Step2Motion: Locomotion Reconstruction from Pressure Sensing Insoles

J. L. Ponton[1] E. Alvarado[2] L. G. Foo[2] N. Pelechano[1] C. Andujar[1] and M. Habermann[2]

[1]Universitat Politècnica de Catalunya, Barcelona, Spain
[2]Max Planck Institute for Informatics, Saarbrücken, Germany

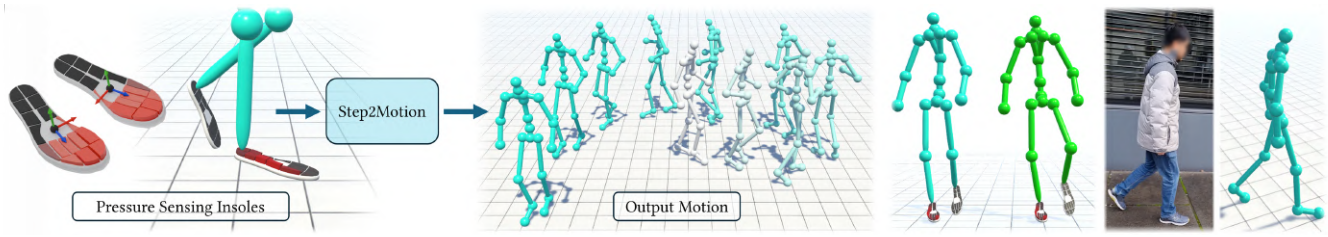

Figure 1: ***Locomotion reconstruction from pressure sensing insoles.*** *(Left) Our method achieves locomotion reconstruction using only data from insoles, each equipped with an IMU and 16 pressure sensors. (Middle) Representation of a user walking with the insole sensors, which are also able to capture the balanced pose when standing on one foot, among others. (Right) Reconstruction of in-the-wild locomotion while the subject wears the insole sensors.*

Abstract

Human motion is fundamentally driven by continuous physical interaction with the environment. Whether walking, running, or simply standing, the forces exchanged between our feet and the ground provide crucial insights for understanding and reconstructing human movement. Recent advances in wearable insole devices offer a compelling solution for capturing these forces in diverse, real-world scenarios. Sensor insoles pose no constraint on the users' motion (unlike mocap suits) and are unaffected by line-of-sight limitations (in contrast to optical systems). These qualities make sensor insoles an ideal choice for robust, unconstrained motion capture, particularly in outdoor environments. Surprisingly, leveraging these devices with recent motion reconstruction methods remains largely unexplored. Aiming to fill this gap, we present ***Step2Motion****, the first approach to reconstruct human locomotion from multi-modal insole sensors. Our method utilizes pressure and inertial data—accelerations and angular rates—captured by the insoles to reconstruct human motion. We evaluate the effectiveness of our approach across a range of experiments to show its versatility for diverse locomotion styles, from simple ones like walking or jogging up to moving sideways, on tiptoes, slightly crouching, or dancing. The complete source code, trained model, data, and supplementary material used in this paper can be found at: https://vcai.mpi-inf.mpg.de/projects/Step2Motion/*

CCS Concepts

• ***Computing methodologies*** → ***Motion capture; Motion processing; Animation;*** *Learning paradigms;*

1. Introduction

Human motion reconstruction plays a crucial role in a wide range of fields, ranging from entertainment-related applications (games, VR) to those involving greater biomechanical complexity (sports, rehabilitation). Consequently, there is a growing demand for high-quality, accurate motion capture systems. However, current technologies often present barriers that limit their widespread use. Optical (external or egocentric) and markerless systems, while accurate, are typically complex and expensive, require specialized equipment, and perform better in controlled environments, limiting their practicality in outdoor recordings [SPS*11; RRC*16; LHR*21; LLW23; SHS*24]. Although less constrained by the capture area, IMU-based systems require specialized suits or external attachments that restrict movement and require frequent calibration [HKA*18; YZX21; YZH*22; JYG*22; YZH*23]. Applications such as sports analytics and injury prevention often require solutions where unconstrained movement and ease of use are vital.

Ideally, such a system should be easy to set up and wear, capable of recording arbitrary motion outside a recording studio (e.g., mountain hiking), and robust enough to handle intense activity without sensor displacement or occlusions (e.g., rugby).

Contact dynamics could play an important role in the search for such new capture systems. Legged motion is fundamentally driven by the continuous action of the feet against the ground, producing, in return, a reaction force distributed over the contact area. Related factors such as center of pressure (CoP), distributed weight, or impact forces caused by the foot strike during the gait cycle are often enough to serve as an accurate descriptor of human movement [RJH08; WWRS14; SP17]. Traditionally, force plates have been used to measure dynamics during gait [BSM14]. More recently, wearable insole devices have proved to be an excellent way to measure gait dynamics in arbitrary environments [SGNA24], although their use has often been limited to motion analysis; both their hardware features and output data make them a potential candidate in reconstruction tasks, suitable for novel entertainment, VR or biomechanics.

This paper introduces **Step2Motion**, a novel deep learning-based approach for reconstructing human locomotion and root motion using pressure and inertial measurements from insole sensors. Our method leverages this information to condition a diffusion-based motion reconstruction model. We develop a new multi-head cross-attention mechanism to effectively incorporate the multi-modal nature of insole data, enabling the network to selectively attend to different sensor modalities based on the body part being reconstructed and capture the complex relationships between insole measurements and human movement.

To our knowledge, this is the first approach to achieve general locomotion reconstruction solely from insole sensor data. While recent work such as Smart Insole [HLP*24] has demonstrated pose estimation from pressure data, our approach differs fundamentally in scope, modality, and hardware. Han et al. [HLP*24] rely on high-density research-grade sensor grids (>600 sensors per foot) to perform discrete activity classification and local pose estimation relative to the pelvis. In contrast, Step2Motion targets the distinct problem of continuous locomotion reconstruction using sparse, consumer-grade hardware (only 16 sensors per foot + IMU). We reconstruct the global root trajectory and synthesize temporally coherent animation via a diffusion framework, rather than performing per-frame keypoint regression.

We demonstrate the versatility of our proposed algorithm and hardware setup by reconstructing diverse locomotion styles, including walking, dancing, jogging, or tiptoeing. Through extensive evaluation and analysis, we provide insight into our design choices and highlight the reconstruction effectiveness of our approach. The main contributions of our paper can be summarized as follows:

- We propose **Step2Motion**—the first method for general human locomotion reconstruction from insole sensors. Our system accurately reconstructs lower-body motion while synthesizing plausible upper-body movements that naturally align with the reconstructed motion.
- We introduce a displacement predictor network that effectively predicts root motion displacement from insole sensors.
- Along with a comprehensive set of evaluations, we recorded a new motion capture dataset paired with insole readings, which we have made publicly available under the following link: https://vcai.mpi-inf.mpg.de/projects/Step2Motion/

2. Related Work

2.1. Motion Capture from Body-worn Sensors

Optical-based motion capture systems [Vic25; Opt25] achieve highly accurate motion reconstruction but are primarily used in indoor environments with multi-camera setups. In recent years, motion capture with body-worn sensors [Xse25] has gained popularity due to its versatility in unconstrained outdoor environments and robustness to occlusions and lighting conditions. Body-worn sensor approaches can be broadly classified by sensor type: Inertial Measurement Units (IMUs), outside-in trackers, and egocentric cameras.

Inertial Measurement Units. IMU-based mocap achieves precise pose reconstruction with multiple IMUs. Efforts to enhance accessibility have reduced sensor counts, starting with six IMUs and offline optimization [vMRBP17]. For real-time use, deep learning architectures have emerged [HKA*18; YZX21]. However, persistent issues remain, like drift errors from integrating acceleration and angular rate. Subsequent methods address drift reduction: Jiang et al. [JYG*22] apply Transformers for temporal data, while Yi et al. [YZH*22] use physics-based optimization, later refining non-inertial effects [YZX24]. Armani et al. [AQJH24] estimate inter-sensor distances to mitigate drift and jitter. Additionally, monocular cameras paired with SLAM algorithms [GMSP21; YZH*23; LJ24] help locate subjects and minimize drift. Unlike full-body IMU suits prone to looseness, our method employs just two IMUs embedded in the insoles. This design leverages the feet's stability as natural anchors, reducing sensor displacement.

Outside-in Trackers. The rising demand for VR has driven the development of high-accuracy positional and rotational sensors. Early efforts, such as [YKL21], used recurrent neural networks for motion reconstruction with four 6DoF sensors. To avoid additional tracking devices, Ponton et al. [PYAP22] proposed a motion-matching system for one or three sensors. Other works with three sensors explored deep learning methods, including transformers [JSQ*22], reinforcement learning [WWY22], diffusion models [DKP*23], and vector quantization [SSH*24]. Recent research focuses on enhancing reconstruction accuracy using six sensors [PYA*23], variable sensor counts [PPA*25], and integrating 6DoF data with IMUs [YKM*24]. Insoles operate seamlessly outdoors and are free from restrictions, unlike positional sensors that depend on external devices. Despite progress in built-in cameras, these systems remain constrained by limited capture space and bright outdoor scenes.

Egocentric Cameras. Egocentric camera-based estimation eliminates the need for additional tracking devices. However, it faces other challenges, like occlusion and lighting variation. Early approaches [SPS*11; RRC*16] tackled these using body-worn and dual fisheye cameras, further refined by a single head-mounted

fisheye camera [XCZ*19]. Advances in deep learning enhanced fisheye-based estimation [JI21; TPAB19; WLX*21; TAP*23; WLX*23]. Zhang et al. [ZYG21] improved lens distortion handling through automatic calibration, while physics-based controllers added realism [YK18; YK19]. Recently, Li et al. [LLW23] simplified motion capture with a two-stage estimation process, bypassing paired egocentric videos and motions. Such cameras often struggle with lower-body accuracy due to visibility constraints. In contrast, our method ensures precise contact estimation and infers accurate locomotion predictions, making it resilient to occlusions and ideal for advancing physics-based motion techniques.

2.2. Pressure Sensors for Character Animation

Foot pressure data is crucial for understanding locomotion, as it reflects the forces exerted on the ground. Traditionally analyzed with force plates [HST*23], this data is paired with kinematic mocap to predict ground reaction forces (GRFs) and the center of pressure (CoP). Such ideas extend to pose generation, allowing animators to draw foot pressure maps and retrieve the corresponding poses from a database [YP03] or use the exerted impact to alter the character's locomotion [APRC22] and environment [AAR*24]. Grimm et al. [GSHG11] employed force plates in a mattress to classify patient poses using nearest-neighbor search based on basic pressure patterns. Additionally, joint data has been employed to predict foot pressure maps [SRF*20].

