Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Human Motion Capture from Loose and Sparse Inertial Sensors with Garment-aware Diffusion Models

Andela Ilic , **Jiaxi Jiang** , **Paul Streli** , **Xintong Liu** and **Christian Holz**
Department of Computer Science
ETH Zürich, Switzerland
firstname.lastname@inf.ethz.ch

Abstract

Motion capture using sparse inertial sensors has shown great promise due to its portability and lack of occlusion issues compared to camera-based tracking. Existing approaches typically assume that IMU sensors are tightly attached to the human body. However, this assumption often does not hold in real-world scenarios. In this paper, we present a new task of full-body human pose estimation using sparse, loosely attached IMU sensors. To solve this task, we simulate IMU recordings from an existing garment-aware human motion dataset. We developed transformer-based diffusion models to synthesize loose IMU data and estimate human poses based on this challenging loose IMU data. In addition, we show that incorporating garment-related parameters while training the model on simulated loose data effectively maintains expressiveness and enhances the ability to capture variations introduced by looser or tighter garments. Experiments show that our proposed diffusion methods trained on simulated and synthetic data outperformed the state-of-the-art methods quantitatively and qualitatively, opening up a promising direction for future research.

1 Introduction

Human pose estimation is essential in healthcare, sports, ergonomics, and AR/VR applications. Traditionally, it has been achieved via images or videos [Lan *et al.*, 2023; Sosa and Hogg, 2023; Pavllo *et al.*, 2019; Wang *et al.*, 2020; Zhang *et al.*, 2022a; Artacho and Savakis, 2020; Chen *et al.*, 2021; Liu *et al.*, 2021], AR/VR devices [Zheng *et al.*, 2023; Jiang *et al.*, 2022a; Dai *et al.*, 2024; Millerdurai *et al.*, 2024] and Inertial Measurement Units (IMUs). In recent years, various models have relied on IMUs as a portable and privacy-friendly solution [Yi *et al.*, 2024; De Marchi *et al.*, 2024]. Some notable examples like PIP [Yi *et al.*, 2022], TIP [Jiang *et al.*, 2022b], TransPose [Yi *et al.*, 2021], and DynaIP [Zhang *et al.*, 2024] focus on full-body pose estimation from a minimal number of IMUs. However, they all share a common limitation: the reliance on IMUs that are tightly strapped to the body, thus sacrificing comfort. In contrast, IMUs loosely attached to garments allow individuals to perform everyday activities while wearing regular clothing without feeling restricted or encumbered. In this way, the tracking of human motion in diverse environments, beyond controlled settings, could be facilitated.

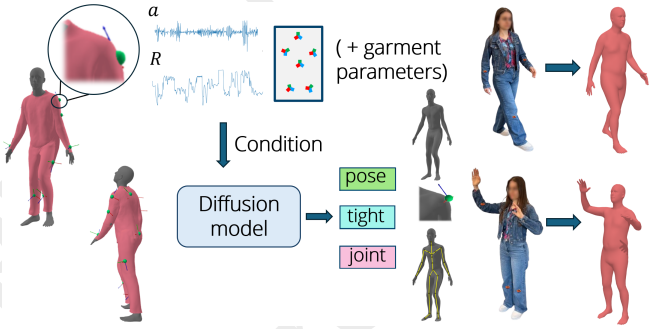

Figure 1: We present a novel task of full-body human pose estimation using sparse IMUs that are loosely attached to garments. To address this challenge, we propose methods for simulating and synthesizing loose IMU signals using diffusion models. Leveraging both simulated and synthetic data, we design diffusion models conditioned on IMU data alone or IMU data and clothing model parameters, enabling robust human motion estimation. We evaluate the effectiveness of our approach across three datasets: a simulated dataset, a real-world upper-body dataset, and a real-world full-body dataset.

Loose Inertial Poser (LIP) [Zuo *et al.*, 2024] explores this setup, focusing on upper-body pose using four IMUs. The Secondary Motion Autoencoder (Semo-AE) they proposed tackles the lack of sufficient real-world training data, as simulated data often fails to capture the full variability seen in actual human movements. However, this model was trained exclusively on recordings from tightly and loosely attached IMUs, disregarding other factors that influence secondary motion, such as the individual's pose or garment-related characteristics. In addition, it does not address full-body tracking.

To overcome these limitations, we propose a real-time method using conditional diffusion models for full- and upper-body pose estimation from loosely worn IMUs. We train our models on simulated, synthetic, and real datasets, using a secondary diffusion model to generate realistic loose-wear data. While moderately loose conditions represented by simulated and synthetic data provide satisfactory performance, training the model on real-world data reveals a problem with limited

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

motion dynamics and convergence of the model towards the mean prediction. This motivated us to shift from the goal of generating realistic-looking synthetic data to incorporating garment characteristics into the model training. We evaluate our models on simulated [Mahmood *et al.*, 2019], real-world upper-body [Zuo *et al.*, 2024], and real-world whole-body [Lorenz *et al.*, 2022] datasets, demonstrating improved generalization and robustness to varying garment looseness.

To summarize, our contributions are as follows:

1. the new task of full-body pose estimation using sparse IMU sensors that are loosely attached. Previous methods assumed tightly attached IMUs or only attempted partial reconstruction, whereas our work tackles a more challenging and realistic scenario.
2. a training strategy that applies three distinct types of loosely attached IMU data: simulated data derived from existing garment-aware human motion datasets, synthetic data generated by our secondary diffusion model, and real-world recordings.
3. a transformer-based diffusion model that estimates full-body poses from loosely attached IMUs. We show that incorporating garment-related parameters significantly enhances pose accuracy, even amid variations in signal noise and differences in garment fit across individuals.
4. several experimental evaluations that show that our diffusion models outperform the state-of-the-art (SOTA) methods. We also provide comprehensive ablation studies to thoroughly analyze the effectiveness of our methods.

