This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

EnvPoser: Environment-aware Realistic Human Motion Estimation from Sparse Observations with Uncertainty Modeling

Songpengcheng Xia[1] Yu Zhang [1] Zhuo Su[2] * Xiaozheng Zheng[2] Zheng Lv[2] Guidong Wang[2]
Yongjie Zhang[2] Qi Wu[1,2] Lei Chu [3] Ling Pei[1] †
[1]Shanghai Jiao Tong University [2]ByteDance [3]University of Southern California

Abstract

Estimating full-body motion using the tracking signals of head and hands from VR devices holds great potential for various applications. However, the sparsity and unique distribution of observations present a significant challenge, resulting in an ill-posed problem with multiple feasible solutions (i.e., hypotheses). This amplifies uncertainty and ambiguity in full-body motion estimation, especially for the lower-body joints. Therefore, we propose a new method, EnvPoser, that employs a two-stage framework to perform full-body motion estimation using sparse tracking signals and pre-scanned environment from VR devices. EnvPoser models the multi-hypothesis nature of human motion through an uncertainty-aware estimation module in the first stage. In the second stage, we refine these multi-hypothesis estimates by integrating semantic and geometric environmental constraints, ensuring that the final motion estimation aligns realistically with both the environmental context and physical interactions. Qualitative and quantitative experiments on two public datasets demonstrate that our method achieves state-of-the-art performance, highlighting significant improvements in human motion estimation within motion-environment interaction scenarios. Project page: `https://xspc.github.io/EnvPoser/`.

1. Introduction

In recent years, motion capture algorithms using sparse tracking signals have achieved significant progress across various research fields, establishing themselves as essential technologies for applications such as motion-sensing games and immersive VR/AR experiences [13, 18, 42, 49]. With the rapid development of VR devices, full-body motion estimation algorithms based on head-mounted display (HMD) have shown great application potential. However, devices like PICO and Quest typically provide sparse tracking signals of the head and hands, presenting challenges in estimating the full-body motion from these limited observations.

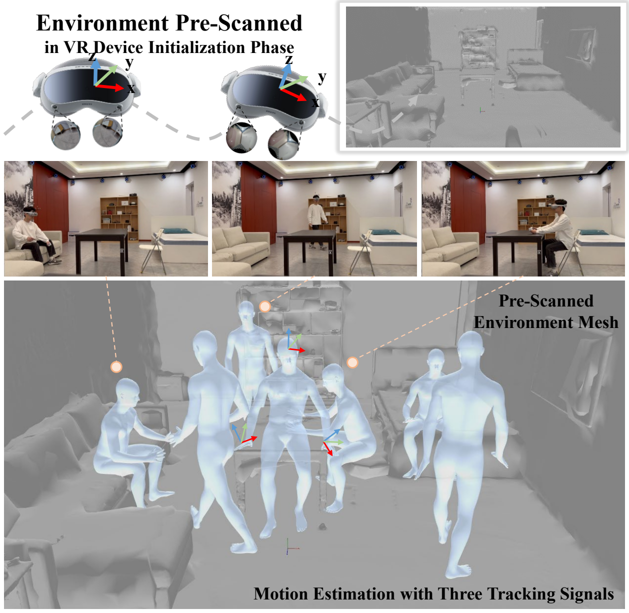

Figure 1. EnvPoser can estimate the full-body motion using three tracking signals (HMD and hand controllers) and a pre-scanned environment mesh.

Current sparse-observation-based human motion estimation methods typically rely on large-scale motion capture datasets, such as AMASS [28], to capture full-body motion through regression [14, 55] or generative models [2, 6, 8]. Alternatively, physics engines and reinforcement learning have been used [39] to synthesize full-body motion, bypassing dataset dependency. Utilizing various advanced architectures or large-scale datasets, these methods [2, 6, 8, 39, 55] attempt to model the mapping from sparse tracking signals to the most probable full-body motion. However, the lack of direct joint observations results in the same sparse input often corresponding to multiple plausible motions, making it a challenge for sparse-observation-

---

*Project Lead

†Corresponding Author

This work was supported by the National Nature Science Foundation of China (NSFC) under Grant 62273229.

based motion estimation algorithms to converge on the most probable full-body motion. To address this one-to-many mapping ambiguity, existing methods [4, 30, 31] attempt to add sensors to regions such as the pelvis and legs to narrow the distribution of possible motion outcomes. Although additional sensors can enhance the accuracy and robustness of full-body motion estimation, this approach may negatively impact user experience due to increased hardware requirements [13, 17, 41, 49, 54].

