that synchronizes everyday consumer wearables (two smartphones, two smartwatches, pressure-sensing insoles, and a VR headset) with high-quality 3D motion from a calibrated 120-camera markerless system [5], time-aligned and spatially registered across modalities. The dataset spans 14 participants (two sessions each), 50 action classes, and over 7 hours ($\approx$ 800k frames) of synchronized recordings, captured using commodity hardware in realistic configurations (see Fig. 1). Leveraging this data, we develop WHIP, a flow-matching generative model that reconstructs full-body motion from arbitrary sensor subsets. WHIP builds on a DiT-style backbone with per-modality cross-attention, conditioned on any available sensors and optional action labels, and is trained with a conditional flow-matching objective to produce temporally consistent motions. Together, dataset and model enable a principled analysis of sensor complementarity: we systematically evaluate single- and multi-

modal subsets, quantify each sensor's contribution, and identify Pareto-optimal combinations under practical hardware constraints.

In summary, our work presents:

- A large-scale dataset combining synchronized everyday wearables.
- WHIP, a flow-matching model that reconstructs motion from any sensors.
- The first cross-modal study that comprehensively compares the complementarity of individual wearable sensors for motion capture.

2 Related Work

Datasets for Wearable Motion Capture. Collecting datasets for wearable motion capture is inherently challenging, requiring precise temporal synchronization and spatial alignment with full-body tracking labels.

IMU-based datasets. Several datasets have paired inertial recordings with high-quality motion capture supervision. Most notably, DIP-IMU [13] and TotalCapture [33] collected data using a full-body Xsens suit, establishing standardized evaluation on a subset of sensors and the widely adopted "six-IMU" configuration that underpins much subsequent IMU-based pose estimation. However, they rely on intrusive full-body suits, limiting applicability in everyday scenarios. IMUPoser [24] reduced this overhead by relying solely on commodity devices—two smartphones, two smartwatches, and a pair of earbuds—but the resulting dataset comprises only two hours of data.

HMD-based datasets. Head-mounted devices (HMDs) are another major source of wearable signals. Ego4D [10] and Ego-Exo4D [11] introduced large-scale egocentric video benchmarks. Nymeria [21] further scaled AR-glasses recordings, combining them with wristband sensors. However, their full-body tracking labels are derived from sparse external cameras, resulting in lower-quality ground truth compared to professional motion capture.

Pressure datasets. Research on plantar pressure and ground reaction forces has primarily targeted biomechanics rather than full-body pose. Large-scale resources pair MoCap with force plates (e.g., GroundLink [12], AddBiomechanics [37]) or pressure mats with video (MMVP [46], MOYO [32]), but none involve wearable insoles. Dedicated insole datasets remain small-scale: recent works coupling insoles with MoCap [26, 36, 40] show promising pose reconstruction yet cover only narrow motion taxonomies. The complementary role of plantar signals with respect to other wearable modalities remains unexplored.

Overall, prior datasets have been siloed by modality, each advancing its own line of work without studying how these signals interact. Our dataset bridges this gap by jointly capturing smartphones, smartwatches, smart insoles, and a VR headset, all synchronized with a multi-view markerless motion capture system.

Wearable Motion Capture Methods. *IMU-based methods.* Sparse IMU configurations have long been explored for pose reconstruction. Early optimization methods fit body models to sensor orientations and accelerations [22], demonstrating feasibility at high computational cost. Deep learning enabled real-time

solutions: DIP [13] regressed joint rotations from six IMUs using recurrent networks, inspiring later refinements with kinematic constraints [45], physics priors [44], terrain awareness [15], or part-based modeling [47]. These advances consolidated six-IMU setups as the research standard, with some work extending to three IMUs [49]. Recent approaches turned to commodity devices: IMUPoser [24] showed that arbitrary combinations of phones, watches, and earbuds suffice for pose estimation, while MobilePoser [42] further added root translation and lightweight physics smoothing. DiffusionPoser [39] similarly targets arbitrary sparse sensors with a generative autoregressive diffusion model, but focuses on inertial inputs rather than the broader mix of consumer wearables studied here.

HMD-based methods. Head-mounted devices provide another path toward minimal sensing. AvatarPoser [14] and QuestSim [38] reconstruct body motion from only head and hand poses, while EgoEgo [17] augments HMDs with egocentric cameras to mitigate lower-body ambiguity. Diffusion models have also been adopted [6, 8] to regularize the reconstruction.

Insole-based methods. Plantar pressure signals have recently gained attention. SolePoser [40] and P2P-Insole [36] combined pressure distributions with foot-mounted IMUs to estimate motion from sparse inputs. However, these methods have only been validated on narrow locomotion tasks.

Despite this progress, each method specializes in a single modality and a fixed sensor set. Real-world use, however, is inherently cross-modal, with sensors appearing in diverse and unpredictable combinations. In the next section, we present the dataset we collected to study this scenario.

3 Dataset

In this work, we aim for a real-world motion tracking model that operates with unobtrusive devices and adapts to whatever sensors are available. This requires properties that existing datasets lack: *sparse, low-overhead* and *diverse sensing*.

To address this gap, we collected a multimodal dataset in which participants wore two smartphones, two smartwatches, a head-mounted device, and pressure sensing insoles, synchronized with professional markerless motion capture (see Fig. 2 and Fig. 1, left). All devices are commodity hardware, reflecting configurations possible in everyday use.

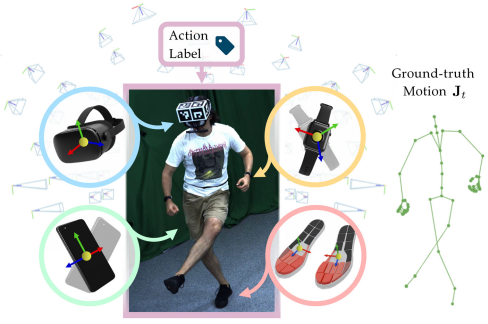


Fig. 2: Wearable capture rig. Two smart insoles (pressure + IMU), two smartwatches and two smartphones (IMUs), and an HMD (6-DoF pose), all synchronized with markerless MoCap.

3.1 Capture System

Our data acquisition takes place in a markerless motion-capture studio equipped with 120 synchronized and calibrated RGB cameras operating at 30 Hz [5]. This system provides ground-truth 3D skeletal motion, $\mathbf{J}_t \in \mathbb{R}^{J\times 3}$, where t denotes the time index, J the number of skeletal joints and 3 indicates the 3D coordinates in the studio reference frame.