2 Related Work

Human pose estimation from wearable sensors. Models like DIP, TransPose, TIP, PIP, and DynaIP focus on real-time full-body pose reconstruction using six IMUs. Researchers have noted that this task is under-constrained [Huang *et al.*, 2018; Mollyn *et al.*, 2023], as many poses can match the same IMU readings. The scarcity of ground-truth data is often addressed by leveraging simulated IMU data derived from mocap datasets. Additionally, modeling temporal dependencies can sometimes impede real-time predictions. TransPose [Yi *et al.*, 2021] highlights signal sparsity and noise, while TIP [Jiang *et al.*, 2022b] introduces Transformers to improve temporal consistency and reduce drift. PIP [Yi *et al.*, 2022] is the first real-time, physics-aware model estimating both pose and forces. DynaIP [Zhang *et al.*, 2024] shifts from synthetic to real IMU data and minimizes the influence of less relevant joints to reduce ambiguity. Beyond standalone IMUs, recent efforts include VR wearables [Jiang *et al.*, 2022a; Jiang *et al.*, 2024b; Jiang *et al.*, 2024a], camera-based systems [Hollidt *et al.*, 2024], and hybrid setups combining IMUs with UWB sensors [Armani *et al.*, 2024; Armani and Holz, 2024].

Pose estimation from loosely attached IMUs. To the best of our knowledge, the only method tailored for loosely worn IMUs is Loose Inertial Poser (LIP) [Zuo *et al.*, 2024]. It introduces a Secondary Motion Autoencoder that treats secondary motion as Gaussian noise, using latent space learning to model variation across body shapes and garment types. A temporal coherence mechanism ensures smooth, realistic motion. Trained on AMASS and evaluated on real upper-body data, LIP achieved $< 20°$ joint rotation error and 10.6 cm position error and outperformed several SOTA models (PIP, TIP, DIP, TransPose) trained on simulated loose IMU data. LIP uses LSTMs, which suit sequential IMU inputs. In contrast, our pipeline leverages modern generative models to better capture uncertainty and complex motion patterns.

Diffusion models for human motions. Diffusion models have recently shown promise for motion generation tasks [Zhang *et al.*, 2022b; Tevet *et al.*, 2022; Tseng *et al.*, 2022; Yuan *et al.*, 2023; Zhou *et al.*, 2025; Kim *et al.*, 2023]. In pose estimation, they have been applied to HMD-based setups [Feng *et al.*, 2024; Du *et al.*, 2023], where sparse upper-body tracking must be used to infer full-body motion. These models perform well under such under-constrained conditions. DiffusionPoser [Wouwe *et al.*, 2024] is the most relevant work here—it uses IMUs and pressure insoles with an unconditional diffusion model that supports flexible sensor configurations and is robust to signal degradation. Unlike DiffusionPoser, which uses tightly worn sensors, our method employs conditional diffusion models trained on both generated and real-world data with loosely attached IMUs.

3 Proposed Method

3.1 Method Overview

We estimate full-body human pose from loosely attached garment-mounted IMUs using acceleration and orientation inputs to predict SMPL joint rotations.

Training data is generated by simulating IMU signals from a garment-aware motion capture dataset. Our approach uses two conditional diffusion models: a primary model for pose estimation and a secondary model for generating realistic loose-wear IMU signals. The secondary model, conditioned on tight-wear IMU data and pose, synthesizes loose-wear signals. These are used to train the primary model, which predicts pose, tight-wear IMU signals, and joint positions. A garment-aware variant further conditions on clothing parameters to enhance realism.

3.2 Diffusion Framework

We adopt a Denoising Diffusion Probabilistic Model (DDPM) [Ho *et al.*, 2020] as the generative backbone. This model operates by progressively corrupting data through a forward process, where Gaussian noise is incrementally added to the data over successive timesteps, forming a series of latent variables z_t. Starting from a clean data sample x_0, the forward diffusion process follows a Markov chain that applies Gaussian noise at each timestep t, defined as:

$$q(z_t|x_0) = \mathcal{N}\left(z_t; \sqrt{\alpha_t}x_0, (1-\alpha_t)\mathbf{I}\right),$$

where α_t is a noise scheduling coefficient that controls the amount of noise added at each step, and $t \in [0,T]$ denotes the progression of time. As t approaches the maximum timestep T, the sample z_T approximates a Gaussian distribution $\mathcal{N}(0,\mathbf{I})$, effectively turning the original data into pure noise.

To recover the original data, a reverse process is learned, which denoises the latents step by step. The reverse diffusion

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

process is modeled as a sequence of conditional probabilities $p_\theta(z_{t-1}|z_t)$, where θ represents the parameters learned by the model. This process is parameterized by a neural network, which predicts either the noise-free data x_0 or the noise component ε_t added during the forward process. For the pose estimation task, we opted to predict the signal itself x_0, as this facilitates a more straightforward application of the objective function.

Figure 2 shows the architecture of the neural network used in all the proposed diffusion models. This is a transformer-based autoencoder with an arbitrary number of transformer blocks in the encoder and decoder networks.

The embedding layer consists of four convolutional layers applied to different parts of the input sequence and the condition, for efficient feature extraction. These parts refer to distinct signals concatenated before being fed into the diffusion model. The use of causal self-attention and residual connections enables effective sequence modeling. To preserve the sequence information, positional encoding is consistently applied before passing the final vector to the first transformer block of the encoder.