Wireless insole sensors provide temporal pressure distribution data per foot and integrate seamlessly into regular footwear. Despite their potential, research in this area remains limited. Most studies use insole sensors for precise foot contact labeling. Mourot et al. [MHCH22] introduced a method to correct foot sliding artifacts, pairing mocap with insole data to train a model estimating vertical ground reaction forces (vGRF) during locomotion. Addressing the inverse problem, Wu et al. [WKKK24] focused on predicting skiing poses from insole data. However, due to the limited input, their upper-body reconstruction appears restricted, and root motion is not shown. While these works focus on a single type of motion, we prove how our method can generalize over different types and motion styles. Han et al. [HLP*24] presented a CNN-based approach for estimating 3D poses using insoles with over 600 pressure sensors per foot, focusing on activity classification rather than motion reconstruction. Our approach uses publicly available insoles, combining cross-modality inputs (pressure and IMU) to produce smooth animation data.

2.3. Diffusion-based Motion Synthesis

Diffusion models [HJA20] have recently been applied to human motion synthesis from text and audio inputs. Approaches include U-Nets [DMGT23] and transformers [TRG*23; KKC23; ZCP*24] to capture temporal dependencies. Subsequent research has significantly advanced motion synthesis and reconstruction using various conditioning signals. For synthesis, Mughal et al. [MDH*24] generated gestures from audio via latent diffusion, while Sun et al. [SZH*24] employed LLMs for part-specific motion from text, and Cohan et al. [CTR*24] developed a text-conditioned motion in-betweening framework. In reconstruction, Zhang et al. [ZLHA24] achieved long-term motion by encoding temporality into denoising. Motion was also reconstructed from noisy RGB(-D) videos by Zhang et al. [ZBX*24], and from IMUs by Van Wouwe et al. [VLF*24], though the latter optionally used insoles only for contact labeling. Despite significant progress in diffusion-based motion generation, prior works have not examined conditioning these models on insole data features, like pressure, acceleration, and total force. Therefore, we introduce the first diffusion-based model that successfully integrates these diverse modalities, overcoming challenges posed by their unique characteristics.

3. Method

Our goal is to recover human locomotion solely from insole measurements (see Figure 1). Such measurements are multimodal, featuring, for example, pressure and acceleration (data is explained in Section 3.1). Additionally, the limited sensing capabilities lead to inherent ambiguities; similar readings can correspond to the same pose. The multimodal nature and this inherent ambiguity present an interesting research question that we formalize in Section 3.2. At the core, we solve this challenging task by introducing a transformer-based diffusion model with an attention mechanism to account for the data multimodality (as shown in Figure 2). We provide a high-level overview in Section 3.3, followed by the detailed architecture in Section 3.4.

3.1. Insole Sensors

In this work, we use Moticon's OpenGo Sensor Insoles, which incorporate an Inertial Measurement Unit (IMU) in addition to 16 pressure sensors per foot. These are worn like regular shoe insoles and are available in different sizes. The raw data provided by each insole sensor includes:

- **Pressure sensors,** $\mathbf{c_p} \in \mathbb{R}^{16}$: Each element represents the pressure applied to one of the 16 sensors distributed across the insole, measured in $1/4$ N/cm^2.
- **Linear acceleration,** $\mathbf{c_a} \in \mathbb{R}^3$ is measured by the IMU in its local coordinate frame, in units of g. To facilitate the network's understanding of movement direction, we transform this information into a fixed world coordinate frame by integrating the angular rate data.
- **Angular rate,** $\mathbf{c_r} \in \mathbb{R}^3$ is measured by the IMU in its local coordinate frame, in units of degree/s.
- **Total force value,** $c_f \in \mathbb{R}^1$ represents the ground reaction force magnitude measured in N.
- **Center of pressure,** $\mathbf{c_c} \in \mathbb{R}^2 \cap [-0.5, 0.5]$ is a normalized CoP location as a percentage of the insole length and width.

3.2. Problem Definition

Human motion can be represented as an ordered sequence of W poses, $(\mathbf{p}^{(i)})_{i=1}^{W}$, of a humanoid skeleton with J joints. Each pose, $\mathbf{p} = (\mathbf{d}, \mathbf{j})$, comprises a world space displacement 3D vector $\mathbf{d} \in \mathbb{R}^3$ defining the root's 3D movement from the previous pose, and a set of relative joint positions $\mathbf{j} \in \mathbb{R}^{(J-1)\times 3}$ represented as root-relative 3D vectors. The global position (i.e., the root position) in frame i is the cumulative sum of displacements: $\sum_{f=0}^{i} \mathbf{d}^{(f)}$.

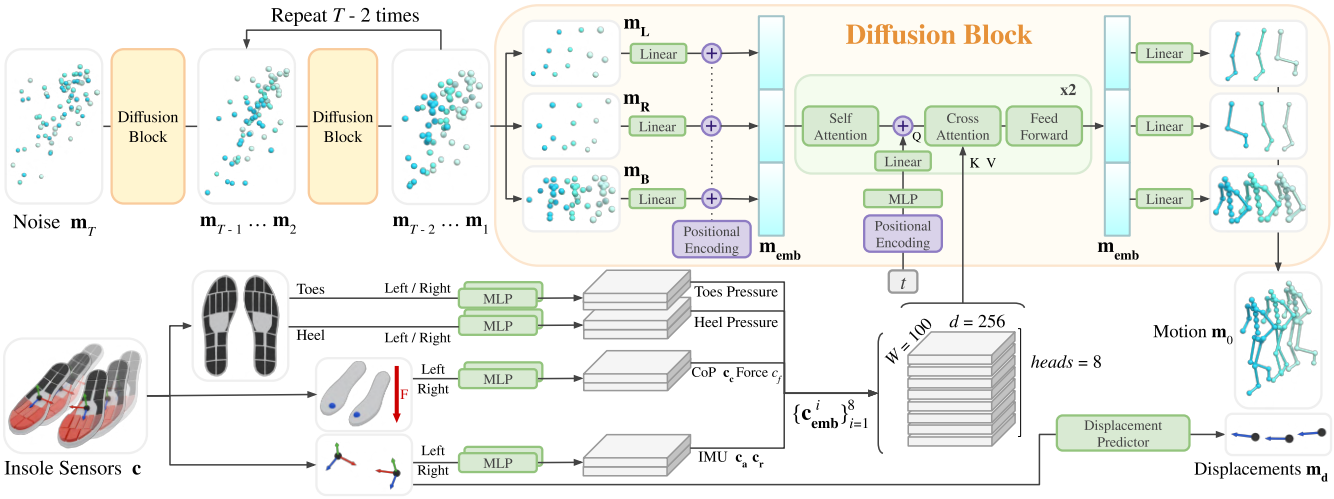

Figure 2: ***Overview of the diffusion-based motion reconstruction process conditioned to insole sensor data.*** *The method starts with some unit gaussian noise* $\mathbf{m}_T$ *and returns the denoised motion sequence* $\mathbf{m}_0$ *after* T *iterations. The diffusion block is executed each iteration: given a noisy input motion sample* $\mathbf{m}_t$ *at timestep* t*, the output is the denoised pose* $\mathbf{m}_{t-1}$ *at timestep* $t-1$*. The input is divided into three representations,* $\mathbf{m_L}$*,* $\mathbf{m_R}$*,* $\mathbf{m_B}$*, to facilitate part-wise attention within the Transformer network. The insole data* $\mathbf{c}$ *is partitioned into eight components to be used as heads for multi-head cross-attention. Sinusoidal positional encodings [VSP*17] are employed to encode motion temporality and the current diffusion timestep. An additional Transformer network predicts the displacements* $\mathbf{m_d}$ *of the corresponding motion* $\mathbf{m}$ *from the IMU readings of both feet.*

In addition to poses, the insole readings $(\mathbf{c}^{(i)})_{i=1}^{W}$ at the i-th frame is a 50-dimensional vector as follows:

$$\mathbf{c}^{(i)} = (\mathbf{c_p^L}, \mathbf{c_a^L}, \mathbf{c_r^L}, c_f^L, \mathbf{c_c^L}, \mathbf{c_p^R}, \mathbf{c_a^R}, \mathbf{c_r^R}, c_f^R, \mathbf{c_c^R}) \in \mathbb{R}^{50} \quad (1)$$

comprising 25 features per foot. Our objective is to synthesize a sequence of W poses $(\mathbf{p}^{(i)})_{i=1}^{W}$ from the corresponding insole readings $(\mathbf{c}^{(i)})_{i=1}^{W}$.

3.3. Overview

Our approach to motion reconstruction from insole data is based on two primary components: a diffusion model for reconstructing poses (Figure 2). We employ a diffusion probabilistic model to synthesize poses due to its demonstrated ability to capture complex distributions and synthesize high-quality data. We operate on a sequence of poses, therefore capturing the temporal consistency of the motion. To further enhance the synthesis, we incorporate the insole information through a carefully designed cross-attention mechanism that allows the network to focus on relevant sensor data based on the specific body part and the reconstructed motion. A separate Transformer network is used to regress the root displacements of the motion from the IMU data provided by the insole sensors. This two-stage pipeline allows for more effective processing of distinct feature types, preventing the displacement information from being ignored, a common issue even with input standardization.

3.4. Network Structure

Diffusion Probabilistic Framework. We employ a diffusion probabilistic model framework [HJA20] consisting of a *forward diffusion* and *reverse diffusion* process. The *forward diffusion* process is a Markovian chain that iteratively adds Gaussian noise to an initial sequence of poses, $\mathbf{m}^{(i)} = (\mathbf{j}^{(i)}, \dots, \mathbf{j}^{(i+W)}) \in \mathbb{R}^{W\times(J-1)\times 3}$, obtained by sliding a window of size W over the complete pose sequence $(\mathbf{p}^{(i)})_{i=1}^{F}$. This process transforms the input $\mathbf{m}_0$ into a sample from a standard Gaussian distribution $\mathbf{m}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, after a number of timesteps T. Formally, the *forward diffusion*, q, is defined as follows:

$$q(\mathbf{m}_t|\mathbf{m}_{t-1}) = \mathcal{N}(\mathbf{m}_t; \sqrt{1-\beta_t}\mathbf{m}_{t-1}, \beta_t \mathbf{I}) \quad (2)$$

$$q(\mathbf{m}_{1:T}|\mathbf{m}_0) = \prod_{t=1}^{T} q(\mathbf{m}_t|\mathbf{m}_{t-1}) \quad (3)$$

where $\{\beta_t \in (0,1)\}_{t=1}^{T}$ controls the variance schedule.