Therefore, an alternative approach to addressing the inherent ambiguity in sparse observations is to incorporate additional environmental constraints, as shown in Fig. 1. Human motion is highly correlated with the surrounding environment, which contains rich context that can provide valuable cues for motion estimation. However, effectively utilizing environmental constraints presents its own challenges, as existing methods [35, 55] often simplify human-environment interactions to basic foot-to-ground contact, overlooking the complexities of interactions. In reality, individuals frequently engage in more complex interactions with their surroundings.

Our key insight is that joint uncertainty estimation can explicitly model the multi-hypothesis nature of full-body motion estimation, and by incorporating environmental information, we can reduce estimation uncertainty, guiding the motion estimation to converge toward the most plausible outcome that aligns with both the environment and sparse input observations. Therefore, we design a two-stage deep learning framework to handle the one-to-many mapping ambiguity challenge. We first explicitly model the multi-hypothesis human motion by the joint uncertainty estimation and its sampling strategy. Then, by integrating environmental information, we guide the uncertainty-based multi-hypothesis motion estimations to converge on a full-body motion that best aligns with both the environment and sparse observations. To address the challenge of effectively utilizing environmental information, we introduce semantic constraints for non-contact motions, providing contextual cues for realistic estimation, and geometric constraints for contact scenarios, ensuring consistency and preventing collisions.

The key contributions of this paper are as follows:

- We propose EnvPoser, a novel framework that significantly advances full-body motion estimation from sparse tracking signals by incorporating pre-scanned environmental context. This framework addresses the complexities of estimating human motion from sparse observations, ensuring robustness across diverse interactive scenarios.
- This work presents a two-stage framework for human motion estimation: an uncertainty-aware module provides initial estimates that account for multi-hypothesis motion, while an environment-aware refinement module integrates semantic and geometric constraints to ensure realistic outcomes in diverse scenarios, setting a new benchmark in motion estimation with environmental interactions.
- Comprehensive quantitative and qualitative experiments demonstrate that our proposed method achieves state-of-the-art performance across two public datasets and excels in scenarios with frequent environmental interactions.

2. Related Work

2.1. Motion estimation with sparse observations

Current Wearable-based human motion estimation approaches mainly adopt two configurations: one uses multiple inertial measurement units (IMUs) placed on body extremities [13, 17, 36, 40, 48, 54], while the other relies on an HMD with two controllers [2, 4, 7, 14, 15, 55].

While methods utilizing six IMUs have shown impressive results, HMD-based techniques leveraging sparse three-point observations show significant potential for practical applications. For instance, AvatarPoser [14] employs a transformer architecture to achieve accurate full-body motion. While Zheng et al. [55] developed a two-stage framework to capture joint dependencies, improving motion accuracy and fluidity. Some studies [1, 5, 8] address the ill-posed problem of sparse-to-dense mapping using generative models. Dittadi et al. [5] and Aliakbarian et al. [1] proposed using Variational Autoencoders (VAE) and normalizing flows, respectively, to estimate full-body poses from three tracking points. Additionally, the diffusion model [6, 8, 35] has also shown notable performance in motion estimation with sparse data. Despite these advances, most methods rely solely on sparse tracking signals, leading to high uncertainty and multiple potential poses for the same input. Therefore, this paper introduces joint uncertainty estimation and constrains physically implausible motions using environmental information.

2.2. Pose regression with uncertainty estimation

The sparse nature of observations in 3D human pose estimation introduces significant uncertainty for unobserved joints, as identical inputs can correspond to multiple pose hypotheses [8, 54, 55]. This challenge is particularly evident in image-based pose estimation [3, 9, 22, 24, 43], where monocular methods face depth ambiguity. For instance, Wehrbein et al. [38] modeled the posterior distribution of 3D pose hypotheses using normalizing flows, while MHFormer [24] employed a multi-hypothesis transformer to capture diverse pose hypotheses through a one-to-many mapping. Recently, Shan et al. [33] proposed diffusion-based methods to aggregate multi-hypothesis predictions with joint-wise re-projection.