3.3 Data Simulation and Generation

We simulate both tightly and loosely attached IMU sensor data for poses from the AMASS dataset. The IMU observations include the acceleration and rotation information from 6 sensors:

$$\text{IMU}(t) = (\{a_1(t), a_2(t), \ldots, a_6(t)\}, \{R_1(t), R_2(t), \ldots, R_6(t)\}).$$

The sensors are positioned on the left and right forearm, sternum, pelvis, and left and right lower leg. For the simulation of clothing effects, we employ TailorNet's [Patel *et al.*, 2020] *Male Shirt* and *Male Pants* models to maintain consistency with the LIP paper. Specifically, we select the SMPL body model [Loper *et al.*, 2015] and utilize the *TallThin* physique and garment style with zero γ parameter. Following LIP's loose data simulation, we select four neighboring vertices from the clothing model to define the IMU orientation through cross-products, while the geometric center of these vertices serves as its position. Similarly, we obtain tight IMU data directly specifying the corresponding vertex on the AMASS body mesh.

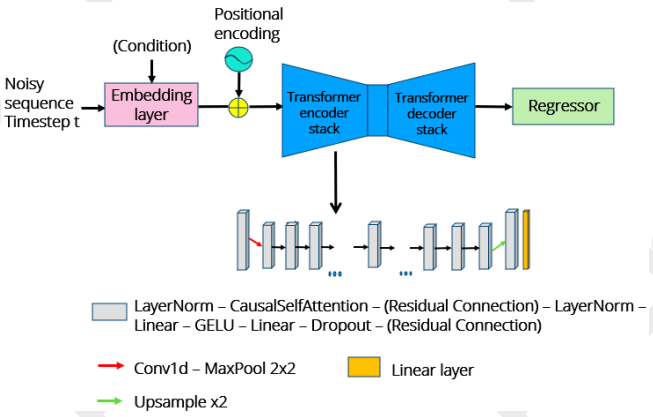

Figure 2: Denoiser architecture used in all diffusion models.

Body Part	Simulated Data [°]	Real-World Data [°]
Left Forearm	21.451	40.890
Right Forearm	21.575	49.545
Sternum	21.094	27.395
Pelvis	12.195	25.224
Left Lower Leg	15.819	35.449
Right Lower Leg	13.512	36.002

Table 1: Orientation difference between tightly and loosely worn IMU sensors' rotations, calculated as the angular difference in degrees. The difference is computed by converting the rotations to rotation matrices, finding the offset using matrix multiplication, and then applying the axis-angle representation to determine the angle.

The orientation difference between the simulated tight and loose data is significantly smaller than that found in real-world data provided by Lorenz et al. [2022], as shown in Table 1. This discrepancy effectively represents tighter-fitting garments, making the simulation less reflective of real-world conditions. To align our dataset better with real-world scenarios, we developed a diffusion model (Figure 3) trained on real-world data provided by Lorenz et al. [2022]. This model generates loose-wear IMU data c_l conditioned on tight-wear IMU data c_t and pose information θ represented as a quaternion. While the LIP model generates synthetic loose data by introducing noise into the tight data, we believe the pose provides essential secondary motion context. We included pose as the condition, as it directly influences how garments move and interact with the body. The objective function is the simple L1 reconstruction loss:

$$\mathscr{L}_{L1} = \|c_l - \hat{c}_l\|_1 = \sum_{i=1}^{N} |c_l^i - \hat{c}_l^i|. \tag{1}$$

Despite the improved offset, we observed that the generated data lacked sufficient diversity, often reflecting scenarios with very wide garment configurations. Therefore, we introduced a parameter $\alpha \sim \text{Uniform}(0,1)$, which modulates the contribution of the simulated and generated data. By varying α, we can adjust the tightness or looseness of the clothing representation, effectively blending more or less of the simulated or generated signal. The final training data is obtained using the following formula:

$$c_i = \alpha c_s + (1-\alpha) c_l, \tag{2}$$

where c_i represents the interpolated synthetic loose data, c_s is the simulated loose data, and c_l is the loose data generated by our diffusion model.

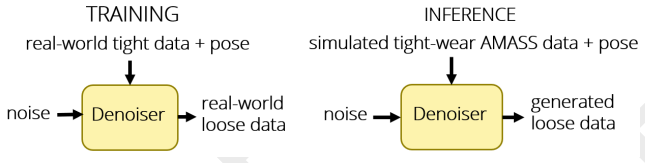

Figure 3: Data generation scheme.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

3.4 Conditional Diffusion Models for Pose Estimation

To estimate the pose based on the observations from loosely attached sensors, we developed a conditional diffusion model (Figure 4). These sensor observations serve as conditions throughout the diffusion process, enabling the model to generate not only pose estimates but also predictions of tight IMU data and joint positions. Given the conditional nature of this model, where predictions fully rely on the provided input, a flexible number of available sensors cannot be accommodated. Therefore, we propose two distinct models—one for the upper-body pose estimation and another for the full-body pose estimation.

The upper-body model predicts the positions and rotations of 14 upper-body joints, including leaf joints, while the whole-body model predicts all 24 joints, as defined by the SMPL body model [Loper *et al.*, 2015]. The loss function for the upper-body model is defined as

$$\begin{aligned}\mathscr{L} = \alpha_{r_r}\mathscr{L}_{r_r} + \alpha_{r_j}\mathscr{L}_{r_j} + \alpha_{p_a}\mathscr{L}_{p_a} \\ + \alpha_{p_j}\mathscr{L}_{p_j} + \alpha_t\mathscr{L}_t + \alpha_c\mathscr{L}_c.\end{aligned} \tag{3}$$

Here, $\mathscr{L}_{r_r}$ and $\mathscr{L}_{r_j}$ represent the reconstruction errors of the root joint rotation and the remaining joints' rotations, respectively. The terms $\mathscr{L}_{p_a}$ and $\mathscr{L}_{p_j}$ capture the position errors for the arms and all other joints, respectively, while $\mathscr{L}_t$ corresponds to the reconstruction error of the tightly attached IMU observations. Finally, $\mathscr{L}_c$ denotes the consistency regularization, calculated as the difference between the model's predictions when conditioned on the original sensor data and a noisy version of the data, where Gaussian noise scaled by 0.3 is added. The coefficients are set as follows: $\alpha_{r_r} = 2$, $\alpha_{r_j} = 1$, $\alpha_{p_a} = 2$, $\alpha_{p_j} = 1$, $\alpha_t = 1$, and $\alpha_c = 3$.