Conversely, the *reverse diffusion* process, p, reconstructs a pose sequence from Gaussian noise. Starting from $\mathbf{m}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, it progressively denoises the input over T timesteps to recover the original pose sequence $\mathbf{m}_0$:

$$p_\theta(\mathbf{m}_{0:T}) = p(\mathbf{m}_T) \prod_{t=1}^{T} p_\theta(\mathbf{m}_{t-1}|\mathbf{m}_t) \quad (4)$$

where $p_\theta(\mathbf{m}_{t-1}|\mathbf{m}_t)$ is approximated utilizing a neural network $f_\theta(\mathbf{m}_{t-1}|\mathbf{m}_t, t, \mathbf{c})$ conditioned on the previous denoised sequence $\mathbf{m}_t$, the timestep t, the corresponding insole readings $\mathbf{c} \in \mathbb{R}^{W\times 50}$, and with parameters θ. Following Tevet et al. [TRG*23] and Du et al. [DKP*23], we directly predict the denoised poses $\mathbf{m}_{t-1}$ at each step.

Body-Partitioned Pose Encoding. Motivated by the success of Transformers [VSP*17] in motion synthesis [TRG*23; MDH*24;

ZDC*24], we use them as the core of f_θ. To leverage the inherent structure of the data, we decompose each pose into three body parts: left leg, right leg, and the remaining body. This does not imply using independent Transformers. Instead, we generate a distinct vector embedding for each body part at every frame. These tokens are concatenated into a single sequence $\mathbf{m}_{emb}$. Consequently, the Transformer applies self-attention globally across both spatial (body parts) and temporal (frames) dimensions. This design enhances communication, enabling the network to be highly selective; for example, the left leg token at frame i can attend to the right leg token at frame $i-5$. This decomposition also allows the Transformer's attention mechanism to focus on relevant body parts based on the insole readings. Since each insole primarily influences the corresponding leg, this separation guides the network to prioritize pertinent information.

Specifically, we partition the pose sequence $\mathbf{m} \in \mathbb{R}^{W\times(J-1)\times 3}$ into (assuming four joints per leg) $\mathbf{m_L} \in \mathbb{R}^{W\times 4\times 3}$, $\mathbf{m_R} \in \mathbb{R}^{W\times 4\times 3}$, and $\mathbf{m_B} \in \mathbb{R}^{W\times(J-1-8)\times 3}$), representing the left leg, right leg, and the rest of the body, respectively. Each part is then projected to the Transformer's embedding dimension d and augmented with sinusoidal positional encodings [VSP*17]. Time embeddings are generated by passing the sinusoidal positional embedding through a two-layer MLP. The final input to the Transformer layers, $\mathbf{m}_{emb} \in \mathbb{R}^{3W\times d}$, is formed by concatenating these embedded representations along the temporal dimension.

We employ two Transformer layers with specific modifications to incorporate the insole conditioning effectively. First, a standard self-attention block processes the embedded pose sequence $\mathbf{m}_{emb}$, serving as the query, key, and value. Next, we encode the diffusion timestep t using sinusoidal positional embeddings passed through a two-layer MLP. This time embedding is then transformed by a linear layer specific to each decoder layer. By adding these timestep embeddings to the output of the self-attention block, we explicitly inform the network about the current stage of the diffusion process.

Insole Multi-head Cross-Attention. Similarly to our body-partitioned pose encoding, we also separately process the cross-modality insole information to allow the attention mechanism to focus on specific sensor data based on the motion.

Given the insole data, $\mathbf{c} \in \mathbb{R}^{W\times 25\times 2}$, for the pose sequence, we generate four components per foot. We generate the first two components by partitioning the pressure sensor readings $\mathbf{c_p}$ into two regions per foot: toes and heel (as visualized in Figure 2). Next, we aggregate the IMU data ($\mathbf{c_a}$ and $\mathbf{c_r}$) into a single component. Lastly, we combine the total force c_f and the center of pressure $\mathbf{c_c}$ into the fourth component. Each of these four components per foot is then processed by a separate three-layer MLP to produce insole embeddings, $\{\mathbf{c}^i_{emb} \in \mathbb{R}^{W\times d}\}_{i=1}^{8}$.

To efficiently integrate the insole embeddings into the motion reconstruction process, we employ a multi-head cross-attention block that treats each of the eight insole components (four per foot) as a separate attention head (see Figure 2). This allows the network to selectively attend to different aspects of the insole data based on the reconstructed motion without increasing the required computation by adding additional cross-attention blocks or increasing the attention matrix size. More formally, our multi-head cross-attention mechanism can be expressed as:

$$\text{MultiHead}(\mathbf{m}_{emb}, \{\mathbf{c}^i_{emb}\}_{i=1}^{8}) = \text{Concat}(\{\mathbf{H}^i\}_{i=1}^{8})\mathbf{W_O} \quad (5)$$

$$\mathbf{H}^i = \text{Attention}(\mathbf{m}_{emb}\mathbf{W}^i_{\mathbf{Q}}, \mathbf{c}^i_{emb}\mathbf{W}^i_{\mathbf{K}}, \mathbf{c}^i_{emb}\mathbf{W}^i_{\mathbf{V}}) \quad (6)$$

where $\mathbf{W}^i_{\mathbf{Q}} \in \mathbb{R}^{d\times d/8}$, $\mathbf{W}^i_{\mathbf{K}} \in \mathbb{R}^{d\times d/8}$, $\mathbf{W}^i_{\mathbf{V}} \in \mathbb{R}^{d\times d/8}$, and $\mathbf{W_O} \in \mathbb{R}^{d\times d}$ are learnable parameter matrices, and $\mathbf{m}_{emb}$ is the output of the self-attention block with the added timestep embedding.

Following the cross-attention, a standard feed-forward block (two linear layers with activation and dropout functions) further processes the representation. Finally, after the two Transformer layers, the output $\mathbf{m}_{emb}$ is projected back to the original pose space, yielding the reconstructed 3D pose $\hat{\mathbf{m}}$. The *reverse diffusion* process is trained with the mean absolute error (MAE) loss between the reconstructed $\hat{\mathbf{m}}$ and the original pose sequence $\mathbf{m}$.

By designing the attention mechanism to map specific sensor groups to distinct heads, we introduce a strong inductive bias that aids interpretability and convergence. This separation allows the network to dynamically shift focus based on the gait phase. For example, during a heel strike, the network can attend heavily to the heel head to resolve contact, whereas during the swing phase, it can shift attention to the IMU head to track leg orientation. This prevents the high-dimensionality of the combined signal from diluting the critical cues provided by specific sensors.

Displacement Predictor. In addition to predicting poses, we independently estimate the displacements, $\mathbf{m_d}^{(i)} = (\mathbf{d}^{(i)}, \ldots, \mathbf{d}^{(i+W)})$, for the sequence of poses $\mathbf{m}^{(i)}$. Instead of using a diffusion process, we directly regress the displacements from the IMU data using two standard Transformer layers with self-attention and feed-forward blocks. This choice is motivated by the observation that a given sequence of IMU readings typically corresponds to a specific displacement pattern, making this a suitable regression task. However, for pose reconstruction, where a one-to-many mapping may exist between pressure readings and corresponding poses, the diffusion process proves more suitable for modeling this inherent ambiguity.

We found that including both IMU and pressure data for displacement prediction often leads to overfitting the training data (see Section 4.4). This happens because the displacement predictor network tends to memorize specific pressure patterns, lowering the performance for unseen motion. Consequently, we opted to utilize only IMU data to regress root displacements. However, larger training datasets might benefit from using both IMU and pressure data.

The IMU data, $(\mathbf{c}^{\mathbf{L}}_{\mathbf{a}}, \mathbf{c}^{\mathbf{L}}_{\mathbf{r}}, \mathbf{c}^{\mathbf{R}}_{\mathbf{a}}, \mathbf{c}^{\mathbf{R}}_{\mathbf{r}}) \in \mathbb{R}^{12}$, is first processed by a two-layer MLP to embed it into the Transformer's embedding dimension. Sinusoidal positional encodings are then added, and a final linear layer projects the output back to the displacement space. This approach allows us to leverage the temporal information within the IMU data to predict the root motion of the skeleton accurately. The displacement predictor is trained with the mean squared error (MSE) loss between the reconstructed displacements $\hat{\mathbf{m}}_{\mathbf{d}}$ and the original displacements $\mathbf{m_d}$.

To improve accuracy and account for accumulated errors over time, we incorporate the cumulative sum of displacements within the loss function. The final loss for the displacement predictor is

defined as

$$\mathcal{L} = \text{MSE}(\mathbf{m_d}, \hat{\mathbf{m}}_\mathbf{d}) + \frac{\lambda}{W} \sum_{k=1}^{W} \text{MSE}\left(\sum_{i=1}^{k} \mathbf{d}^{(i)}, \sum_{i=1}^{k} \hat{\mathbf{d}}^{(i)} \right), \quad (7)$$

where $\hat{\mathbf{d}}^{(i)}$ is the predicted displacement and $\lambda = 0.001$ in our experiments. This additional term ensures that the model is penalized for accumulating errors at any point within the temporal window, leading to more accurate displacement predictions throughout the motion sequence.

3.5. Training and Implementation Details

We implemented **Step2Motion** in PyTorch and used the Adam optimizer [KB17] for training. The experiments were conducted on a single NVIDIA GeForce RTX 4090 GPU. Both networks were trained with a batch size of 256 and a learning rate of 10^{-3}. The pose diffusion model was trained for 500 epochs, while the displacement predictor was trained for 200 epochs. For the diffusion model, we followed the standard training procedure by sampling a random diffusion timestep t from a uniform distribution and training the network to predict the denoised motion at the previous timestep, $\mathbf{m}_{t-1}$. Data augmentation was performed by randomly rotating the motion along the world's vertical axis.

We set the Transformer's embedding dimension to 256 and the feedforward block's dimension to 512. The diffusion process utilized 200 diffusion steps (T) with variances β_t increasing linearly from 0.0001 to 0.02. We employed a window size W of 100 poses, sampled at 30 Hz. We use GeLU activation functions similar to previous work [MDH*24].

To reconstruct temporally coherent motion of arbitrary lengths, we implement an autoregressive approach inspired by the diffusion inpainting technique [LDR*22; MDH*24].

4. Experiments and Evaluation

In this section, we comprehensively evaluate our method on a publicly available dataset, employing both quantitative and qualitative assessments. We compare **Step2Motion** with established baseline models and perform ablation studies to provide insight into our design choices. Pose accuracy is analyzed in Section 4.3 and displacement prediction in Section 4.4. Next, we also show the system's capabilities in diverse settings in Section 4.5. Finally, we analyze the use of pressure sensors and IMUs in Section 4.6.