In our paper, we aim to use uncertainty estimation to model the multi-hypothesis nature of motion from sparse observations. Uncertainty in deep learning can be categorized into two types [19]: aleatoric (data-related) and epistemic (model-related). Li et al. [23] refined 3D poses by incorporating joint uncertainties through uncertainty guided-sampling and uncertainty-guided self-attention, while Zhang et al. [53] introduced a probabilistic framework that encodes aleatoric uncertainty using a robust negative log-likelihood loss, with epistemic uncertainty guiding model refinement. In wearable motion estimation, Yang et al. [46] leveraged uncertainty to weight IMU-derived features, enhancing pose corrections with additional text annotations.

2.3. Scene-aware motion estimation and generation

Recent advancements in human motion estimation and generation have expanded from solely analyzing body motion to integrating interactions with the surrounding environment, especially vision-based approaches [11, 12, 16, 34, 47, 52]. For instance, PROX [10] incorporated scene constraints into monocular human pose estimation to address depth ambiguity, while Shen et al. [34] used sparse 3D CNNs to estimate absolute positions and dense scene contacts, refining human mesh recovery via cross-attention with 3D scene cues. EgoHMR [52] leveraged ego-camera images and diffusion models within pre-scanned environments to recover full-body motion. Similarly, S2Fusion [35] utilized sparse observations by learning human motion priors on synthetic datasets and conditioned motion generation on scene encoding and VR inputs. However, S2Fusion primarily focuses on lower body and environment interactions. QuestEnvsim [21] employed reinforcement learning to generate realistic poses in highly constrained environments using HMDs. Additionally, recent studies further emphasize the integration of scene context and historical motion to produce plausible human motion [27, 44, 45].

These scene-aware methods underscore the importance of environmental context for enhancing motion estimation. Our approach builds on these foundations by comprehensively leveraging human-environment interactions, addressing both contact and non-contact scenarios to improve motion estimation with sparse tracking signals.

3. Method

3.1. Problem Statement

Our goal is to estimate the full-body motion sequence $\boldsymbol{\theta} = \{\theta_t\}_{t=1}^{T} \in \mathcal{R}^{T\times 132}$ (the first 22 joints of SMPL [26] model with the 6D representation of rotations) using the sparse tracking signals $\boldsymbol{X} = \{\boldsymbol{x}_t\}_{t=1}^{T} \in \mathcal{R}^{T\times N_c}$ from the HMD and hand controllers, along with a pre-scanned 3D environment point cloud $\boldsymbol{S}$, where T denotes the length of input, N_c = 36 means the channel of input. Each sparse tracking signal at time step t consists of the head and hand positions $\boldsymbol{p}_t$, rotations $\boldsymbol{R}_t$, and linear velocities $\boldsymbol{v}_t$, represented as $\boldsymbol{x}_t = [\boldsymbol{p}_t, \boldsymbol{R}_t, \boldsymbol{v}_t]$. Moreover, we utilize a cropped 3D environment point cloud $\boldsymbol{V_S} \in \mathcal{R}^{N_S\times 3}$, where N_S, the number of sampled points, is set to 1000. With the human translation, this point cloud refines pose estimation by incorporating environmental context, leveraging both contact and non-contact interactions between the body and the environment to reduce the uncertainty inherent in sparse observations.

As shown in Fig. 2, our approach is a two-stage model comprising two core technical modules: 1) an uncertainty-aware initial motion estimation module; and 2) an environment-aware motion refinement module that leverages semantic and geometric information from the environment. We first train the uncertainty-aware initial motion estimation module (Stage I) on the AMASS dataset, with a focus on explicitly capturing the multi-hypothesis nature of motion estimation through uncertainty quantification. Subsequently, we jointly train both modules (Stage II) on the motion-environment interaction datasets, applying environmental semantic and geometric constraints to refine the multi-hypothesis estimates from Stage I into the most plausible motion estimation.

3.2. Uncertainty-aware Initial Human Motion Estimation Module

To obtain the joint uncertainty quantification and explicitly model the multi-hypothesis nature of motion estimation, we propose an uncertainty-aware initial human motion estimation module. This module is designed with three primary objectives: (1) to integrate historical motion states and sparse observations for the current initial motion estimation; (2) to effectively extract features from sparse signals; and (3) to quantify uncertainty in motion estimation, explicitly modeling the multi-hypothesis nature of motion under identical input conditions. Accordingly, this module comprises three sub-modules: sparse input signal and historical motion embedding, transformer-based motion feature extraction, and human motion and joint uncertainty regression.