The loss function for the whole-body model follows the same structure, with $\mathscr{L}_{p_{al}}$ (position error of arms and legs) replacing $\mathscr{L}_{p_a}$ (position error of arms) to account for the position error of all extremities. Similarly, $\mathscr{L}_{r_j}$ now incorporates the rotation estimation errors for the lower body as well. All loss functions are based on L1 loss. Additionally, we assign higher weights to the root joint's rotation and the positions of the extremities, as these proved to be particularly challenging for accurate estimation.

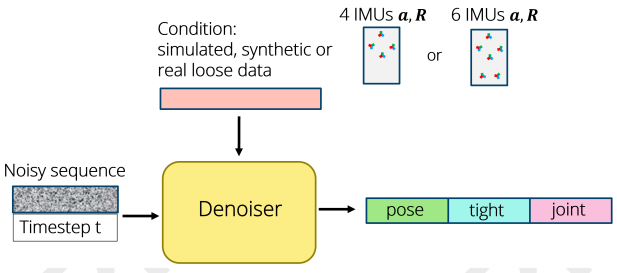

Figure 4: Conditional diffusion model.

3.5 Garment-Aware Conditional Model

The conditional model trained on real-world whole-body IMU data revealed issues with very loose garment configurations, particularly with the root joint sensor, which is placed on top of the jacket and over the pants, leading to highly noisy recordings. Similarly, sensors placed on the extremities often provide misleading observations due to garment movement. To mitigate this intrinsic uncertainty, we try to enhance the conditional model by integrating garment-related information. To achieve this, we utilize seven simulations generated using TailorNet's *MaleShirt* model. Six of these simulations correspond to a *TallThin* physique, denoted as [180, 22], where 180 represents height in centimeters and 22 refers to the body mass index (BMI). These simulations encompass six garment style variations, with γ values of 0, 5, 10, 15, 20, and 24. The final simulation represents a *ShortFat* physique, denoted as [160, 30], with a γ value of 0. We focus on four upper-body IMU sensors for these simulations.

The training scheme remains consistent with that shown in Figure 4, except for the condition dimension, which now expands to 39 (comprising 36 loose sensor values, γ, height, and BMI). As a result, the training dataset is expanded by a factor of seven, yielding approximately 47 million samples.

3.6 Real-Time Inference

We aim to achieve real-time inference and seamless predictions, which is particularly challenging when working with diffusion models. On one hand, these models require time to denoise the input; on the other hand, they do not keep track of previous estimations. As a result, motion predictions are disconnected, leading to significant jitter. For all these reasons, we opted for the frame-by-frame prediction strategy. We use a sliding window of length N (equal to the training sequence length) and shift it by one frame at a time, using the last frame's prediction as the estimate for the current timestep. To enhance temporal consistency and reduce jitter, we incorporate a masking strategy during inference, as proposed by Van Wouwe et al. [2024], which suggests fixing past predictions to produce smoother, more coherent movements.

4 Experiments

4.1 Datasets

AMASS dataset. AMASS is a large-scale motion capture dataset with more than 40 hours of diverse motions. Specifically, we use the 10 subsets (*BMLrub*, *CMU*, *Eyes-Japan*, *HumanEva*, *MPIHDM05*, *MPIMoSh*, *SFU*, *SSM*, *Transitions_mocap*, *BMLmovi*) for training, and the rest 5 subsets *TotalCapture*, *ACCAD*, *TCD_handMocap*, *BMLhandball*, and *MPI_Limits*, for the evaluation. These five datasets ensure diversity in motion capture, ranging from everyday activities in *ACCAD* to extreme poses in *MPI_Limits*, from detailed hand motions in *TotalCapture* and *TCD_handMocap* to dynamic sports data in *BMLhandball*.

LIP dataset. The LIP dataset, provided by the authors of the LIP paper [Zuo *et al.*, 2024], contains IMU signals recorded from four sensors placed on the upper body: left forearm, right forearm, back, and waist. The dataset includes two wearing styles: zipped and unzipped, and participants performed

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

walking, running, jumping, boxing and ping-pong, along with five sessions of free-form movement. In total, the dataset comprises 212,496 frames recorded at 30 frames per second.

Whole-body dataset. Since the LIP dataset is limited to upper-body motions, we utilized a whole-body dataset provided by Lorenz et al. [2022], with 12 IMUs placed on loose-fitting clothing to capture whole-body movements. Ten participants performed various upper and lower body movements and a 15-minute activity sequence. The pelvis sensor was attached to the jacket, not the trousers, and the work suit sleeves were unbuttoned. The dataset includes 1,137,418 frames recorded at 60 fps. We utilize only the rotation data from this dataset, since global acceleration measurements are not available. To obtain the SMPL parameters, we apply inverse kinematics to the global joint orientations provided in the dataset and then map with the corresponding joint names.