4.1. Datasets

We use two datasets to train and evaluate **Step2Motion**, UnderPressure [MHCH22], and our own recorded locomotion dataset focusing on motion diversity, which we made publicly available under the link: https://vcai.mpi-inf.mpg.de/projects/Step2Motion/

UnderPressure Dataset. Public dataset comprising recordings of nine subjects performing various locomotion tasks such as walking, jogging, or jumping. It provides approximately 5 hours of motion capture data paired with insole readings. However, this dataset has very limited motion variety (normal pace walking and jogging), and subjects were instructed to perform actions in a very specific manner.

Step2Motion Dataset. To assess the performance in scenarios of greater motion diversity, we recorded our dataset. It includes markerless motion capture data paired with insole readings for 8 subjects performing diverse locomotion activities. Participants performed a continuous 25-minute sequence where the type of movement changed approximately every minute based on auditory instructions (e.g., walk, jog, squat). Subjects were explicitly instructed to perform the motion naturally without rigid constraints to ensure diversity. For each type of movement, the participant performs that activity, then transitions between that activity and walking interchangeably, and finally performs the activity stationary for a specified time. Total duration across subjects encompasses approximately 3.6 hours of data.

All motion sequences used in the evaluation and depicted in the figures were not included during training. Furthermore, the testing sets for UnderPressure [MHCH22] and our dataset exclusively utilize motion data from subjects not included in the training sets. For instance, in the UnderPressure dataset, we test on Subject 4, who was intentionally selected due to significantly different body proportions compared to the training set (S1–S9). Subject 4 weighs 65 kg with a height of 167 cm, whereas all training subjects average significantly higher weight and height (e.g., S1 is 91 kg/175 cm; S7 is 88 kg/184 cm).

4.2. Metrics

We use four metrics in our evaluations, three of which to measure pose quality: *Mean Per Joint Positional Error (MPJPE)*, which calculates the mean Euclidean distance between corresponding joints in centimeters; $MPJPE_{\text{Legs}}$, identical to the previous one but considering the four joints of each leg only; and *Mean Per Joint Velocity Error* ($MPJVE_{\text{Legs}}$), which calculates the mean velocity error across the legs joints in centimeters per second. To analyze pose quality, the root is aligned with the ground truth. The fourth metric, *Mean Root Positional Error (MRPE)*, measures the accuracy of the displacement predictor with the mean Euclidean distance error of the root joint in centimeters.

4.3. Analysis of Pose Accuracy

We first performed an ablation study to evaluate the performance of **Step2Motion** in reconstructing general full-body locomotion from insole sensor data. For this evaluation, acceleration is transformed from local to world space using ground truth rotational data, rather than through integration of angular rate. This isolates the performance of our method from the inherent challenges of IMU integration. The quantitative results of both datasets are summarized in Table 1, and qualitative comparisons are provided in Figure 3. Note that we separately train our system in the UnderPressure [MHCH22] database and our dataset.

We benchmark our method against two established architectures—**MLP** and **Transformer**—for regression problems. The Transformer model is a stack of Transformer encoders. These two baselines take a window of insole data as input and directly predict the corresponding window of poses. This comparison allows us to assess the benefits of using a diffusion-based approach over direct regression and to validate the architectural modifications we

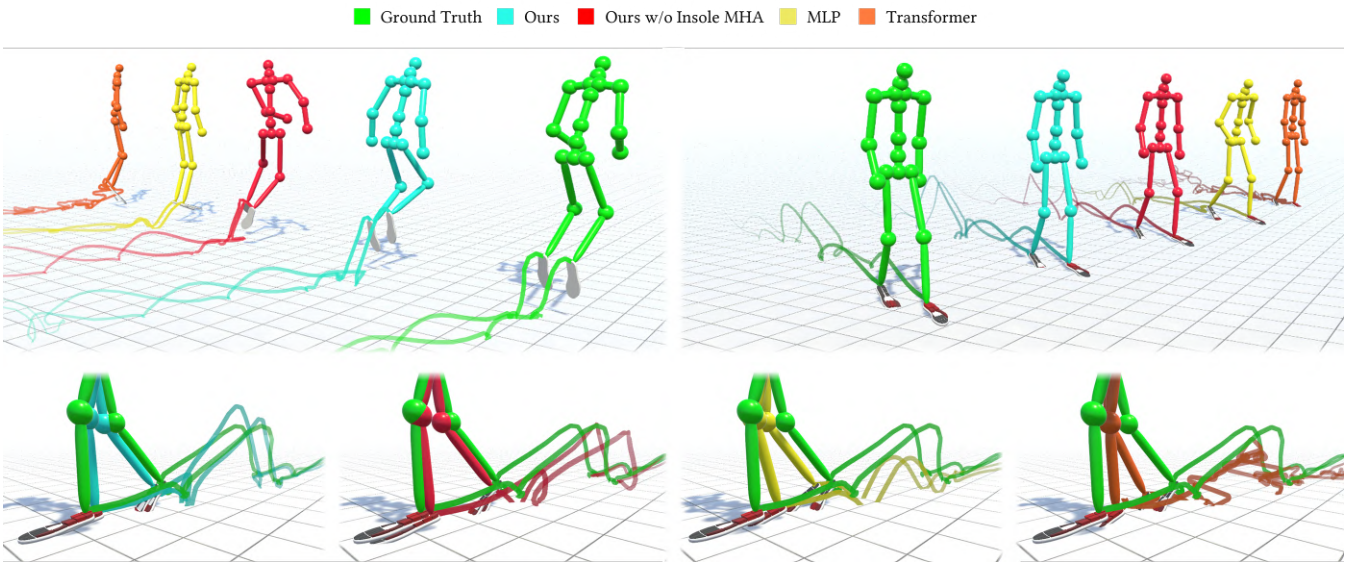

Figure 3: ***Comparison on a jump followed by walking motion sequence.*** *Root motion is aligned to the ground truth (with an offset for visualization) to highlight pose differences. Our full method (cyan) accurately captures the jump trajectory and overall motion, while the* ***Transformer*** *baseline (orange) exhibits significant jitter and the* ***MLP*** *baseline (yellow) produces overly smooth motion. Removing the insole multi-head cross-attention (red) leads to a degradation in accuracy. The bottom row provides close-up views.*

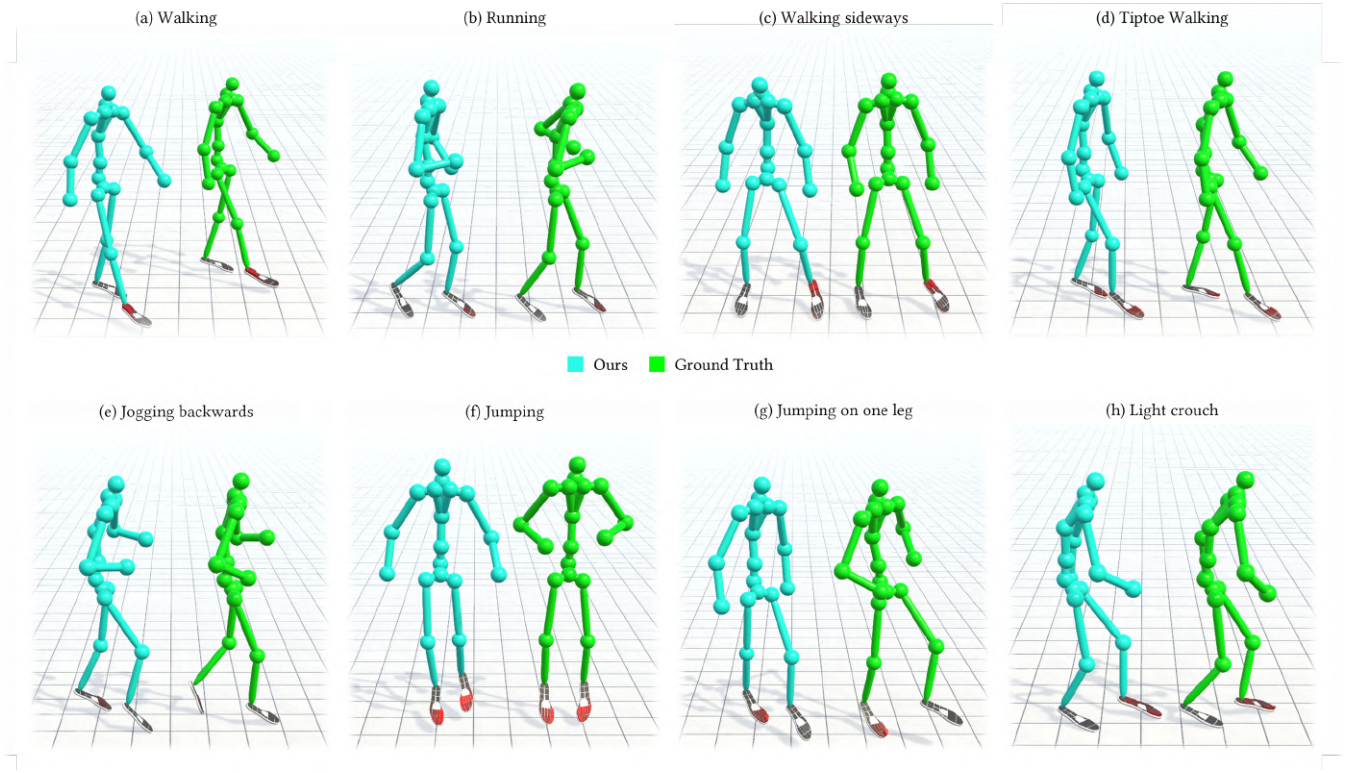

Figure 4: ***Reconstructing locomotion styles.*** *Our method allows the reconstruction of different motion styles only using insole data, from walking and running to crouching, walking sideways, or moving on tiptoes.*

Table 1: ***Pose accuracy evaluation*** *of our method, with and without insole multi-head cross-attention, compared with common deep learning architectures:* ***MLP*** *and* ***Transformer*** *(see Section 4.3). The standard deviation is shown in parentheses. Errors are reported in cm and cm/s.*

Method	UnderPressure Dataset [MHCH22]			Step2Motion Dataset		
	MPJPE ↓	$MPJPE_{\text{Legs}}$ ↓	$MPJVE_{\text{Legs}}$ ↓	*MPJPE* ↓	$MPJPE_{\text{Legs}}$ ↓	$MPJVE_{\text{Legs}}$ ↓
MLP	7.7(7.3)	9.9(9.5)	65.2(73.3)	11.9(11.5)	13.7(12.7)	52.3(61.8)
Transformer	10.7(9.1)	13.5(10.9)	131.1(134.2)	14.2(11.9)	15.0(11.7)	140.2(142.7)
Ours w/o Insole MHA	7.4(7.5)	7.2(8.9)	26.4(30.0)	12.2(11.7)	13.0(11.9)	52.5(61.6)
Ours	**7.2(7.1)**	**6.5(8.2)**	**26.1(29.2)**	**11.4(11.7)**	**12.3(11.9)**	**50.4(62.0)**

introduce for effectively leveraging insole data. Additionally, we perform an ablation study by removing the insole multi-head cross-attention and replacing it with a standard cross-attention module. This setting (**Ours w/o Insole MHA**) effectively evaluates a standard Transformer architecture within our diffusion framework.