Sparse Input Signal and Historical Human Motion Embedding. To integrate historical motion states and sparse observations for the current initial motion estimation, we adopt an auto-regressive structure in the uncertainty-aware initial human motion estimation module. Our model takes sparse observations $\boldsymbol{X} \in \mathcal{R}^{T\times 36}$ and historical motion $\boldsymbol{X}_{hm} \in \mathcal{R}^{T\times 132}$ as inputs, embedding these inputs through linear layers to obtain the observation representation $\boldsymbol{Z}_X$ and motion historical state representation $\boldsymbol{Z}_{hm}$. Then, we concatenate the sparse observation embedding and historical motion embedding to obtain the shallow motion representation $\boldsymbol{Z}_S = [\boldsymbol{Z}_X, \boldsymbol{Z}_{hm}]$, which can be used for subsequent transformer-based feature extraction.

Transformer-based Motion Feature Extraction.

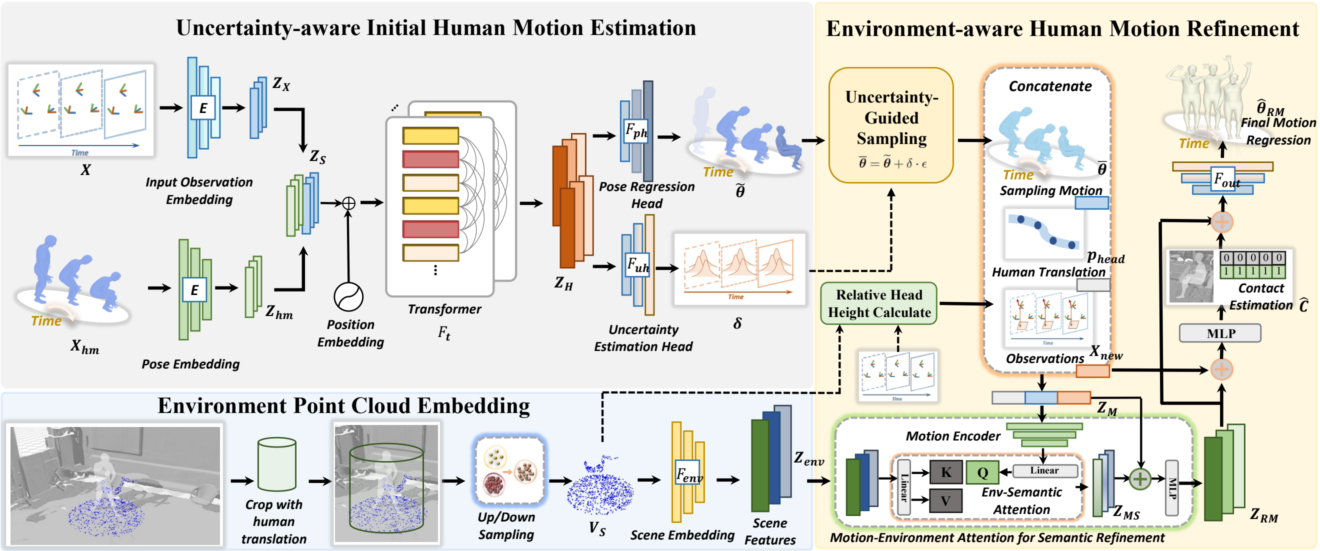

Figure 2. Overview of EnvPoser: A Two-Stage Motion Estimation Model. Stage I involves training the uncertainty-aware initial estimation module on the AMASS dataset to produce initial motion estimates with uncertainty quantification. Stage II refines these estimates by training on motion-environment datasets, incorporating semantic and geometric environmental constraints.

Based on the shallow motion representation $\boldsymbol{Z}_S$, we employ a transformer-based feature extraction module [17] for motion features learning. To retain sequence order information, temporal positional encoding is applied to the shallow motion representation [37]. Using multi-head self-attention and a position-wise feed-forward network within the transformer encoder F_t, this module efficiently captures intricate relationships in the sparse tracking signals and historical motion, resulting in high-dimensional motion features $\boldsymbol{Z}_H = F_t(\boldsymbol{Z}_S)$. While the transformer architecture is effective for extracting motion features, the sparsity of the input data—limited to tracking signals from the head and hands—introduces high uncertainty in estimating unobserved joints, particularly in the lower body. This uncertainty arises from the lack of direct observations for these joints, making their estimation inherently ambiguous. To address this challenge, we incorporate a mechanism to quantify the uncertainty in joint estimations.