Smartphones and Watches. Participants carry two smartphones in the left and right front trouser pockets and wear a smartwatch on each wrist. Each device records an orientation estimate $\mathbf{R}^S_{D,t} \in \mathrm{SO}(3)$, angular velocity $\boldsymbol{\omega}^S_{D,t} \in \mathbb{R}^3$, and linear acceleration $\mathbf{a}^S_{D,t} \in \mathbb{R}^3$, all in the local device frame, where $D \in \{P, W\}$ denotes the device type and $S \in \{L, R\}$ the body side. While these devices can expose data in a shared global frame by leveraging a magnetometer, we find such estimates often unreliable and noisy, consistent with prior reports [3, 24]. We therefore operate in the device-local frames.

Headset. We leverage a Meta Quest 3 headset [23] to obtain 6-DoF head tracking from its onboard SLAM, streamed via a Unity application as $\mathbf{H}_t \in \mathrm{SE}(3)$.

Insoles. Participants also wear Moticon OpenGo pressure-sensing insoles [25]. The per-foot measurements are $\mathbf{I}^S_t = \left[\, \mathbf{p}^S_t,\ \boldsymbol{\omega}^S_{I,t},\ \mathbf{a}^S_{I,t} \,\right] \in \mathbb{R}^{22}$ where $\mathbf{p}^S_t \in \mathbb{R}^{16}$ are pressures from 16 plantar zones, $\boldsymbol{\omega}^S_{I,t} \in \mathbb{R}^3$ is the angular velocity, and $\mathbf{a}^S_{I,t} \in \mathbb{R}^3$ the linear acceleration from the embedded IMU. Unlike the phones and watches, the insole IMU provides raw sensor signals.

As these sensors record at different rates, we subsample all signals to 30 Hz and align them to the motion capture system.

3.2 System Calibration

Each device records in its own local coordinate system and maintains an independent internal clock. To make the signals mutually consistent, they must be spatially aligned and temporally synchronized.

To that end, at the start of every session, participants perform a one-minute calibration sequence designed to excite all sensors, from which we resolve spatial alignment and temporal synchronization. In the following, we describe the calibration procedures for each device.

Headset. Following Camiletto et al. [4], we attach an ArUco board to the HMD to align it with the MoCap. By tracking the board's pose $\mathbf{N}_t \in \mathrm{SE}(3)$ from the studio cameras and collecting the HMD inside-out tracking $\mathbf{H}_t \in \mathrm{SE}(3)$, we solve

$$\mathbf{T}^*_c,\ \mathbf{T}^*_r,\ t^*_0 = \underset{\mathbf{T}_c,\,\mathbf{T}_r,\,t_0}{\arg\min} \sum_{t=1}^{T} \left\| \mathbf{N}_t - \mathbf{T}_c\, \mathbf{H}_{t+t_0}\, \mathbf{T}_r \right\|^2, \tag{1}$$

where $\mathbf{T}_c$ and $\mathbf{T}_r$ are the coordinate and mounting transformations, and t_0 is the clock offset, jointly recovering spatial and temporal alignment.

Insoles. The Moticon OpenGo insoles $\mathbf{I}_t^S$ [25] are initialized using the manufacturer's calibration. Absolute timestamps are obtained via the companion app's QR-code mechanism, which we use for alignment with the studio clock. This alignment, however, is too coarse: we observe clock drift both relative to the studio and between the two insoles. To account for this, we ask participants to perform a jumping sequence towards the end of the recording, which allows us to manually adjust offsets and correct for both absolute and relative drift.

Phones and Watches. We align each phone/watch orientation to the MoCap anatomical joint rotation (hip for phones, wrist for watches). For each device type $D \in \{P, W\}$ and side $S \in \{L, R\}$, we estimate: (i) a coordinate mapping $\mathbf{R}_c^{D,S} \in \mathrm{SO}(3)$ from the device-local frame to the studio frame, (ii) a mounting offset $\mathbf{R}_r^{D,S} \in \mathrm{SO}(3)$ between device and joint, and (iii) a clock offset $t_0^{D,S}$ between the device and ground truth. We do so by solving the following optimization, omitting the superscripts for clarity:

$$\min_{\mathbf{R}_c,\, \mathbf{R}_r,\, t_0} \sum_{t=1}^{T} d_{\mathrm{SO}(3)}(\mathbf{R}_c\, \mathbf{R}_{t+t_0}\, \mathbf{R}_r,\ \mathbf{M}_t)^2\,, \tag{2}$$

where $\mathbf{R}_t \in \mathrm{SO}(3)$ denotes the device-reported orientation, $\mathbf{M}_t \in \mathrm{SO}(3)$ is the corresponding MoCap joint rotation, and $d_{\mathrm{SO}(3)}$ is the geodesic distance on $\mathrm{SO}(3)$. The same problem is solved independently for each device/side pair.

To compensate for sensor drift and for shifts in how sensors are worn, we recompute $\mathbf{R}_r^{D,S}$ for each 20s action block. While this is an idealized setting, it compensates for misalignment that would accumulate over long sessions. In practice, the magnitude of this misalignment depends on both the sensor and how it is worn: wrist-strapped devices remain comparatively stable, whereas phones can shift inside pockets and introduce larger orientation biases. Dynamic on-body calibration methods, such as Transformer IMU Calibrator [50], are a promising way to relax this assumption in future deployments.

3.3 Dataset Structure

Our dataset includes 14 participants, each contributing two sessions to increase variability in sensor placement. Every session begins with a one-minute calibration, followed by 50 predefined actions (20 s each, separated by 4 s pauses) spanning daily and sports-related movements. The headset is operated in passthrough mode so that participants can naturally interact with their environment.

All modalities are time-synchronized at discrete steps. Each 20 s segment is annotated with the respective action label. The resulting dataset comprises over 7 hours of multimodal recordings ($\approx$ 800k synchronized frames). To assess generalization, we hold out one participant and three actions for evaluation.