To the best of our knowledge, this is the first approach for general locomotion reconstruction from insole sensors. The closest work is SolePoser [WKKK24], which focuses on the specific activity of skiing and does not synthesize arm and root motion. Due to their focus on skiing and the unavailability of their code, we do not include it in our comparison.

Figure 3 visualizes the foot joint trajectories for each baseline, our method, and the mocap ground truth, for a sequence involving a jump followed by walking. Note that the root position is fixed to the ground truth to focus on pose quality. The **Transformer** baseline struggles to capture temporal consistency, producing highly jittery results with unpredictable trajectories, likely due to the limitations of the attention mechanism without the iterative refinement of a diffusion process. The **MLP** baseline reconstructs mostly correct poses but tends towards overly smooth and averaged motions (see Figure 5-top, which shows the same jump but in the peak height), as evident in the reduced amplitude of the jump trajectory. In contrast, our method accurately preserves the jump trajectory, particularly when using our full approach, with curves closely resembling the ground truth. For more qualitative ablations, we also refer to our supplemental video.

These observations are supported by the quantitative results in Table 1. The **Transformer** baseline consistently performs the worst across all metrics. The **MLP** baseline exhibits relatively low *MPJPE* due to its tendency to produce average poses, but performs poorly when specifically analyzing leg movement (positions and velocities), indicating that high-frequency details are lost. Incorporating the insole multi-head cross-attention mechanism consistently improves all metrics, particularly the positional accuracy of the legs.

The insole multi-head cross-attention mechanism is particularly crucial, as each modality should be analyzed differently depending on the action, rather than treating all sensor readings uniformly. In general, while walking, the IMU provides strong motion cues. However, during stationary actions such as squatting, the pressure distribution becomes the primary source of information. Figure 5 highlights these scenarios. In the middle image, the full model correctly captures the squatting motion. This highlights the model's ability to prioritize pressure data from the insole when IMU readings are minimal. Additionally, the image at the bottom shows that when the model cannot separately pay attention to the readings of the inertial sensors, it struggles to understand the direction of movement, which is determined primarily by the IMUs and not the pressure sensors. This leads to wrong global rotations, similar to a key limitation observed in the base **Transformer** model, which similarly uses the standard attention mechanism. This highlights the effectiveness of our approach in capturing fine-grained movements and reconstructing more realistic and detailed locomotion.

4.4. Analysis of Root Displacement Accuracy

In this section, we evaluate the performance of **Step2Motion** in predicting the root position from per-frame displacements. Following the methodology in Section 4.3, we train the model separately on the UnderPressure [MHCH22] dataset and our dataset. Quantitative results are presented in Table 2. Note that our dataset presents larger displacement errors due to the increased motion diversity and the length of the test animation clip (around 16 min).

We propose different baselines and ablation experiments for displacement prediction compared to those used in Section 4.3 for pose reconstruction, as we employ a separate network for this task. We estimate per-frame displacement using **Double Integration** of the IMU acceleration and averaging the results (providing a non-data-driven approximation of root motion) as well as using an **MLP**. The **Combined** baseline predicts the displacement within the diffusion process by adding a displacement term to the body-partitioned pose encoding, eliminating the need for a separate displacement predictor. Finally, we compare these baselines with our method under different settings: using **Only Pressure** to predict the root displacement (i.e., the 16 pressure sensors, the center of pressure, and the total force per foot); **w/o Cumsum Loss**, by removing the second loss term in Equation 7 and making our system less aware of accumulated errors during training; and using both, **Pressure + IMU**, for the displacement predictor. Differently from Section 4.3, the **Transformer** baseline is excluded from the displacement evaluation, as our displacement predictor uses a standard transformer encoder.

Analysis of Table 2 reveals that the double integration of acceleration signals yields poor results. This is likely due to the insole accelerometers producing inaccurate readings during foot-ground contact (due to sudden contact with the ground), leading to incorrect root motion estimation. A higher sampling frequency may help alleviate this issue, as we currently sample accelerations at 30 Hz.

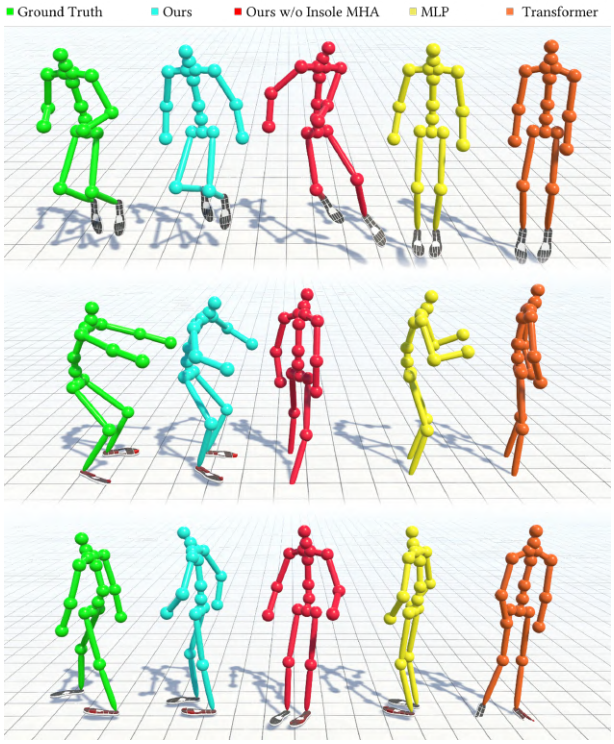

Figure 5: ***Qualitative comparison on various motion sequences.*** *Root motion is aligned to the ground truth (with a slight offset for visualization purposes) to highlight pose quality differences. These animations were not used for training. The top image shows the same jump as Figure 3 but captured at the highest point. The middle and bottom images show squatting and walking motion, respectively. Our full method (cyan) accurately reconstructs various motions, including in-place movements like squatting. Removing the Insole MHA (red) leads to errors, particularly in sequences where distinct sensor information is crucial, such as the squat (relying on pressure) and walking (relying on IMU). The* ***MLP*** *(yellow) and* ***Transformer*** *(orange) baselines exhibit limitations in capturing these motions and overall, smoothed-out results.*

Furthermore, providing only pressure information proves to be insufficient for an accurate prediction. Acceleration and angular rate from IMUs are strong predictors of root motion, as demonstrated in Figure 6, where pressure-only predictions perform well only in scenarios with minimal horizontal root motion (e.g., squatting).

Alternative architectures, such as MLPs or integrating displacement prediction into the diffusion process, also prove ineffective. The former likely struggles to capture temporal dependencies, while the latter may be hindered by the difficulty of combining pose data with displacements. Despite input standardization, the network appears to have difficulty processing these distinct feature types. We can observe in Figure 6 that, for both architectures, the error increases monotonically in most cases.

The rest of the experiments show similar results with minor variations. Removing the cumulative sum term in the displacement loss leads to worse performance, particularly noticeable in Figure 6, where the error exhibits a consistent increase over time. Including both pressure and IMU information yields comparable results to using only IMU data, with the most significant differences observed in our dataset, particularly during idle poses like squatting. We assume that including pressure information may increase the risk of overfitting, while relying solely on IMU data improves generalization. However, this behavior may change with larger training datasets. Overall, our full approach, with or without pressure data, consistently achieves the best results.

Table 2: ***Root displacement accuracy evaluation****. Mean Root Positional Error (MRPE) is reported in meters, with standard deviation in parentheses.*

Method	UnderPressure Dataset ↓ [MHCH22]	Step2Motion Dataset ↓
Double Integration	6.56(6.40)	102.5(76.9)
MLP	1.97(1.19)	12.6(8.15)
Combined	3.14(2.04)	14.7(9.45)
Ours (only Pressure)	4.57(2.65)	10.1(7.06)
Ours w/o Cumsum Loss	1.07(0.69)	26.7(13.6)
Ours (Pressure+IMU)	**0.94(0.56)**	10.1(6.26)
Ours (only IMU)	0.97(0.54)	**6.41(4.03)**

4.5. Analysis of Specific Motion Types

In this section, we examine **Step2Motion** in diverse settings, starting with locomotion reconstruction using our dataset and extending our analysis to more challenging motions, such as dancing, walking on tiptoes, or *in-the-wild* capture. We refer the reader to the companion video for additional qualitative results.

Locomotion. We demonstrate our method on different locomotion tasks. Sections 4.3 and 4.4 provide a full accuracy evaluation. Our method reconstructs from basic locomotion tasks like walking, as visualized in Figure 1 (middle), to more complex movements, such as dancing, shown in Figure 9, thereby demonstrating its ability to predict accurate gait cycles and plausible full-body motion. Additional reconstructions for diverse tasks are shown in Figure 4.