Human Motion and Joint Uncertainty Regression. With the high-dimensional motion features $\boldsymbol{Z}_H$, we employ a pose regression network $F_{ph}(\cdot)$ to predict the human motion $\widetilde{\boldsymbol{\theta}} = F_{ph}(\boldsymbol{Z}_H)$. Given the sparse nature of the input signals, capturing joint uncertainty is crucial in mapping the limited observations to a full-body motion estimation. To achieve this, we utilize a heteroscedastic neural network [19, 23] to estimate the joint uncertainty. Specifically, an uncertainty regression head $F_{uh}(\cdot)$ is employed on the high-dimensional motion feature to estimate the joint uncertainty $\boldsymbol{\delta} \in \mathcal{R}^{T\times 132}$ of the predicted motion.

Using the predicted motion and its associated uncertainty, we resample the full-body motion $\overline{\boldsymbol{\theta}}$ around the predicted mean motion $\widetilde{\boldsymbol{\theta}}$ according to a Gaussian distribution $\mathcal{N}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\delta})$ with the uncertainty δ. This resampling technique introduces a multi-hypothesis representation, characterizing the inherent ambiguity in the unobserved joints. To ensure differentiability throughout the model training process, we apply the reparameterization trick on this sampling operation, drawing a parameter ϵ from a standard Gaussian distribution $\mathcal{N}(0,1)$, represented by $\overline{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}} + \boldsymbol{\delta} \cdot \epsilon$.

Through the uncertainty regression head, we can obtain the joint uncertainty estimation for predicted human motion, enabling a multi-hypothesis motion representation with the sampling strategy. We pre-train the first stage of our method on the large synthetic dataset AMASS [28]. Based on the predicted mean human motion $\widetilde{\boldsymbol{\theta}}$ and uncertainty $\boldsymbol{\delta}$, we train the first stage module with the following loss function:

$$L_M = \|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2, \tag{1}$$

where $\boldsymbol{\theta}$ represents the ground-truth of human motion. Additionally, to ensure accurate uncertainty prediction, we also incorporate an uncertainty estimation loss as [19] stated:

$$L_\delta = \|\frac{\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}}{\boldsymbol{\delta}}\|_2 + \log(\|\boldsymbol{\delta}\|_2). \tag{2}$$

Thus, based on the uncertainty-aware initial human motion module, the objective function for Stage I is defined as follows:

$$L_{S-I} = \lambda_M L_M + \lambda_\delta L_\delta, \tag{3}$$

where the λ_M and λ_δ are the hyper-parameters for Stage I training.

Training Stage I on the AMASS dataset, along with the design for motion uncertainty quantification, allows our network to effectively capture the multi-hypothesis nature of motion estimation arising from sparse observations. To address the inherent one-to-many ambiguity, we introduce environmental information as a refinement stage. By incorporating both semantic and geometric environmental information, our method guides the multi-hypothesis motion estimates to converge toward the solution that best aligns with the sparse observations and environmental constraints.

3.3. Environment-aware Human Motion Refinement with Semantic and Geometry Constraint

Based on the initial motion estimation and its uncertainty produced by the first-stage uncertainty-aware module, we further refine the multi-hypothesis human motion estimates using the pre-scanned environment point clouds collected by real-world devices. To effectively utilize the environmental information, our design is guided by a key insight: human motion within an environment involves both object interactions (*contact scenarios*) and collision avoidance (*non-contact scenarios*). In this module, the environment point clouds are processed through cropping, sampling, and encoding. We then apply constraints based on the environment's semantic and geometric properties to ensure that the full-body motion estimation realistically aligns with the surrounding context.

Environmental Point Cloud Embedding. To enhance computational efficiency and focus on relevant interaction areas, we first clip the environmental point clouds based on the human's global position. A circular sampling approach is used to minimize the impact of body orientation, selecting a 1-meter radius around the body and uniformly distributing N_S points within this area, denoted as $\boldsymbol{V_S} \in \mathcal{R}^{N_S \times 3}$. The cropped environmental point cloud $\boldsymbol{V_S}$ is then processed by the environment encoder $F_{env}(\cdot)$ to obtain the environmental point cloud embedding $\boldsymbol{Z}_{env}$, represented by:

$$\boldsymbol{Z}_{env} = F_{env}(\boldsymbol{V_S}), \tag{4}$$

where the environment encoder $F_{env}(\cdot)$ is implemented using the vanilla PointNet++ [32]. The extracted environmental embedding would be used in the environment semantic-aware motion refinement.