Tab. 1 situates our dataset among prior efforts in wearable motion capture. Unlike earlier benchmarks that rely on dense IMU suits, lack cross-modality coverage, or forgo high-quality ground truth, our dataset combines consumer-grade smartphones, smartwatches, insoles, and a VR headset, all synchronized

Table 1: Comparison of wearable MoCap datasets. HQ-GT: high-quality ground truth from multi-camera or marker-based systems. cIMUs: commodity IMUs from consumer devices, as opposed to full IMU suits.

Dataset	Insoles	HMD	Action	HQ-GT	cIMUs	Subjects	Hours
TotalCapture [33]	✗	✗	✓	✓	✗	5	9
DIP-IMU [13]	✗	✗	✗	✗	✗	10	1.5
VT-NMD [9]	✗	✗	✗	✗	✗	16	40
CIP [30]	✗	✗	✗	✗	✗	31	16
UrbanWalk [20]	✓	✗	✓	✗	✗	20	9
UnderPressure [26]	✓	✗	✓	✗	✗	10	5.6
Ego-Exo4D [11]	✗	✓	✓	✗	✓	740	221
Nymeria [21]	✗	✓	✓	✗	✓	264	300
IMUPoser [24]	✗	✗	✓	✓	✓	10	1.2
Ours	✓	✓	✓	✓	✓	14	7

with professional multi-camera MoCap, thus providing both reliable 3D labels and systematic cross-modal coverage.

4 Method

We now present WHIP (Watch, Headset, Insole, Phone), a generative framework for reconstructing full-body motion from arbitrary combinations of consumer wearables, which serves as our baseline for studying sensor complementarity. Leveraging the dataset from Sec. 3, our architecture activates a dedicated processing pathway for each available sensor modality, scaling computation with the inputs present. The model synthesizes plausible human motions conditioned on any available sensor subset and optional action labels.

Formally, we denote a motion sequence as $X = \{\mathbf{J}_t\}_{t=1}^{T}$, where $\mathbf{J}_t \in \mathbb{R}^{J\times 3}$. Here, J is the number of skeletal joints and T is the sequence length. We use $t \in \{1, \dots, T\}$ for frame indices and reserve $\tau \in [0, 1]$ for the flow-matching time parameter. The conditioning variables are $\mathcal{C} = \{\mathbf{W}, \mathbf{H}, \mathbf{I}, \mathbf{P}, \mathcal{A}\}_{\mathcal{M}}$. Here, $\mathcal{M}$ denotes an arbitrary subset of the available inputs: smartphones ($\mathbf{P}$), smartwatches ($\mathbf{W}$), insoles ($\mathbf{I}$), headset pose ($\mathbf{H}$), and action labels ($\mathcal{A}$).

4.1 Flow Matching Framework

To capture the inherent ambiguity of mapping sparse sensor signals to human motion, we adopt a probabilistic formulation based on *Flow Matching* [18]. Our objective is to learn a velocity field $u_\tau : [0, 1] \times \mathbb{R}^D \to \mathbb{R}^D$ that transports samples from a Gaussian prior to the target motion distribution, conditioned on the available sensors. Here $D = T \times J \times 3$ denotes the dimensionality of a motion sequence.

The velocity field is induced by the probability path p_τ between the prior $p_0 = \mathcal{N}(0, I)$ and the data distribution p_1. The evolution of a sample $x \sim p_0$ is

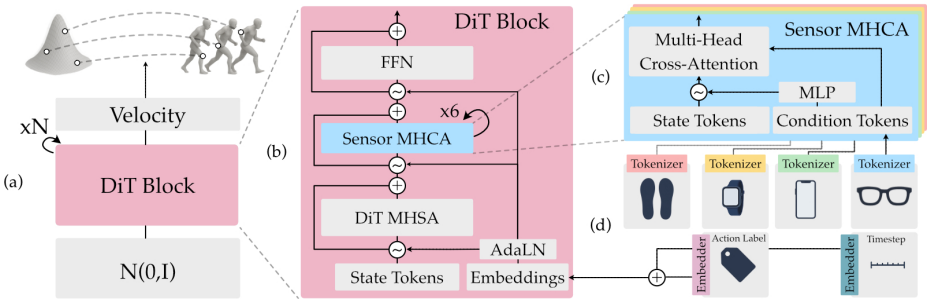


Fig. 3: WHIP model architecture. (a) High-level overview, from noise to the velocity field. (b) DiT transformer block. (c) Sensor-specific cross-attention modules. (d) Conditioning information.

consequently governed by the ordinary differential equation

$$\frac{d}{d\tau}\psi_\tau(x) = u_\tau(\psi_\tau(x)), \tag{3}$$

where ψ_τ denotes the flow map from p_0 to p_τ.

Given sensor observations $\mathcal{C}$ and ground-truth motion $X_1 \sim p_1(\cdot|\mathcal{C})$, we construct a linear interpolation path inducing the conditional velocity field

$$u_\tau(X_\tau \mid X_1, \mathcal{C}) = \frac{X_1 - X_\tau}{1-\tau}. \tag{4}$$

4.2 Model Architecture

WHIP builds upon the Diffusion Transformer (DiT) architecture [31] with modifications for conditioning on heterogeneous sensors. Fig. 3 illustrates the architecture.

Conditioning Signals. Conditioning is processed through two pathways: global embeddings for sequence-level attributes and sensor tokens for time-aligned wearable signals (Fig. 3d).

Global embeddings are sequence-level conditions comprising flow time $\tau \in [0, 1]$ and action labels $a \in \mathcal{A}$. The flow time τ is encoded through an MLP, while action labels are mapped to learnable embeddings, following DiT [31]. Embeddings then modulate the transformer via adaptive layer normalization (adaLN).

Sensor tokens encode the wearable signals independently at each time step. For each modality m, we form

$$\mathbf{C}^m = \{\mathbf{c}_t^m\}_{t=1}^T, \qquad \mathbf{c}_t^m = \mathrm{MLP}_m(\mathbf{s}_t^m), \tag{5}$$

where $\mathbf{s}_t^m$ collects the modality-specific raw sensor inputs at time t. For modalities with multiple instances such as IMUs or insoles, the same MLP is applied independently to each device, yielding separate token sequences per instance.