4.2 Evaluation Metrics

Following the evaluation protocol of LIP [Zuo *et al.*, 2024], we compute the angular error, which is defined as the Mean Per Joint Rotation Error (MPJRE) in degrees. We also calculate the position error as the Mean Per Joint Position Error (MPJPE) in centimeters. To measure the smoothness of the predicted motions, we include two additional metrics: the Mean Per Joint Velocity Error (MPJVE) in cm/s, and Jitter, which quantifies the changes in acceleration over time, expressed in 10^2 m/s^3. For the upper-body evaluation, we average results across 11 joints, while the whole-body evaluation considers all 24 SMPL model joints.

4.3 Evaluation Results

In this section, we compare our proposed methods and the state-of-the-art pose estimation methods using loosely attached IMUs. Note that LIP has previously demonstrated superior performance over methods designed for tightly-attached IMUs [Mollyn *et al.*, 2023; Yi *et al.*, 2022; Huang *et al.*, 2018; Yi *et al.*, 2021; Jiang *et al.*, 2022b].

Table 2 presents the evaluation metrics for the best-performing models across all three datasets. The terms *Sim*, *Syn*, and *Real* in our model names indicate whether the model was trained on simulated, synthetic, or real loose data.

Table 3 compares models trained and tested on simulated AMASS data. We evaluate methods designed for loosely attached sensors (Ours, LIP) against those developed for tight-fitting setups (IMUPoser [Mollyn *et al.*, 2023], PIP [Yi *et al.*, 2022]) under varying levels of sensor dropout.

Figures 5, 6 and 7 illustrate the qualitative performance of our conditional models, including the upper-body and whole-body garment-unaware models, as well as the upper-body garment-aware model.

Based on the results presented in Table 2, we observe that models trained on simulated and synthetic data perform exceptionally well on the five AMASS datasets. However, their performance gradually declines on the LIP dataset, culminating in significantly reduced accuracy on the whole-body real-world dataset. In contrast, models trained on real-world data achieve good performance on the whole-body dataset but struggle with the AMASS and LIP datasets, likely due to the tighter fit conditions. This indicates that all diffusion models and both LIP models perform significantly better when the distribution of test data closely matches that of the training data, revealing a problem with adaptation to different garment conditions or body shapes.

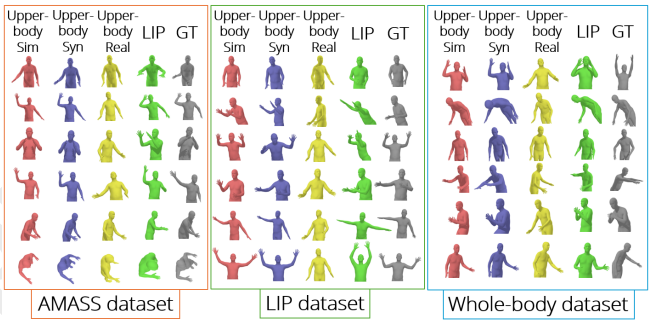

Figure 5: Qualitative performance of the upper-body models on all three evaluation datasets.

Figure 6: Qualitative performance of the whole-body models on the simulated and real-world datasets.

In very challenging scenarios with noisy IMU observations, the models trained on real-world data (Figures 5 and 6) tend to converge towards the mean predictions (neutral pose with legs and arms in relaxed positions). This raises the question of whether the data augmentation approach and emphasis on realistic data are optimal, given that noise and ambiguity in the recorded signals not only mislead the model but also reduce motion dynamics.

We found that incorporating garment-related parameters while training the model on simulated loose data is an effective way to maintain the expressiveness of motion through less noisy observations and still account for potential uncertainties with the addition of these three values.

While we introduced height and BMI to determine whether a *TallThin* or *ShortFat* physique was used for the clothing simulation, it remains to be seen whether these parameters in real-world scenarios pertain more to garment-related character-

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Datasets	Metrics	Upper-body Sim (Ours)	Upper-body Syn (Ours)	Upper-body LIP	Whole-body Sim (Ours)	Whole-body Syn (Ours)	Whole-body LIP	Upper-body Real (Ours)	Whole-body Real (Ours)
AMASS	MPJRE	**14.46 ± 6.00**	15.90 ± 6.75	15.05 ± 5.81	**13.65 ± 4.47**	14.61 ± 4.72	13.67 ± 4.47	27.52 ± 6.89	23.28 ± 4.94
	MPJPE	**9.58 ± 6.33**	12.12 ± 9.50	10.58 ± 6.46	**10.36 ± 5.32**	12.54 ± 6.38	12.78 ± 8.76	15.40 ± 9.57	18.16 ± 9.32
	MPJVE	**26.66**	28.18	27.65	**27.54**	30.64	31.40	31.35	32.89
	Jitter	63.03	54.65	95.83	55.76	47.00	93.37	44.02	34.34
	GT Jitter	137.07	137.07	137.07	140.76	140.76	140.76	137.07	140.76
LIP	MPJRE	20.68 ± 5.66	**19.93 ± 5.68**	20.85 ± 6.36	—	—	—	28.31 ± 5.99	—
	MPJPE	12.17 ± 6.01	**10.65 ± 5.86**	11.49 ± 6.70	—	—	—	16.50 ± 7.42	—
	MPJVE	32.88	**32.68**	41.99	—	—	—	35.26	—
	Jitter	112.51	88.40	233.40	—	—	—	61.89	—
	GT Jitter	33.41	33.41	33.41	—	—	—	33.41	—
Whole Body	MPJRE	31.21 ± 7.52	**30.27 ± 8.21**	33.18 ± 8.84	**24.72 ± 5.70**	25.16 ± 5.95	25.97 ± 7.02	—	—
	MPJPE	28.80 ± 16.02	**25.21 ± 14.10**	26.39 ± 15.59	28.65 ± 14.47	28.60 ± 14.37	**26.23 ± 14.5**	—	—
	MPJVE	22.68	**21.12**	23.97	**26.97**	28.10	30.48	—	—
	Jitter	23.61	15.06	23.03	34.80	30.32	50.59	—	—
	GT Jitter	43.97	43.97	43.97	54.70	54.70	54.70	—	—

Table 2: Numerical comparisons between our methods and the state-of-the-art method LIP. We show results for the upper-body as well as the whole-body settings. We also provide the performance of the models trained only on the real-world data to show the benefit of our data simulation and synthesis strategies.