To further validate the relationship between foot pressure and the reconstructed motion, we examine the acceleration curves generated by our method and the baselines compared to the pressure data. Specifically, we selected the pressure sensor on the left insole's toe region. During the walking gait, increased exerted pressure by the toes signals the beginning of the swing phase. This pressure peaks during the toe-off phase, immediately before the increase in toe acceleration. Therefore, we expect to observe an increase in acceleration between the peak pressure and its subsequent decrease. This relationship is demonstrated in Figure 7. A jogging animation is visualized, where our method closely follows the ground truth acceleration, exhibiting distinct peaks in acceleration that align with the release of pressure. In contrast, the **MLP** baseline produces a

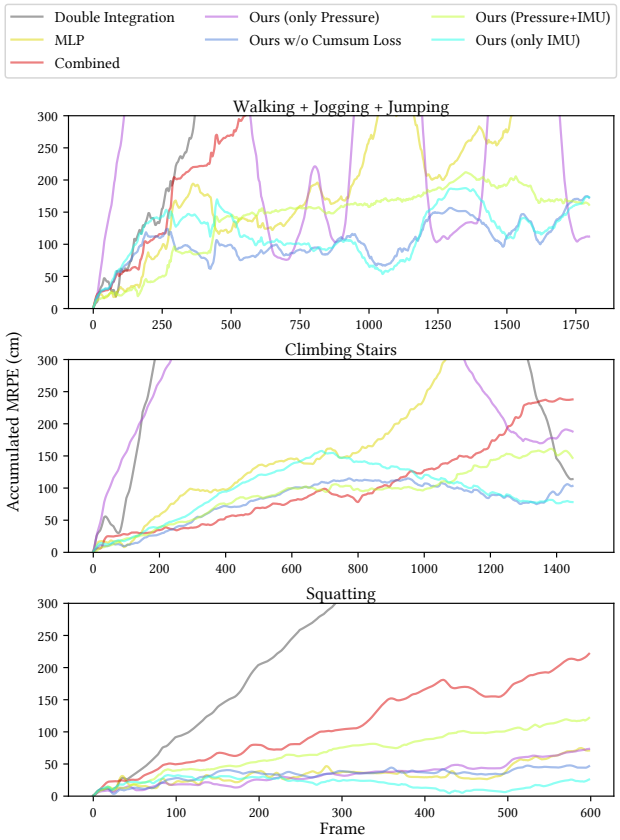

Figure 6: ***Accumulated root positional error over time.*** *The figure demonstrates the limitations of double integration, pressure-only inputs, MLP architectures, and integrating displacement prediction into the diffusion process. Our full method, with or without pressure data, consistently shows the lowest error.*

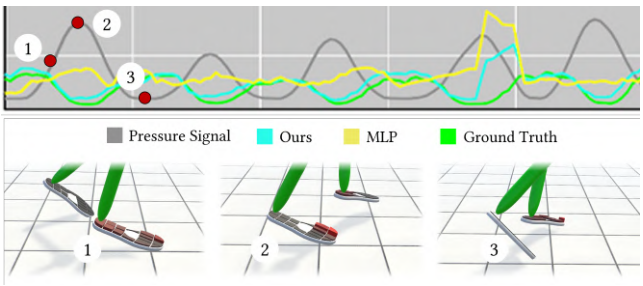

Figure 7: ***Comparison of acceleration against pressure curves for a jogging sequence.*** *Our method (cyan) accurately captures jogging acceleration patterns, with increases aligning with pressure release, unlike the overly smooth* ***MLP*** *baseline (yellow). The* ***Transformer*** *baseline is excluded due to large fluctuations. The visualization includes left toes pressure (white) and ground truth acceleration (green).*

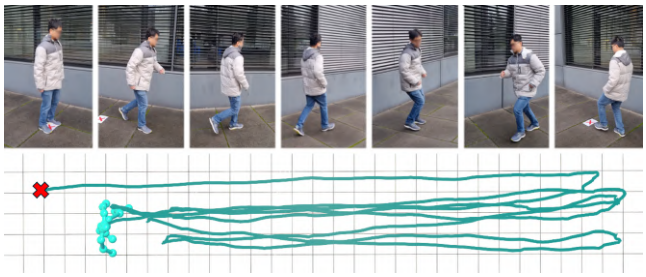

Figure 8: ***Visualization of positional drift in in-the-wild capture.*** *A user jogged repeatedly from the red cross marker to a point 7.5 meters away and back, completing four cycles. The total distance covered was 60 meters. The final drift, indicated by the offset between the red cross and the character's end position, is approximately 0.75 meters (1.25 % of the total distance).*

relatively flat acceleration curve, averaging the predicted acceleration throughout the motion sequence.

In-the-wild capture. To assess the performance of our method in real outdoor settings, we conduct an *in-the-wild* experiment, capturing insole data from a user performing various locomotion tasks while wearing the insole sensors. The motion is then reconstructed using our system trained on our dataset. In this case, linear accelerations from the IMU are transformed to a fixed world frame by integrating the angular rates. Figure 1 (right) shows a reconstructed locomotion sequence, which demonstrates **Step2Motion**'s ability to recover plausible motion sequences in outdoor environments.

We also analyze the displacement accuracy for *in-the-wild* capture. Figure 8 shows the root trajectory of a user repetitively jogging from a marked point to a location 7.5 meters away and back, completing four cycles for a total distance of 60 meters. The final drift of our full approach is approximately 0.75 meters (1.25 % of the total distance). The displacement predictor baselines (see Section 4.4) achieve considerably worse results: **Double Integration** (9.8 meters, 16.33 %), **MLP** (3.85 meters, 6.41 %), **Ours (only Pressure)** (no movement in the whole sequence), **Combined** (4.6 meters, 7.66 %), **Ours w/o Cumsum Loss** (2.45 meters, 4.08 %), and **Ours (Pressure+IMU)** (3.25 meters, 5.41 %).

Figure 9: ***Dancing animation reconstructed by Step2Motion.*** *Trained on dancing motion, it accurately reconstructs the lower-body movements and temporally coherent upper-body motion for unseen motion during training.*

Dancing. To assess **Step2Motion** in a more challenging reconstruction task, we trained our model to predict *dancing* motions. Due to the scarcity of paired insole sensors and motion capture data, we collected and trained our system in a limited dataset of dance motions (approximately 15 minutes), focusing on a specific dance style. We train on 12 minutes of data and use 3 minutes for testing. Figure 9 and the companion video showcase the results. **Step2Motion** effectively reconstructs temporally coherent dance sequences, including natural arm movements that align with the lower-body motion. A quantitative evaluation of these dance data is provided in Section 4.6, analyzing the importance of utilizing both pressure and IMU information for accurate reconstruction. The dancing data used in this analysis will be released together with the **Step2Motion** dataset.

4.6. Influence of Multimodal Insole Data

Section 4.3 focuses on analyzing pose reconstruction accuracy on a publicly available dataset, primarily evaluating locomotion and jumping animations. To further assess the potential of combining pressure and IMU data, we now analyze dance motions (see *Dancing* in Section 4.5), which present a greater challenge and require both data modalities for meaningful pose reconstruction.

Table 3 presents the results of evaluating our full approach, **Ours**, against variations using only IMU data, **Ours (only IMU)**; only pressure data, **Ours (only Pressure)**; and replacing the insole multi-head cross-attention with a standard cross-attention module, **Ours w/o Insole MHA**. The latter effectively removes the specialized mechanism for incorporating insole sensor data into the diffusion process. Our results demonstrate that only the full approach successfully reconstructs dance motions. In the other three scenarios, during training, the validation loss stops decreasing in early epochs, indicating the model's inability to learn a relationship between the sensor readings and the target motion. This results in significantly worse performance when using only pressure or IMU information and when excluding our insole cross-attention.

Table 3: ***Accuracy evaluation of pose reconstruction on dancing motion.*** *This analysis demonstrates the efficacy of* ***Step2Motion*** *in combining pressure and IMU information for reconstructing complex motions like dancing, which are challenging to reconstruct using IMU data alone. Errors are reported in cm and cm/s.*

Method	Dancing		
	MPJPE ↓	$MPJPE_{\text{Legs}}$ ↓	$MPJVE_{\text{Legs}}$ ↓
w/o Insole MHA	31.6(25.5)	34.4(16.2)	307.1(193.4)
Only Pressure	25.7(22.1)	26.1(14.5)	296.9(246.3)
Only IMU	21.3(18.7)	22.7(14.9)	320.9(275.9)
Ours	**8.2(10.5)**	**6.9(6.6)**	**34.2(40.5)**

5. Limitations and Future Work

A main limitation of our method stems from the inherent drift associated with IMU sensors and the limited measurements on body parts far from the feet, i.e., head and arms, affecting the recovery of certain motions like squats. This primarily affects the accuracy of root motion and arm pose prediction, but can also influence the predicted global orientation when significant drift occurs in angular rate measurements. Since insole data primarily dictates lower-body kinematics, upper-body reconstruction is inherently probabilistic. Our model generates *plausible* motion based on learned correlations (e.g., arm swing), rather than deterministic tracking.

We also observe foot sliding, which is common in generative approaches. While our method implicitly learns contact from pressure, it does not explicitly enforce hard constraints. Future work could leverage the physical nature of pressure data to enforce foot-lock constraints during a post-optimization stage.

While using only the IMUs helps prevent overfitting in displacement prediction, more motion data paired with insole readings (ideally incorporating other sensor modalities like egocentric cameras or additional IMUs) could help mitigate this issue. Additionally, generating synthetic insole readings from existing motion sequences could substantially increase the amount of training data, which would require an accurate physics-based model for foot-ground interactions and soft-tissue deformation. Another potential solution involves using motion priors trained on large motion capture databases and fine-tuning them for insole sensors. However, the domain gap is significant, and large research datasets lack sufficient diversity and quality. We found that training a specialized architecture from scratch yielded better alignment with the insoles signal. We leave further exploration for future work, as our initial tests suggested that significant work would be required to successfully use pre-trained models.

Integrating the pressure readings in the attention mechanism is still open for further study. Our approach divides pressure into two key regions: toes and heel, motivated by their significant roles in the push-off and heel strike phases during gait. This could be improved by adopting a finer sole pattern to increase accuracy.

6. Conclusions

This work represents a significant first step towards general locomotion reconstruction only from insole sensors. To the best of our knowledge, **Step2Motion** is the first approach to tackle this challenging problem, demonstrating that insole readings can be excellent, feature-rich human motion descriptors.

Our approach leverages the power of diffusion models, combined with a dedicated displacement prediction network and a carefully designed cross-attention mechanism, to effectively capture the complex relationship between human movement and the different sensors included in the insoles. Through extensive evaluations on a public database and our dataset, we demonstrated **Step2Motion**'s ability to accurately reconstruct lower-body animation while synthesizing plausible upper-body movement.

This work opens exciting possibilities for future research in motion capture and animation. The ability to reconstruct detailed human motion using only insole sensors could greatly benefit applications in sports analysis, rehabilitation, and entertainment, enabling more accessible and versatile motion capture in unconstrained environments.

7. Acknowledgements

This work has received funding from MCIN/AEI/10.13039/501100011033/FEDER, UE (Spain) in the framework of the project PID2021-122136OB-C21, and with the support of the Department of Research and Universities of the Government of Catalonia (2021 SGR 01035). Jose Luis Ponton was also funded by the Spanish Ministry of Universities (FPU21/01927 and EST24/00555).

References

[AAR*24] ALVARADO, EDUARDO, ARGUDO, OSCAR, ROHMER, DAMIEN, et al. "TRAIL: Simulating the Impact of Human Locomotion on Natural Landscapes". *The Visual Computer* (June 2024). DOI: 10.1007/s00371-024-03506-z 3.