State Tokenization. We represent motion trajectories as sequences of per-frame tokens $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^{T}$ with $\mathbf{x}_t \in \mathbb{R}^d$, where d is the transformer's feature dimension. To facilitate conditioning, we partition each token into motion and sensor-reserved subspaces: $\mathbf{x}_t = [\mathbf{x}_t^{\text{mot}}; \mathbf{x}_t^{\text{sens}}]$ with $\mathbf{x}_t^{\text{mot}} \in \mathbb{R}^{d_m}$ and $\mathbf{x}_t^{\text{sens}} \in \mathbb{R}^{d_s}$. A linear map embeds the motion into the motion subspace, while the sensor channels are obtained by projecting each conditioning signal into a lower-dimensional subspace.

Transformer Blocks. Our model backbone consists of 8 stacked DiT blocks operating in a latent dimension of $d = 768$ with 12 attention heads. Each WHIP block extends the standard DiT architecture with a multi-modal cross-attention mechanism for sensor conditioning (see Fig. 3c).

We instantiate a dedicated cross-attention module for each conditioning modality $\mathbf{C}^m$. These modules operate conditionally: if the corresponding sensor is present, the module runs; if absent, the module is not evaluated and contributes no update (i.e., it is excluded from the sum). The motion tokens $\mathbf{X}$ serve as queries to the conditioning sequences, with the resulting updates aggregated as:

$$\mathbf{X}' = \mathbf{X} + \frac{1}{\sqrt{|\mathcal{M}|}} \sum_{m \in \mathcal{M}} \text{CrossAttn}_m(\mathbf{X}, \mathbf{C}^m), \tag{6}$$

where $\mathcal{M}$ denotes the set of active modalities. The $1/\sqrt{|\mathcal{M}|}$ normalization factor keeps the residual-connection variance consistent regardless of the number of active inputs.

This per-modality design makes the model robust to missing sensor inputs: during training, we randomly drop modalities per sample, and at inference, absent modalities are simply skipped.

4.3 Training and Inference

We optimize the model using the conditional flow matching objective [18]: $\mathcal{L}(\theta) = \mathbb{E}_\tau\left[\|u_\tau^\theta(X_\tau, \mathcal{C}) - u_\tau(X_\tau|X_1, \mathcal{C})\|^2\right]$, where $\tau \sim \mathcal{U}[0,1]$, $X_0 \sim \mathcal{N}(0, I)$, and the target velocity is as defined in Sec. 4.1. During inference we integrate the ODE in Eq. (3) with the learned velocity field from $X_0 \sim \mathcal{N}(0, I)$ over $\tau \in [0, 1]$, using an explicit Euler solver with N discretization steps.

5 Experiments

Implementation Details. We process motion sequences in sliding windows of $T = 90$ frames (3 seconds at 30 fps) using a batch size of 64. The WHIP transformer comprises 8 blocks with latent dimension $d = 768$ and 12 attention heads, each augmented with 6 dedicated cross-attention modules: HMD, insoles, left/right watches, and left/right phones. Each conditioning signal is projected to a per-modality latent dimension of 64. We train with AdamW [19] using a learning rate of 10^{-3} and warmup-stable-decay (WSD) scheduling for 110k

Table 2: Comparison of regression baselines and our generative model across different sensor subsets. **All** uses the full set of sensors, **HMD** the headset only, **Ins** the insoles only, **W+P** the average over all watch–phone combinations, and the **Avg** column reports averaged across all possible sensor configurations.

	Method	Params	MPJPE ↓		MRE ↓	PA-MPJPE ↓	N-MPJPE ↓		
			All	HMD			Ins	W+P	Avg
Unseen Actor	MLP	54.6M	127.8	147.5	203.2	95.0	157.0	123.8	120.1
	Transformer	57.4M	78.3	135.9	125.9	61.8	**147.6**	98.9	93.9
	Ours	59.4M	**56.7**	**107.4**	**54.6**	**46.5**	151.7	**83.0**	**77.9**
Unseen Action	MLP	54.6M	128.5	140.9	146.5	106.6	142.3	128.7	125.1
	Transformer	57.4M	83.3	140.5	80.3	74.1	130.6	115.8	106.0
	Ours	59.4M	**60.0**	**106.6**	**37.5**	**55.7**	**121.8**	**99.5**	**87.0**

steps. To ensure robust performance across diverse sensor configurations, we apply dropout during training, retaining each available sensor with probability 0.5. This sampling scheme exposes the model to heterogeneous subsets while maintaining a consistent marginal frequency for every modality, which we find to be a good balance between sparse single-sensor cases and fully instrumented setups.

Metrics. Following standard practice, we evaluate using: Mean Per Joint Position Error (MPJPE) after root alignment, Mean Root Error (MRE), scale-normalized MPJPE (N-MPJPE), and Procrustes-aligned MPJPE (PA-MPJPE).

Dataset and Evaluation. We evaluate generalization along two axes by withholding one subject and three action categories from the training set. Unlike deterministic regression methods, our generative model learns a conditional distribution $p_\theta(X|\mathcal{C})$ over plausible motions. While it is common to report best-of-K sample performance for such models, this would unfairly favor generative methods over regression baselines. We instead evaluate by averaging multiple predictions, effectively using the Bayes estimator under L_2 loss:

$$\hat{X} = \mathbb{E}_X[X] \approx \frac{1}{K}\sum_{k=1}^{K} X_k, \quad X_k \sim p_\theta(X|\mathcal{C}), \tag{7}$$

where we empirically approximate the posterior mean using $K = 10$ samples.

5.1 Generative vs. Regressive Baselines

We first compare our generative model against regression-based baselines that directly map sensor inputs to body poses. For a fair comparison, all methods share the same embedding networks, and we match the number of learnable parameters across models. The baselines operate on flattened latent vectors obtained from the modality-specific encoders. When a sensor is unavailable, its corresponding latent vector is masked to zero. We evaluate against two standard regression approaches: an MLP and a transformer encoder, both matched in parameter count to WHIP. Results in Tab. 2 show that our generative approach

Table 3: Comparison with prior methods, all retrained on our dataset (N-MPJPE, mm; lower is better). Wtc = smartwatch. **IMUs only** uses orientation-only IMU signals with no headset tracking: *All* = all phones and watches, *Head+Wtc* = headset orientation + watches, *Avg* = mean over all IMU subsets. **With HMD** adds the headset's 6-DoF pose: *Alone* = headset only, then *+2 Wtc* and *+All* progressively add the two watches and the full sensor set.