Number of missing sensors	Metrics	AMASS				Whole-body real-world dataset	
		Whole-body Sim (Ours)	Whole-body LIP	IMUPoser	PIP	IMUPoser	PIP
0	MPJRE	13.65 ± 4.47	13.67 ± 4.47	12.08 ± 4.77	17.38 ± 5.03	38.86 ± 5.30	43.43 ± 7.27
	MPJPE	10.36 ± 5.32	12.78 ± 8.76	7.37 ± 4.43	9.62 ± 4.84	44.03 ± 8.77	44.62 ± 9.53
1	MPJRE	14.02 ± 4.58	15.41 ± 4.86	12.11 ± 4.86	20.47 ± 4.35	–	–
	MPJPE	10.71 ± 5.47	14.78 ± 8.84	7.54 ± 4.49	23.84 ± 8.51	–	–
2	MPJRE	14.56 ± 4.47	17.14 ± 4.89	12.75 ± 5.04	24.84 ± 4.39	–	–
	MPJPE	11.76 ± 5.49	17.26 ± 9.06	8.52 ± 4.76	37.82 ± 9.91	–	–
3	MPJRE	17.13 ± 4.34	19.71 ± 4.62	16.05 ± 5.99	30.42 ± 4.56	–	–
	MPJPE	16.48 ± 6.58	21.31 ± 9.19	12.26 ± 5.81	47.14 ± 9.82	–	–

Table 3: Mean per-joint rotation and position errors on AMASS and a real-world whole-body dataset with varying numbers of missing sensors. Loose-fitting models (Ours, LIP) are compared to tight-fitting ones (IMUPoser, PIP).

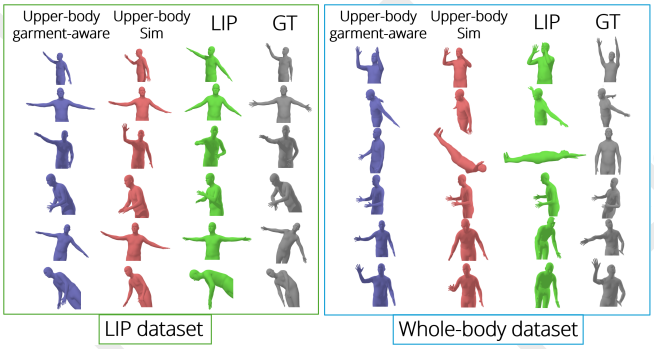

Figure 7: Qualitative performance of the garment-aware conditional upper-body model on the real-world datasets.

istics or the participant's body shape. Since we lacked detailed information about the participants' physiques, we were unable to assess this aspect or evaluate each participant's performance across a long sequence and varying garment styles. However, our experiments on shorter motion sequences, using the optimal selection of the three parameters, demonstrate an improved quantitative performance, as well as a clear qualitative advantage. This is illustrated in Figure 7. Both suitable and unsuitable garment-related information can impact pose estimation accuracy, potentially improving or degrading it, with joint angle errors fluctuating by ±5 degrees and joint position errors by up to ±10 cm (compared to the garment-unaware model). Finally, the selection of the three parameters began by choosing values within the training range—height between 160 and 180 cm, BMI between 22 and 30, and the γ parameter taking one of six predefined values. When the parameters were assigned unrealistic real-world height and BMI values, the motion became more restricted and less dynamic, resembling the behavior of the model trained on real-world data with loosely fitting garments (Upper-body Real).

Experiments with missing sensors (Table 3) show that our model handles such conditions better than others. IMUPoser's performance reflects conservative outputs with limited motion, yielding low error but reduced realism. It also exhibits noticeably more jitter, even when all sensors are available. PIP, on the other hand, performs well only when all sensors are present; its predictions degrade significantly with each additional missing sensor. All models exhibit increased jitter as the number of missing sensors grows[1]. The last two columns highlight the superior generalization ability of our models in real-world loose scenarios, where PIP and IMUPoser perform worse on Lorenz et al. [2022] dataset.

4.4 Ablation Study

Consistency regularization loss. We attribute our model's generalization ability and reduced jitter to the inclusion of this

[1]Visual results are available at https://drive.google.com/file/d/1QHEm2AaGohlv22clJdT9ElAMEb5aAtVc/view?usp=sharing.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Dataset	Metrics	Upper-body Sim	Whole-body Sim
AMASS dataset	MPJRE	12.956 ± 5.674	12.693 ± 4.573
	MPJPE	7.486 ± 5.255	8.130 ± 4.620
	MPJVE	24.445	25.150
	Jitter	130.772	106.773
	GT Jitter	137.073	140.756
Real-world upper-body dataset	MPJRE	21.701 ± 6.390	—
	MPJPE	13.214 ± 7.645	—
	MPJVE	43.738	—
	Jitter	289.034	—
	GT Jitter	33.412	—
Real-world whole-body dataset	MPJRE	32.398 ± 7.362	25.758 ± 6.074
	MPJPE	27.246 ± 15.289	28.647 ± 14.420
	MPJVE	23.978	29.276
	Jitter	50.890	72.931
	GT Jitter	43.967	54.698

Table 4: Performance metrics of the conditional models trained on simulated data without $\mathscr{L}_c$.

loss function. It compels the model to generate similar outputs for slightly varied and noisier versions of the observations, thereby enhancing its robustness. This is demonstrated by the degraded performance on real-world datasets and the enhanced performance on simulated datasets—which share the same characteristics as the training dataset—when the loss is omitted from the objective function. Table 4 presents the performance of the upper- and whole-body models trained on simulated data without this loss function.