[APRC22] ALVARADO, EDUARDO, PALIARD, CHLOÉ, ROHMER, DAMIEN, and CANI, MARIE-PAULE. "Real-Time Locomotion on Soft Grounds With Dynamic Footprints". *Frontiers in Virtual Reality* 3 (Mar. 2022), 3. ISSN: 2673-4192. DOI: 10.3389/FRVIR.2022.801856 3.

[AQJH24] ARMANI, RAYAN, QIAN, CHANGLIN, JIANG, JIAXI, and HOLZ, CHRISTIAN. "Ultra Inertial Poser: Scalable Motion Capture and Tracking from Sparse Inertial Sensors and Ultra-Wideband Ranging". *ACM SIGGRAPH 2024 Conference Papers*. SIGGRAPH '24. Denver, CO, USA: Association for Computing Machinery, 2024. ISBN: 9798400705250. DOI: 10.1145/3641519.3657465 2.

[BSM14] BECKHAM, GEORGE, SUCHOMEL, TIM, and MIZUGUCHI, SATOSHI. "Force plate use in performance monitoring and sport science testing". *New Studies in Athletics* 29.3 (2014), 25–37 2.

[CTR*24] COHAN, SETAREH, TEVET, GUY, REDA, DANIELE, et al. "Flexible Motion In-betweening with Diffusion Models". *ACM SIGGRAPH 2024 Conference Papers*. SIGGRAPH '24. Denver, CO, USA: Association for Computing Machinery, 2024. ISBN: 9798400705250. DOI: 10.1145/3641519.3657414 3.

[DKP*23] DU, YUMING, KIPS, ROBIN, PUMAROLA, ALBERT, et al. "Avatars Grow Legs: Generating Smooth Human Motion from Sparse Tracking Inputs with Diffusion Model". *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, 481–490. DOI: 10.1109/CVPR52729.2023.00054 2, 4.

[DMGT23] DABRAL, RISHABH, MUGHAL, MUHAMMAD HAMZA, GOLYANIK, VLADISLAV, and THEOBALT, CHRISTIAN. "MoFusion: A Framework for Denoising-Diffusion-Based Motion Synthesis". *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, 9760–9770. DOI: 10.1109/CVPR52729.2023.00941 3.

[GMSP21] GUZOV, VLADIMIR, MIR, AYMEN, SATTLER, TORSTEN, and PONS-MOLL, GERARD. "Human POSEitioning System (HPS): 3D Human Pose Estimation and Self-localization in Large Scenes from Body-Mounted Sensors". *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2021, 4316–4327. DOI: 10.1109/CVPR46437.2021.00430 2.

[GSHG11] GRIMM, ROBERT, SUKKAU, JOHANN, HORNEGGER, JOACHIM, and GREINER, GÜNTHER. "Automatic Patient Pose Estimation Using Pressure Sensing Mattresses". *Bildverarbeitung für die Medizin 2011: Algorithmen - Systeme - Anwendungen Proceedings des Workshops vom 20. - 22. März 2011 in Lübeck*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, 409–413. ISBN: 978-3-642-19335-4. DOI: 10.1007/978-3-642-19335-4_84 3.

[HJA20] HO, JONATHAN, JAIN, AJAY, and ABBEEL, PIETER. "Denoising diffusion probabilistic models". *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS '20. Vancouver, BC, Canada: Curran Associates Inc., 2020. ISBN: 9781713829546 3, 4.

[HKA*18] HUANG, YINGHAO, KAUFMANN, MANUEL, AKSAN, EMRE, et al. "Deep Inertial Poser: Learning to Reconstruct Human Pose from Sparse Inertial Measurements in Real Time". *ACM Transactions on Graphics* 37.6 (Dec. 2018), 185:1–185:15. ISSN: 0730-0301. DOI: 10.1145/3272127.3275108 1, 2.

[HLP*24] HAN, ISAAC, LEE, SEOYOUNG, PARK, SANGYEON, et al. "Smart Insole: Predicting 3D human pose from foot pressure". *2nd NeurIPS Workshop on Touch Processing: From Data to Knowledge*. 2024. URL: https://openreview.net/forum?id=DX8C7rAi7O 2, 3.

[HST*23] HAN, XINGJIAN, SENDERLING, BEN, TO, STANLEY, et al. "GroundLink: A Dataset Unifying Human Body Movement and Ground Reaction Dynamics". *SIGGRAPH Asia 2023 Conference Papers*. New York, NY, USA: Association for Computing Machinery, 2023, 1–10. DOI: 10.1145/3610548.3618247 3.

[JI21] JIANG, HAO and ITHAPU, VAMSI KRISHNA. "Egocentric Pose Estimation from Human Vision Span". *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2021, 10986–10994. DOI: 10.1109/ICCV48922.2021.01082 3.

[JSQ*22] JIANG, JIAXI, STRELI, PAUL, QIU, HUAJIAN, et al. "AvatarPoser: Articulated Full-Body Pose Tracking from Sparse Motion Sensing". *Computer Vision – ECCV 2022*. Ed. by AVIDAN, SHAI, BROSTOW, GABRIEL, CISSÉ, MOUSTAPHA, et al. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022, 443–460. ISBN: 978-3-031-20065-6. DOI: 10.1007/978-3-031-20065-6_26 2.

[JYG*22] JIANG, YIFENG, YE, YUTING, GOPINATH, DEEPAK, et al. "Transformer Inertial Poser: Real-time Human Motion Reconstruction from Sparse IMUs with Simultaneous Terrain Generation". *SIGGRAPH Asia 2022 Conference Papers*. SA '22. New York, NY, USA: ACM, Nov. 2022, 1–9. ISBN: 978-1-4503-9470-3. DOI: 10.1145/3550469.3555428 1, 2.

[KB17] KINGMA, DIEDERIK P. and BA, JIMMY. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG]. URL: https://arxiv.org/abs/1412.6980 6.

[KKC23] KIM, JIHOON, KIM, JISEOB, and CHOI, SUNGJOON. "FLAME: free-form language-based motion synthesis & editing". *AAAI'23/IAAI'23/EAAI'23*. AAAI Press, 2023. ISBN: 978-1-57735-880-0. DOI: 10.1609/aaai.v37i7.25996 3.

[LDR*22] LUGMAYR, ANDREAS, DANELLJAN, MARTIN, ROMERO, ANDRES, et al. "RePaint: Inpainting using Denoising Diffusion Probabilistic Models". *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, 11451–11461. DOI: 10.1109/CVPR52688.2022.01117 6.

[LHR*21] LIU, LINGJIE, HABERMANN, MARC, RUDNEV, VIKTOR, et al. "Neural Actor: Neural Free-view Synthesis of Human Actors with Pose Control". *ACM Trans. Graph.(ACM SIGGRAPH Asia)* (2021) 1.

[LJ24] LEE, JIYE and JOO, HANBYUL. "Mocap Everyone Everywhere: Lightweight Motion Capture with Smartwatches and a Head-Mounted Camera". *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, 1091–1100. DOI: 10.1109/CVPR52733.2024.00110 2.

[LLW23] LI, JIAMAN, LIU, C. KAREN, and WU, JIAJUN. "Ego-Body Pose Estimation via Ego-Head Pose Estimation". *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2023, 17142–17151. DOI: 10.1109/CVPR52729.2023.01644 1, 3.

[MDH*24] MUGHAL, MUHAMMAD HAMZA, DABRAL, RISHABH, HABIBIE, IKHSANUL, et al. "ConvoFusion: Multi-Modal Conversational Diffusion for Co-Speech Gesture Synthesis". *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, 1388–1398. DOI: 10.1109/CVPR52733.2024.00138 3, 4, 6.

[MHCH22] MOUROT, LUCAS, HOYET, LUDOVIC, CLERC, FRANÇOIS LE, and HELLIER, PIERRE. "Underpressure: Deep learning for foot contact detection, ground reaction force estimation and footskate cleanup". *Computer Graphics Forum* 41.8 (2022), 195–206 3, 6, 8, 9.

[Opt25] OptiTrack. *OptiTrack Motion Capture Systems*. en. 2025. URL: http://www.optitrack.com (visited on 01/18/2025) 2.

[PPA*25] Ponton, Jose Luis, Pujol, Eduard, Aristidou, Andreas, et al. "DragPoser: Motion Reconstruction from Variable Sparse Tracking Signals via Latent Space Optimization". *Computer Graphics Forum* 44.2 (2025), e70026. DOI: https://doi.org/10.1111/cgf.70026 2.

[PYA*23] Ponton, Jose Luis, Yun, Haoran, Aristidou, Andreas, et al. "SparsePoser: Real-Time Full-Body Motion Reconstruction from Sparse Data". *ACM Trans. Graph.* 43.1 (Oct. 2023). ISSN: 0730-0301. DOI: 10.1145/3625264 2.

[PYAP22] Ponton, Jose Luis, Yun, Haoran, Andujar, Carlos, and Pelechano, Nuria. "Combining Motion Matching and Orientation Prediction to Animate Avatars for Consumer-Grade VR Devices". *Computer Graphics Forum* 41.8 (Sept. 2022), 107–118. ISSN: 1467-8659. DOI: 10.1111/cgf.14628 2.

[RJH08] Ren, Lei, Jones, Richard K., and Howard, David. "Whole Body Inverse Dynamics over a Complete Gait Cycle Based Only on Measured Kinematics". *Journal of Biomechanics* 41.12 (Aug. 2008), 2750–2759. DOI: 10.1016/J.JBIOMECH.2008.06.001 2.

[RRC*16] Rhodin, Helge, Richardt, Christian, Casas, Dan, et al. "EgoCap: Egocentric Marker-Less Motion Capture with Two Fisheye Cameras". *ACM Transactions on Graphics* 35.6 (Dec. 2016), 162:1–162:11. ISSN: 0730-0301. DOI: 10.1145/2980179.2980235 1, 2.

[SGNA24] Santos, Vítor Miguel, Gomes, Beatriz B., Neto, Maria Augusta, and Amaro, Ana Martins. "A Systematic Review of Insole Sensor Technology: Recent Studies and Future Directions". *Applied Sciences* 14.14 (2024). ISSN: 2076-3417. DOI: 10.3390/app14146085 2.

[SHS*24] Shetty, Ashwath, Habermann, Marc, Sun, Guoxing, et al. "Holoported Characters: Real-time Free-viewpoint Rendering of Humans from Sparse RGB Cameras". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, 1206–1215 1.