Method	IMUs only			With HMD		
	All	Head+Wtc	Avg	Alone	+2 Wtc	+All
IMUPoser	75.5	91.8	96.9	–	–	–
EgoEgo	–	–	–	108.4	–	–
EgoAllo	–	–	–	125.9	–	–
MocapEvery	–	–	–	141.2	86.1	–
AvatarPoser	–	–	–	–	78.2	–
AGRoL	–	–	–	–	79.8	–
HMDPoser	–	–	–	–	–	74.6
Ours	**61.7**	**82.0**	**88.2**	**104.6**	**68.7**	**52.6**

consistently outperforms both regression baselines across different sensor configurations, demonstrating the benefit of modeling motion as a generative process.

5.2 Comparison to Existing Methods

While this is the first work to address such a wide range of sensors, we can compare against existing methods on individual subsets by retraining them on our data, adapting each to our conventions and skeleton. For more details, we refer to the supplemental material. In particular, for IMUPoser [24] and HMDPoser [7], we omit IMUs on feet, pelvis, and headphones, as they are not available in our dataset. For head-pose baselines [16, 17, 43], we retrain the motion-estimation component on our data and remove visual components whenever present, retaining only the head-pose inputs available in our setting. For 3-point tracking methods [8, 14], we zero out the controller positions since our watches provide only orientation. Despite being trained for a more diverse task, i.e., reconstructing skeletal motion from arbitrary sensor combinations, our baseline model, WHIP, still outperforms previous works even though they have been exclusively trained for the respective set of sensors (see Tab. 3).

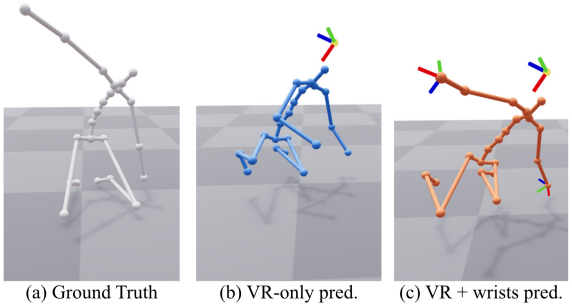
5.3 Cross-Dataset Evaluation on Nymeria

We further evaluate WHIP on Nymeria [21] to test whether the model transfers to another wearable-motion dataset. Since Nymeria does not include our exact sensor suite, we adapt the inputs by using wristband tracking in place of smartwatches, the pelvis IMU in place of phones, and labeled contacts in place of pressure insoles. Because Nymeria has no canonical train/test split and prior work such as Ego4o [35] does not release its split, our results are not directly

comparable to published numbers. Nevertheless, the trends in Tab. 4 are consistent with our dataset: adding wrist, pelvis, and contact cues progressively improves reconstruction, especially for the lower body.

Table 4: WHIP on Nymeria. N-MPJPE (mm).

Input	Full	Upper	Lower
VR	87.8	85.7	91.2
+ wrists	76.7	67.3	91.2
+ pelvis	71.9	65.0	82.6
+ contact	**68.7**	**63.8**	**76.3**



(a) Ground Truth (b) VR-only pred. (c) VR + wrists pred.

Fig. 4: WHIP on Nymeria. Wrist signals recover arm motion missed by VR-only reconstruction.

5.4 In-the-Wild Qualitative Results

To demonstrate applicability beyond controlled settings, we recorded additional sequences in unconstrained environments using a manual calibration procedure. Fig. 5 shows qualitative results: despite the domain shift from our studio-captured training data, WHIP produces plausible motion reconstructions, suggesting that the learned sensor-to-motion mapping generalizes to unconstrained environments.

5.5 Analyzing the Interplay of Modalities

Next, we provide a statistical analysis of each individual input modality, focusing on how different sensors complement one another.

For a specific set of sensors $\mathcal{C} \subseteq \mathcal{M}$ and a given metric, let $v(\mathcal{C})$ be the error (lower is better). With $|\mathcal{M}| = 7$ inputs (HMD tracking, insoles, left/right watch, left/right phone, and action label) there are $2^7 = 128$ possible combinations; we evaluate all of them and report metrics for each.

Importance of Individual Sensors. To evaluate the importance of a specific modality $m \in \mathcal{M}$, we compute the expected error reduction when added to a subset $S \subseteq \mathcal{M} \setminus \{m\}$, which we refer to as the *marginal score* of m:

$$\text{Marginal}(m) := \mathbb{E}_{S \subseteq \mathcal{M} \setminus \{m\}} \big[v(S) - v(S \cup \{m\}) \big]. \tag{8}$$

The term $v(S) - v(S \cup \{m\})$ is the *conditional* improvement of m given a subset of sensors S. Results are reported in Fig. 6. Values are in millimeters for MPJPE and N-MPJPE; higher values indicate greater benefit from including that sensor.

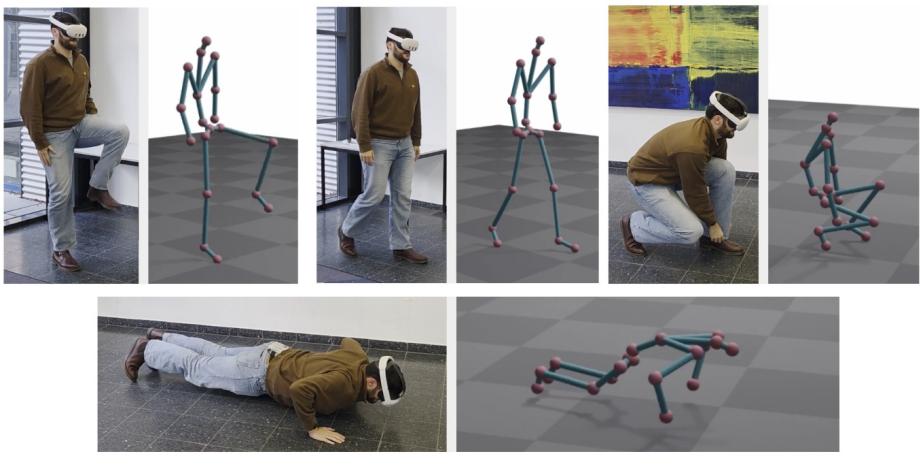


Fig. 5: Qualitative results on in-the-wild recordings. Despite the domain shift, WHIP produces plausible motion reconstructions.