Importance of conditioning on all available signals. We focused on conditional models that fully rely on provided IMU recordings, a dependency often perceived as their shortcoming. In contrast, an unconditional model can handle missing data, making it particularly useful in scenarios where unreliable signals are received from the root sensor and extremities. This approach allows for estimating ambiguous signals rather than conditioning predictions on noisy observations. Therefore, we trained an unconditional model to explore this potential improvement. During training, we adopted an inpainting method by masking the data from the root sensor, as the root joint estimations of the conditional models exhibited the largest angle errors. This approach is illustrated in Figure 8.

We generate a mask indicating for each element whether the loose-wear IMU observation is provided (0) or needs to be predicted by the model (1). Following the inpainting approach proposed by Cherel et al. [2024], the model input consists of the masked ground truth data x_{masked} (Equation 4), the noisy input, and the generated mask, concatenated along the sequence dimension. Temporal information t is also included in the input to the denoiser. The model's output is a comprehensive vector that includes pose estimates, loose-wear IMU data, tight-wear IMU data, and joint positions.

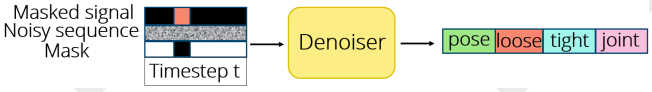

Figure 8: Unconditional inpainting-based diffusion model.

Metrics	Conditional	Unconditional
MPJRE	24.722 ± 5.701	25.933 ± 5.755
MPJPE	28.648 ± 14.47	31.021 ± 11.437
MPJVE	26.974	35.223
Jitter	34.800	150.606
GT Jitter	54.698	54.698
Root angle error	**82.123**	**103.849**

Table 5: Comparison between the conditional whole-body model (trained/evaluated on all six sensors) and an inpainting-based unconditional diffusion model (trained/evaluated with missing root sensor).

$$\mathbf{x}_{\text{masked}} = (1 - \text{mask})\mathbf{x}_0. \tag{4}$$

The objective function mirrors that of the conditional whole-body model, with the addition of the L1 loss term, $\mathscr{L}_{\text{loose\_recon}}$, to account for errors in the reconstruction of the missing root sensor's data. While including all observations led to similar performance in both the conditional and unconditional models, the results in Table 5 suggest that completely excluding the root sensor's data from all frames during evaluation is not advisable, as this data is not consistently unreliable. The presence of reliable observations is likely a contributing factor to the better performance of the model when all data is available. By omitting certain data entirely, the model is granted the freedom to predict the pose of the affected joint by considering other sensors' data that could be misleading itself (such as data obtained from the extremities), which can lead to inaccurate estimates, particularly pronounced when root rotation is minimal.

Pose, tightly attached IMU data, and joint positions in output vector. We adopted the training scheme presented in Figure 4, which involves estimating not only the pose but also data from tightly attached sensors and joint positions. Training the model solely for pose estimation led to abrupt motions. Unlike the loss function in Equation 3, which includes six distinct losses, the pose-only model required additional loss components to reduce jitter. These included velocity, acceleration, and jerk losses over one, three, and five frames, as well as velocity losses for joint positions over the same intervals.

5 Conclusion

In this paper, we introduce a new task of full-body pose estimation from sparse and loosely attached IMU sensor data. Our approach builds on a multi-source training strategy that leverages simulated, synthetic, and real-world IMU data. We simulate IMU observations using garment-aware motion datasets and develop a diffusion model to synthesize realistic loose-wear IMU signals. This framework establishes a foundation for a comprehensive comparison of model performance across different data sources, allowing us to systematically evaluate the strengths and limitations of each. In addition, we demonstrate that incorporating garment-related parameters during training on simulated loose data enhances model expressiveness and helps address the inherent ambiguities of loosely attached IMU sensors. Experiments show our diffusion models outperform the state-of-the-art methods, providing a new direction for future research.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

References

[Armani and Holz, 2024] Rayan Armani and Christian Holz. Accurately tracking relative positions of moving trackers based on uwb ranging and inertial sensing without anchors. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12515–12521, 2024.

[Armani *et al.*, 2024] Rayan Armani, Changlin Qian, Jiaxi Jiang, and Christian Holz. Ultra inertial poser: Scalable motion capture and tracking from sparse inertial sensors and ultra-wideband ranging. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.

[Artacho and Savakis, 2020] Bruno Artacho and Andreas Savakis. Unipose: Unified human pose estimation in single images and videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7035–7044, 2020.

[Chen *et al.*, 2021] Tianlang Chen, Chen Fang, Xiaohui Shen, Yiheng Zhu, Zhili Chen, and Jiebo Luo. Anatomy-aware 3d human pose estimation with bone-based pose decomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):198–209, 2021.

[Cherel *et al.*, 2024] Nicolas Cherel, Andrés Almansa, Yann Gousseau, and Alasdair Newson. Diffusion-based image inpainting with internal learning, 2024.

[Dai *et al.*, 2024] Peng Dai, Yang Zhang, Tao Liu, Zhen Fan, Tianyuan Du, Zhuo Su, Xiaozheng Zheng, and Zeming Li. 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*, pages 874–884, 2024.

[De Marchi *et al.*, 2024] Mirco De Marchi, Cristian Turetta, Graziano Pravadelli, and Nicola Bombieri. Combining 3d human pose estimation and imu sensors for human identification and tracking in multi-person environments. *IEEE Sensors Letters*, 2024.