[SP17] Shahabpoor, Erfan and Pavic, Aleksandar. "Measurement of Walking Ground Reactions in Real-Life Environments: A Systematic Review of Techniques and Technologies". *Sensors* 17.9 (2017). ISSN: 1424-8220. DOI: 10.3390/s17092085 2.

[SPS*11] Shiratori, Takaaki, Park, Hyun Soo, Sigal, Leonid, et al. "Motion Capture from Body-Mounted Cameras". *ACM Transactions on Graphics* 30.4 (July 2011), 1–10. ISSN: 0730-0301, 1557-7368. DOI: 10.1145/2010324.1964926 1, 2.

[SRF*20] Scott, Jesse, Ravichandran, Bharadwaj, Funk, Christopher, et al. "From Image to Stability: Learning Dynamics from Human Pose". *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII*. Glasgow, United Kingdom: Springer-Verlag, 2020, 536–554. ISBN: 978-3-030-58591-4. DOI: 10.1007/978-3-030-58592-1_32 3.

[SSH*24] Starke, Sebastian, Starke, Paul, He, Nicky, et al. "Categorical Codebook Matching for Embodied Character Controllers". *ACM Trans. Graph.* 43.4 (July 2024). ISSN: 0730-0301. DOI: 10.1145/3658209 2.

[SZH*24] Sun, Haowen, Zheng, Ruikun, Huang, Haibin, et al. "LGTM: Local-to-Global Text-Driven Human Motion Diffusion Model". *ACM SIGGRAPH 2024 Conference Papers*. SIGGRAPH '24. Denver, CO, USA: Association for Computing Machinery, 2024. ISBN: 9798400705250. DOI: 10.1145/3641519.3657422 3.

[TAP*23] Tome, Denis, Alldieck, Thiemo, Peluse, Patrick, et al. "SelfPose: 3D Egocentric Pose Estimation From a Headset Mounted Camera". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.6 (June 2023), 6794–6806. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2020.3029700 3.

[TPAB19] Tome, Denis, Peluse, Patrick, Agapito, Lourdes, and Badino, Hernan. "xR-EgoPose: Egocentric 3D Human Pose From an HMD Camera". *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2019, 7727–7737. DOI: 10.1109/ICCV.2019.00782 3.

[TRG*23] Tevet, Guy, Raab, Sigal, Gordon, Brian, et al. "Human Motion Diffusion Model". *The Eleventh International Conference on Learning Representations*. 2023. URL: https://openreview.net/forum?id=SJ1kSyO2jwu 3, 4.

[Vic25] Vicon. *Vicon Motion Systems*. en-US. 2025. URL: https://www.vicon.com (visited on 01/18/2025) 2.

[VLF*24] Van Wouwe, Tom, Lee, Seunghwan, Falisse, Antoine, et al. "DiffusionPoser: Real-Time Human Motion Reconstruction From Arbitrary Sparse Sensors Using Autoregressive Diffusion". *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, 2513–2523. DOI: 10.1109/CVPR52733.2024.00243 3.

[vMRBP17] Von Marcard, Timo, Rosenhahn, Bodo, Black, Michael J., and Pons-Moll, Gerard. "Sparse Inertial Poser: Automatic 3D Human Pose Estimation from Sparse IMUs". *Comput. Graph. Forum* 36.2 (May 2017), 349–360. ISSN: 0167-7055. DOI: 10.1111/cgf.13131 2.

[VSP*17] Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, et al. "Attention is all you need". *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, 6000–6010. ISBN: 9781510860964 4, 5.

[WKKK24] Wu, Erwin, Khirodkar, Rawal, Koike, Hideki, and Kitani, Kris. "SolePoser: Full Body Pose Estimation using a Single Pair of Insole Sensor". *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. UIST '24. Pittsburgh, PA, USA: Association for Computing Machinery, 2024. ISBN: 9798400706288. DOI: 10.1145/3654777.3676418 3, 8.

[WLX*21] Wang, Jian, Liu, Lingjie, Xu, Weipeng, et al. "Estimating Egocentric 3D Human Pose in Global Space". *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2021, 11480–11489. DOI: 10.1109/ICCV48922.2021.01130 3.

[WLX*23] Wang, Jian, Luvizon, Diogo, Xu, Weipeng, et al. "Scene-Aware Egocentric 3D Human Pose Estimation". *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2023, 13031–13040. DOI: 10.1109/CVPR52729.2023.01252 3.

[WWRS14] Wannop, John W., Worobets, Jay T., Ruiz, Rodrigo, and Stefanyshyn, Darren J. "Footwear Traction and Three-Dimensional Kinematics of Level, Downhill, Uphill and Cross-Slope Walking". *Gait & Posture* 40.1 (May 2014), 118–122. ISSN: 0966-6362. DOI: 10.1016/j.gaitpost.2014.03.004 2.

[WWY22] Winkler, Alexander, Won, Jungdam, and Ye, Yuting. "QuestSim: Human Motion Tracking from Sparse Sensors with Simulated Avatars". *SIGGRAPH Asia 2022 Conference Papers*. Daegu Republic of Korea: ACM, Nov. 2022, 1–8. ISBN: 978-1-4503-9470-3. DOI: 10.1145/3550469.3555411 2.

[XCZ*19] Xu, Weipeng, Chatterjee, Avishek, Zollhöfer, Michael, et al. "Mo2Cap2: Real-time Mobile 3D Motion Capture with a Cap-mounted Fisheye Camera". *IEEE Transactions on Visualization and Computer Graphics* 25.5 (May 2019), 2093–2101. ISSN: 1941-0506. DOI: 10.1109/TVCG.2019.2898650 3.

[Xse25] Xsens. *Xsens Motion Capture*. en. 2025. URL: https://www.movella.com/products/motion-capture (visited on 01/18/2025) 2.

[YK18] Yuan, Ye and Kitani, Kris. "3D Ego-Pose Estimation via Imitation Learning". *Computer Vision – ECCV 2018*. Ed. by Ferrari, Vittorio, Hebert, Martial, Sminchisescu, Cristian, and Weiss, Yair. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, 763–778. ISBN: 978-3-030-01270-0. DOI: 10.1007/978-3-030-01270-0_45 3.

[YK19] YUAN, YE and KITANI, KRIS. "Ego-Pose Estimation and Forecasting As Real-Time PD Control". *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2019, 10081–10091. DOI: 10.1109/ICCV.2019.01018 3.

[YKL21] YANG, DONGSEOK, KIM, DOYEON, and LEE, SUNG-HEE. "Lobstr: Real-time lower-body pose prediction from sparse upper-body tracking signals". *Comp. Graph. Forum* 40.2 (2021), 265–275. DOI: 10.1111/cgf.142631 2.

[YKM*24] YANG, DONGSEOK, KANG, JIHO, MA, LINGNI, et al. "DivaTrack: Diverse Bodies and Motions from Acceleration-Enhanced Three-Point Trackers". *Computer Graphics Forum* 43.2 (2024), e15057. DOI: https://doi.org/10.1111/cgf.15057 2.

[YP03] YIN, KANGKANG and PAI, DINESH K. "FootSee: an interactive animation system". *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. SCA '03. San Diego, California: Eurographics Association, 2003, 329–338. ISBN: 1581136595 3.

[YZH*22] YI, XINYU, ZHOU, YUXIAO, HABERMANN, MARC, et al. "Physical Inertial Poser (PIP): Physics-aware Real-time Human Motion Tracking from Sparse Inertial Sensors". *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, June 2022, 13157–13168. ISBN: 978-1-66546-946-3. DOI: 10.1109/CVPR52688.2022.01282 1, 2.

[YZH*23] YI, XINYU, ZHOU, YUXIAO, HABERMANN, MARC, et al. "EgoLocate: Real-time Motion Capture, Localization, and Mapping with Sparse Body-mounted Sensors". *ACM Transactions on Graphics* 42.4 (July 2023), 76:1–76:17. ISSN: 0730-0301. DOI: 10.1145/3592099 1, 2.

[YZX21] YI, XINYU, ZHOU, YUXIAO, and XU, FENG. "TransPose: Real-Time 3D Human Translation and Pose Estimation with Six Inertial Sensors". *ACM Transactions on Graphics* 40.4 (July 2021), 86:1–86:13. ISSN: 0730-0301. DOI: 10.1145/3450626.3459786 1, 2.

[YZX24] YI, XINYU, ZHOU, YUXIAO, and XU, FENG. "Physical Non-inertial Poser (PNP): Modeling Non-inertial Effects in Sparse-inertial Human Motion Capture". *ACM SIGGRAPH 2024 Conference Papers*. SIGGRAPH '24. Denver, CO, USA: Association for Computing Machinery, 2024. ISBN: 9798400705250. DOI: 10.1145/3641519.3657436 2.

[ZBX*24] ZHANG, SIWEI, BHATNAGAR, BHARAT LAL, XU, YUANLU, et al. "RoHM: Robust Human Motion Reconstruction via Diffusion". *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, 14606–14617. DOI: 10.1109/CVPR52733.2024.01384 3.

[ZCP*24] ZHANG, MINGYUAN, CAI, ZHONGANG, PAN, LIANG, et al. "MotionDiffuse: Text-Driven Human Motion Generation With Diffusion Model". *IEEE Trans. Pattern Anal. Mach. Intell.* 46.6 (Jan. 2024), 4115–4128. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2024.3355414 3.

[ZDC*24] ZHOU, WENYANG, DOU, ZHIYANG, CAO, ZEYU, et al. "EMDM: Efficient Motion Diffusion Model for Fast and High-Quality Motion Generation". *ECCV 2024*. Milan, Italy: Springer-Verlag, 2024, 18–38. ISBN: 978-3-031-72626-2. DOI: 10.1007/978-3-031-72627-9_2 5.

[ZLHA24] ZHANG, ZIHAN, LIU, RICHARD, HANOCKA, RANA, and ABERMAN, KFIR. "TEDi: Temporally-Entangled Diffusion for Long-Term Motion Synthesis". *ACM SIGGRAPH 2024 Conference Papers*. SIGGRAPH '24. Denver, CO, USA: Association for Computing Machinery, 2024. ISBN: 9798400705250. DOI: 10.1145/3641519.3657515 3.

[ZYG21] ZHANG, YAHUI, YOU, SHAODI, and GEVERS, THEO. "Automatic Calibration of the Fisheye Camera for Egocentric 3D Human Pose Estimation from a Single Image". *2021 IEEE Winter Conference on Applications of Computer Vision*. WACV. Waikoloa, HI, USA: IEEE, Jan. 2021, 1771–1780. DOI: 10.1109/WACV48630.2021.00181 3.