The figure shows that (i) HMD contributes the most across all contexts, partly because it is the only sensor that recovers metric scale. When evaluated with N-MPJPE, its contribution, although still substantial, is less pronounced. (ii) Smartwatches are the second most important sensor overall and especially effective for arm reconstruction. (iii) Phones and pressure insoles primarily benefit the reconstruction of leg joints, where capturing pelvis motion and foot–ground contacts becomes central. (iv) Action labels provide smaller and less consistent gains than physical signals for pose accuracy.

Pairwise Interactions. To characterize how different sensor modalities cooperate or overlap, we quantify their *pairwise interaction* I. Given two sensor modalities a, b, and a subset of the remaining sensors $X \subseteq \mathcal{M} \setminus \{a, b\}$, we measure how their joint contribution deviates from the sum of their individual contributions:

$$I(a,b) = \mathbb{E}_X\big[v(X_{ab}) - v(X_a) - v(X_b) + v(X)\big] \tag{9}$$

where $X_a := X \cup \{a\}$, $X_b := X \cup \{b\}$, and $X_{ab} := X \cup \{a, b\}$. Since adding a sensor can only reduce (or maintain) the error, $I(a,b) \leq 0$: it is zero when the two sensors are perfectly complementary, i.e., their joint gain equals the sum of their individual gains, and increasingly negative as their contributions overlap (redundancy). However, since $I(a,b)$ is averaged over all sensor subsets, the marginal gains of adding sensors inevitably diminish, and perfect complementarity is unattainable. Fig. 7 summarizes the results on N-

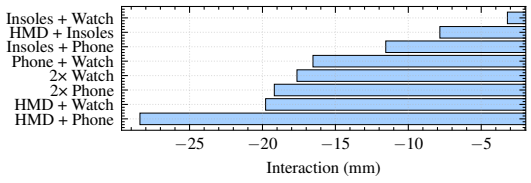


Fig. 7: Pairwise sensor interactions on N-MPJPE. Higher values indicate greater synergy.

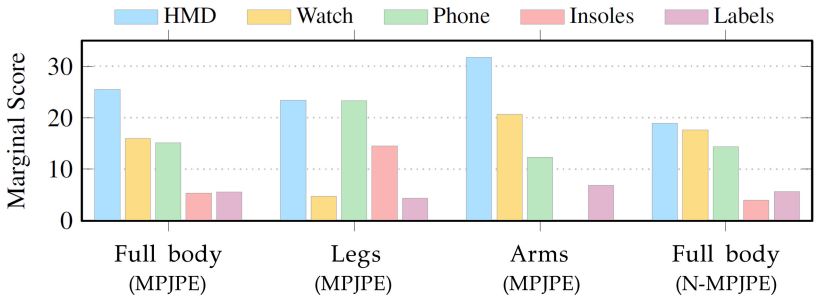


Fig. 6: Sensor importance via *marginal scores* across four contexts: full body, legs, arms, and normalized full body. Bars show the error reduction contributed by each modality (larger is better).

MPJPE. We observe that, while the insoles contribute modestly when considered alone, they exhibit the strongest synergy when paired with other modalities, providing information most distinct from other sensors.

Best-k-out-of-all Sensors. To quantify the trade-off between accuracy and instrumentation, we compute a Pareto frontier of reconstruction performance over sensor combinations. For each device count k, we identify the subset $\mathcal{C}$ with the lowest error in terms of N-MPJPE, yielding a discrete frontier. Fig. 8 shows all subsets as a scatter, with the Pareto envelope highlighted.

We observe: (i) with a single device, *HMD* dominates the frontier; (ii) among two devices, *Watch+Phone* outperforms pairings that include *HMD*, suggesting complementary cues across upper/lower body; (iii) adding *HMD* as a third device produces a marked gain; (iv) additional devices yield diminishing returns, with *Insoles* providing a modest improvement at full instrumentation.

6 Limitations and Future Work

Despite the breadth of our multimodal dataset, we did not record additional VR-centric signals such as hand-tracking from controllers, which could further improve body- and hand-focused reconstruction. We plan to enrich future recordings with additional sensor modalities. Our sensor-complementarity analysis uncovers consistent trends (e.g., the HMD's strong single-device performance, and the benefit of pairing insoles with smartwatches), but different conclusions may emerge when focusing on specific scenarios or activity types, where the relative importance of individual sensors could shift. Finally, WHIP was developed and evaluated for robustness and accuracy rather than low-latency on-device operation, leaving real-time and low latency deployment as future work.

7 Conclusion

In this work, we introduce a large, task-diverse multimodal dataset of consumer-grade wearables synchronized with ground-truth 3D motion, surpassing existing

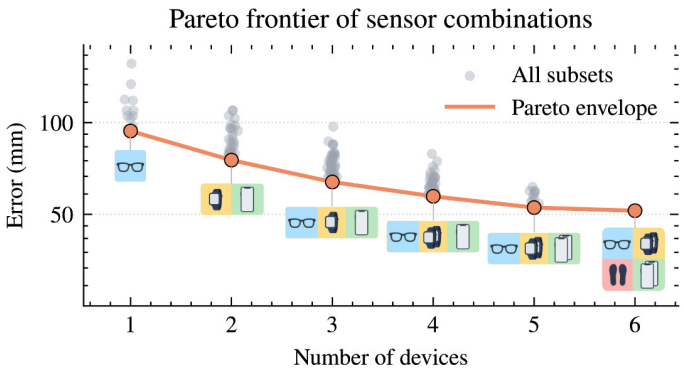


Fig. 8: Pareto frontier of N-MPJPE versus number of devices. Each point represents a sensor combination; the frontier highlights the best subset for each device count.

datasets in sensor diversity and scale. Our dataset enables unobtrusive, real-world configurations and supports a systematic analysis of sensor complementarity, quantifying how modalities interact and reinforce one another. Together, these contributions provide a valuable resource for consumer-oriented motion reconstruction methods. In addition, our proposed generative method, WHIP, demonstrates reliable reconstruction from sparse signals and consistent gains from complementary device combinations while remaining robust to missing modalities. Together, the dataset and model advance consumer-grade motion capture and provide a practical foundation for future work on pose reconstruction and broader in-the-wild generalization.

Acknowledgements. This work was funded by the Saarbrücken Research Center for Visual Computing and Artificial Intelligence (VIA).