[Du *et al.*, 2023] Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, and Artsiom Sanakoyeu. Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model, 2023.

[Feng *et al.*, 2024] Han Feng, Wenchao Ma, Quankai Gao, Xianwei Zheng, Nan Xue, and Huijuan Xu. Stratified avatar generation from sparse observations, 2024.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[Hollidt *et al.*, 2024] Dominik Hollidt, Paul Streli, Jiaxi Jiang, Yasaman Haghighi, Changlin Qian, Xintong Liu, and Christian Holz. Egosim: An egocentric multi-view simulator and real dataset for body-worn cameras during motion and activity. *Advances in Neural Information Processing Systems*, 37:106607–106627, 2024.

[Huang *et al.*, 2018] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J. Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time, 2018.

[Jiang *et al.*, 2022a] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *European conference on computer vision*, pages 443–460. Springer, 2022.

[Jiang *et al.*, 2022b] Yifeng Jiang, Yuting Ye, Deepak Gopinath, Jungdam Won, Alexander W. Winkler, and C. Karen Liu. Transformer inertial poser: Real-time human motion reconstruction from sparse imus with simultaneous terrain generation. In *SIGGRAPH Asia 2022 Conference Papers*, SA '22. ACM, November 2022.

[Jiang *et al.*, 2024a] Jiaxi Jiang, Paul Streli, Xuejing Luo, Christoph Gebhardt, and Christian Holz. Manikin: biomechanically accurate neural inverse kinematics for human motion estimation. In *European Conference on Computer Vision*, pages 128–146. Springer, 2024.

[Jiang *et al.*, 2024b] Jiaxi Jiang, Paul Streli, Manuel Meier, and Christian Holz. Egoposer: Robust real-time egocentric pose estimation from sparse and intermittent observations everywhere. In *European Conference on Computer Vision*, pages 277–294. Springer, 2024.

[Kim *et al.*, 2023] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8255–8263, 2023.

[Lan *et al.*, 2023] Gongjin Lan, Yu Wu, Fei Hu, and Qi Hao. Vision-based human pose estimation via deep learning: A survey. *IEEE Transactions on Human-Machine Systems*, 53(1):253–268, February 2023.

[Liu *et al.*, 2021] Junfa Liu, Juan Rojas, Yihui Li, Zhijun Liang, Yisheng Guan, Ning Xi, and Haifei Zhu. A graph attention spatio-temporal convolutional network for 3d human pose estimation in video. In *2021 IEEE international conference on robotics and automation (ICRA)*, pages 3374–3380. IEEE, 2021.

[Loper *et al.*, 2015] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

[Lorenz *et al.*, 2022] Michael Lorenz, Gabriele Bleser-Taetz, Takayuki Akiyama, Takehiro Niikura, Didier Stricker, and Bertram Taetz. Towards artefact aware human motion capture using inertial sensors integrated into loose clothing. 05 2022.

[Mahmood *et al.*, 2019] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes, 2019.

[Millerdurai *et al.*, 2024] Christen Millerdurai, Hiroyasu Akada, Jian Wang, Diogo Luvizon, Christian Theobalt, and Vladislav Golyanik. Eventego3d: 3d human motion capture from egocentric event streams. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1186–1195, 2024.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

[Mollyn *et al.*, 2023] Vimal Mollyn, Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja. Imuposer: Full-body pose estimation using imus in phones, watches, and earbuds. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, volume 38 of *CHI '23*, page 1–12. ACM, April 2023.

[Patel *et al.*, 2020] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style, 2020.

[Pavllo *et al.*, 2019] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training, 2019.

[Sosa and Hogg, 2023] Jose Sosa and David Hogg. Self-supervised 3d human pose estimation from a single image, 2023.

[Tevet *et al.*, 2022] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model, 2022.

[Tseng *et al.*, 2022] Jonathan Tseng, Rodrigo Castellon, and C. Karen Liu. Edge: Editable dance generation from music, 2022.

[Wang *et al.*, 2020] Manchen Wang, Joseph Tighe, and Davide Modolo. Combining detection and tracking for human pose estimation in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11088–11096, 2020.

[Wouwe *et al.*, 2024] Tom Van Wouwe, Seunghwan Lee, Antoine Falisse, Scott Delp, and C. Karen Liu. Diffusionposer: Real-time human motion reconstruction from arbitrary sparse sensors using autoregressive diffusion, 2024.

[Yi *et al.*, 2021] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Transpose: Real-time 3d human translation and pose estimation with six inertial sensors, 2021.

[Yi *et al.*, 2022] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors, 2022.

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

[Yuan *et al.*, 2023] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16010–16021, 2023.

[Zhang *et al.*, 2022a] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13232–13242, 2022.

[Zhang *et al.*, 2022b] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022.

[Zhang *et al.*, 2024] Yu Zhang, Songpengcheng Xia, Lei Chu, Jiarui Yang, Qi Wu, and Ling Pei. Dynamic inertial poser (dynaip): Part-based motion dynamics learning for enhanced human pose estimation with sparse inertial sensors, 2024.

[Zheng *et al.*, 2023] Xiaozheng Zheng, Zhuo Su, Chao Wen, Zhou Xue, and Xiaojie Jin. Realistic full-body tracking from sparse observations via joint-level modeling, 2023.

[Zhou *et al.*, 2025] Wenyang Zhou, Zhiyang Dou, Zeyu Cao, Zhouyingcheng Liao, Jingbo Wang, Wenjia Wang, Yuan Liu, Taku Komura, Wenping Wang, and Lingjie Liu. Emdm: Efficient motion diffusion model for fast and high-quality motion generation. In *European Conference on Computer Vision*, pages 18–38. Springer, 2025.

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