References

1. Abdul Razak, A.H., Zayegh, A., Begg, R.K., Wahab, Y.: Foot plantar pressure measurement system: A review. Sensors **12**(7), 9884–9912 (2012). https://doi.org/10.3390/s120709884, https://www.mdpi.com/1424-8220/12/7/9884
2. Akada, H., Wang, J., Shimada, S., Takahashi, M., Theobalt, C., Golyanik, V.: Unrealego: A new dataset for robust egocentric 3d human motion capture. In: European Conference on Computer Vision (ECCV) (2022)
3. Apple Inc.: Getting processed device-motion data. https://developer.apple.com/documentation/coremotion/getting-processed-device-motion-data (2025), accessed: 2026-06-29
4. Boscolo Camiletto, A., Wang, J., Alvarado, E., Dabral, R., Beeler, T., Habermann, M., Theobalt, C.: Frame: Floor-aligned representation for avatar motion from egocentric video. CVPR (2025)
5. The Captury GmbH, Saarbrücken, Germany: The Captury: Markerless Motion Capture System (2023), available at https://www.thecaptury.com. Accessed: 2026-06-29

6. Castillo, A., Escobar, M., Jeanneret, G., Pumarola, A., Arbeláez, P., Thabet, A., Sanakoyeu, A.: Bodiffusion: Diffusing sparse observations for full-body human motion synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 4223–4233 (2023). https://doi.org/10.1109/ICCVW60793.2023.00456
7. Dai, P., Zhang, Y., Liu, T., Fan, Z., Du, T., Su, Z., Zheng, X., Li, Z.: Hmd-poser: On-device real-time human motion tracking from scalable sparse observations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 874–884 (2024). https://doi.org/10.1109/CVPR52733.2024.00089
8. Du, Y., Kips, R., Pumarola, A., Starke, S., Thabet, A., Sanakoyeu, A.: Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 481–490 (2023). https://doi.org/10.1109/CVPR52729.2023.00054
9. Geissinger, J.H., Asbeck, A.T.: Motion inference using sparse inertial sensors, self-supervised learning, and a new dataset of unscripted human motion. Sensors **20**(21), 6330 (2020). https://doi.org/10.3390/s20216330
10. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., Martin, M., Nagarajan, T., Radosavovic, I., Ramakrishnan, S.K., Ryan, F., Sharma, J., Wray, M., Xu, M., Xu, E.Z., Zhao, C., Bansal, S., Batra, D., Cartillier, V., Crane, S., Do, T., Doulaty, M., Erapalli, A., Feichtenhofer, C., Fragomeni, A., Fu, Q., Gebreselasie, A., Gonzalez, C., Hillis, J., Huang, X., Huang, Y., Jia, W., Khoo, W., Kolar, J., Kottur, S., Kumar, A., Landini, F., Li, C., Li, Y., Li, Z., Mangalam, K., Modhugu, R., Munro, J., Murrell, T., Nishiyasu, T., Price, W., Puentes, P.R., Ramazanova, M., Sari, L., Somasundaram, K., Southerland, A., Sugano, Y., Tao, R., Vo, M., Wang, Y., Wu, X., Yagi, T., Zhao, Z., Zhu, Y., Arbelaez, P., Crandall, D., Damen, D., Farinella, G.M., Fuegen, C., Ghanem, B., Ithapu, V.K., Jawahar, C.V., Joo, H., Kitani, K., Li, H., Newcombe, R., Oliva, A., Park, H.S., Rehg, J.M., Sato, Y., Shi, J., Shou, M.Z., Torralba, A., Torresani, L., Yan, M., Malik, J.: Ego4d: Around the world in 3,000 hours of egocentric video (2022), https://arxiv.org/abs/2110.07058
11. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., et al.: Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19383–19400 (2024)
12. Han, X., Senderling, B., To, S., Kumar, D., Whiting, E., Saito, J.: Groundlink: A dataset unifying human body movement and ground reaction dynamics. In: SIGGRAPH Asia 2023 Conference Papers. p. 1–10. SA '23, ACM (Dec 2023). https://doi.org/10.1145/3610548.3618247, http://dx.doi.org/10.1145/3610548.3618247
13. Huang, Y., Kaufmann, M., Aksan, E., Black, M.J., Hilliges, O., Pons-Moll, G.: Deep inertial poser learning to reconstruct human pose from sparse inertial measurements in real time. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **37**(6), 185:1–185:15 (Nov 2018)
14. Jiang, J., Streli, P., Qiu, H., Fender, A., Laich, L., Snape, P., Holz, C.: Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In: European Conference on Computer Vision (ECCV). pp. 443–460 (2022). https://doi.org/10.1007/978-3-031-20065-6_26

15. Jiang, Y., Ye, Y., Gopinath, D.E., Won, J., Winkler, A.W., Liu, C.K.: Transformer inertial poser: Real-time human motion reconstruction from sparse imus with simultaneous terrain generation. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–9 (2022). https://doi.org/10.1145/3550469.3555428
16. Lee, J., Joo, H.: Mocap everyone everywhere: Lightweight motion capture with smartwatches and a head-mounted camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
17. Li, J., Liu, K., Wu, J.: Ego-body pose estimation via ego-head pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17142–17151 (2023)
18. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
19. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR) (2019)
20. Losing, V., Hasenjäger, M.: A multi-modal gait database of natural everyday-walk in an urban environment. Scientific Data **9**(473) (2022). https://doi.org/10.1038/s41597-022-01580-3
21. Ma, L., Ye, Y., Hong, F., Guzov, V., Jiang, Y., Postyeni, R., Pesqueira, L., Gamino, A., Baiyya, V., Kim, H.J., Bailey, K., Fosas, D.S., Liu, C.K., Liu, Z., Engel, J., Nardi, R.D., Newcombe, R.: Nymeria: A massive collection of multimodal egocentric daily motion in the wild (2024), https://arxiv.org/abs/2406.09905
22. von Marcard, T., Rosenhahn, B., Black, M.J., Pons-Moll, G.: Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. Computer Graphics Forum **36**(2), 349–360 (2017). https://doi.org/10.1111/cgf.13131
23. Meta Platforms, Inc.: Meta Quest 3 VR Headset. https://www.meta.com/quest/quest-3/ (2025), accessed: 2025-09-01
24. Mollyn, V., Arakawa, R., Goel, M., Harrison, C., Ahuja, K.: 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. CHI '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3544548.3581392, https://doi.org/10.1145/3544548.3581392
25. Moticon ReGo AG: Moticon OpenGo Sensor Insoles. https://moticon.com/opengo (2025), accessed: 2025-09-01
26. Mourot, L., Hoyet, L., Clerc, F.L., Hellier, P.: Underpressure: Deep learning for foot contact detection, ground reaction force estimation and footskate cleanup. Computer Graphics Forum **41**(8), 195–206 (2022). https://doi.org/10.1111/cgf.14635
27. novel GmbH: loadsol: Wireless in-shoe plantar pressure sensors. https://novel.de/products/loadsol (2025), accessed: 2025-09-01
28. novel GmbH: pedar: In-shoe plantar pressure measurement system. https://novel.de/products/pedar (2025), accessed: 2025-09-01
29. NURVV: NURVV Run Smart Insoles. https://www.prnewswire.com/news-releases/nurvv-debuts-first-ever-tech-to-accurately-measure-running-power-using-smart-insoles-301365702.html (2021), accessed: 2026-06-24
30. Palermo, M., Cerqueira, S., André, J., Pereira, A., Santos, C.P.: Complete inertial pose dataset: from raw measurements to pose with low-cost and high-end marg sensors. arXiv preprint arXiv:2202.06164 (2022). https://doi.org/10.48550/arXiv.2202.06164
31. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)

32. Tripathi, S., Müller, L., Huang, C.H.P., Omid, T., Black, M.J., Tzionas, D.: 3D human pose estimation via intuitive physics. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4713–4725 (2023)
33. Trumble, M., Gilbert, A., Malleson, C., Hilton, A., Collomosse, J.: Total capture: 3d human pose estimation fusing video and inertial sensors. In: 2017 British Machine Vision Conference (BMVC) (2017)
34. Vicon Motion Systems Ltd., Oxford, UK: Vicon Motion Capture System (2023), available at https://www.vicon.com. Accessed: 2026-06-29
35. Wang, J., Dabral, R., Luvizon, D., Cao, Z., Liu, L., Beeler, T., Theobalt, C.: Ego4o: Egocentric human motion capture and understanding from multi-modal input. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22668–22679 (June 2025)
36. Watanabe, A., Aisuwarya, R., Jing, L.: P2p-insole: Human pose estimation using foot pressure distribution and motion sensors. arXiv preprint arXiv:2505.00755 (2025)
37. Werling, K., Kaneda, J., Tan, T., Agarwal, R., Skov, S., Van Wouwe, T., Uhlrich, S., Bianco, N., Ong, C., Falisse, A., et al.: Addbiomechanics dataset: Capturing the physics of human motion at scale. In: European Conference on Computer Vision. pp. 490–508. Springer (2024)
38. Winkler, A., Won, J., Ye, Y.: Questsim: Human motion tracking from sparse sensors with simulated avatars. In: SIGGRAPH Asia 2022 Conference Papers. SA '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3550469.3555411, https://doi.org/10.1145/3550469.3555411
39. Wouwe, T.V., Lee, S., Falisse, A., Delp, S.L., Liu, C.K.: 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 (CVPR). pp. 2513–2523 (2024). https://doi.org/10.1109/CVPR52733.2024.00243
40. Wu, E., Khirodkar, R., Koike, H., Kitani, K.: Soleposer: Full body pose estimation using a single pair of insole sensor. In: Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology. UIST '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3654777.3676418, https://doi.org/10.1145/3654777.3676418
41. Wu, E., Peng, Y., Khirodkar, R., Koike, H., Kitani, K.: Dual-modal 3d human pose estimation using insole foot pressure sensors. In: 2024 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct). pp. 131–135 (2024). https://doi.org/10.1109/ISMAR-Adjunct64951.2024.00036
42. Xu, V., Gao, C., Hoffmann, H., Ahuja, K.: 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. pp. 1–11 (2024). https://doi.org/10.1145/3654777.3676461
43. Yi, B., Ye, V., Zheng, M., Li, Y., Müller, L., Pavlakos, G., Ma, Y., Malik, J., Kanazawa, A.: Estimating body and hand motion in an ego-sensed world. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7072–7084 (June 2025)
44. Yi, X., Zhou, Y., Habermann, M., Shimada, S., Golyanik, V., Theobalt, C., Xu, F.: 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 (CVPR). pp. 13157–13168 (2022). https://doi.org/10.1109/CVPR52688.2022.01282

45. Yi, X., Zhou, Y., Xu, F.: Transpose: real-time 3d human translation and pose estimation with six inertial sensors. ACM Trans. Graph. **40**(4) (Jul 2021). https://doi.org/10.1145/3450626.3459786, https://doi.org/10.1145/3450626.3459786
46. Zhang, H., Ren, S., Yuan, H., Zhao, J., Li, F., Sun, S., Liang, Z., Yu, T., Shen, Q., Cao, X.: Mmvp: A multimodal mocap dataset with vision and pressure sensors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21842–21852 (2024). https://doi.org/10.1109/CVPR52733.2024.02063
47. Zhang, Y., Xia, S., Chu, L., Yang, J., Wu, Q., Pei, L.: Dynamic inertial poser (dynaip): Part-based motion dynamics learning for enhanced human pose estimation with sparse inertial sensors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1889–1899 (2024)
48. Zhao, D., Wei, Z., Mahmud, J., Frahm, J.M.: Egoglass: Egocentric-view human pose estimation from an eyeglass frame. In: 2021 International Conference on 3D Vision (3DV). pp. 32–41 (2021). https://doi.org/10.1109/3DV53792.2021.00014
49. Zhu, Z., Zhao, Y., Hu, Y., Wang, G., Qiu, H., Zheng, B., Yan, C., Xu, F.: Progressive inertial poser: Progressive real-time kinematic chain estimation for 3d full-body pose from three imu sensors. IEEE Transactions on Instrumentation and Measurement (2025)
50. Zuo, C., Huang, J., Jiang, X., Yao, Y., Shi, X., Cao, R., Yi, X., Xu, F., Guo, S., Qin, Y.: Transformer IMU calibrator: Dynamic on-body IMU calibration for inertial motion capture. ACM Transactions on Graphics **44**(4), 1–14 (July 2025). https://doi.org/10.1145/3730937