Environment Semantic-Aware Constraint for Motion Refinement. Capturing semantic information from the environmental point cloud provides crucial guidance for motion generation [27, 44, 47], as human motion is highly correlated with environmental context. Therefore, leveraging this motion-environment correlation as a valuable supplement has the potential to enhance the sparse-observation-based human motion estimation. Inspired by [27, 34], we employ a cross-attention module between the initial motion estimates and the cropped environmental point clouds to refine the human motion estimation. This module comprises an environment-embedding network, a motion information embedding network and an environment-motion cross-attention network.

The motion information embedding network $F_{motion}(\cdot)$ takes the sampled motion estimation $\widetilde{\boldsymbol{\theta}}$, head translation $\boldsymbol{p}^{head}$ and the extended sparse observations $\boldsymbol{X}_{new} = \{\boldsymbol{X}, \boldsymbol{h}, \boldsymbol{\theta}^{up}\} \in \mathcal{R}^{T\times 40}$ as input. The extended sparse observations $\boldsymbol{X}_{new}$ include the relative head height $\boldsymbol{h}$ and head up-vector $\boldsymbol{\theta}^{up}$, obtained via a relative head height calculation module. Following [17, 20], this module calculates terrain height by using the height of the feet when stationary and nearby environmental points, and then determines $\boldsymbol{h}$ as the difference between the absolute head height from the HMD and the estimated terrain height. These inputs are concatenated and processed through linear layers to produce the motion information embedding $\boldsymbol{Z}_M = F_{motion}(concat(\overline{\boldsymbol{\theta}}, \boldsymbol{p}^{head}, \boldsymbol{X}_{new}))$.

For the environment-motion cross-attention network, the Key $\boldsymbol{K}$ and Value $\boldsymbol{V}$ vectors can be obtained by multiplying the weight matrix $\boldsymbol{W}_k, \boldsymbol{W}_v$ with the environmental embedding $\boldsymbol{Z}_{env}$, while the motion information embedding $\boldsymbol{Z}_M$ multiplies the weight matrices $\boldsymbol{W}_q$ to get the Query $\boldsymbol{Q}$ vectors. Inspired by [25, 27], we incorporate a spatial salience $\boldsymbol{s}_{spatial} = MLP(\boldsymbol{V_S})$ that expects our cross-attention model to be able to perceive the distance between environmental points clouds and human bodies. Using the spatial salience $\boldsymbol{s}_{spatial}$ in conjunction with attention scores $Attn(Q, K)$, we derive a motion-environment representation $\boldsymbol{Z}_{ME}$, which could be represented by

$$\boldsymbol{Z}_{ME} = (Attn(\boldsymbol{Q}, \boldsymbol{K}) + \boldsymbol{s}_{spatial}) \cdot \boldsymbol{V}. \tag{5}$$

Then, by concatenating the motion-environment representation $\boldsymbol{Z}_{ME}$ with the motion information embedding $\boldsymbol{Z}_M$ and passing them through an MLP network, we obtain the environment-refined motion representation $\boldsymbol{Z}_{RM} = MLP(concat(\boldsymbol{Z}_{ME}, \boldsymbol{Z}_M))$.

Using the environment-refined motion representation $\boldsymbol{Z_{RM}}$, which integrates environmental semantics and spatial priors between the human and environment, the model can then regress the probability of human-environment contact and generate the final motion estimation.

Motion Regression with Contact Estimation. With the environment-refined motion representation $\boldsymbol{Z}_{RM}$, we estimate the contact probability $\boldsymbol{C} = \{\boldsymbol{c}\}_{t=1}^{T} \in \mathcal{R}^{T\times 22}$ and regress the final human motion $\hat{\boldsymbol{\theta}}_{RM}$. We take the environment-refined motion representation $\boldsymbol{Z}_{RM}$ and the extended sparse observations $\boldsymbol{X}_{new}$ as input, and apply two-layers MLP layers on the concatenated input to get the contact probability prediction $\hat{\boldsymbol{C}} = MLP(concat(\boldsymbol{X}_{new}, \boldsymbol{Z_{RM}}))$.

Contact probability prediction is typically treated as a binary classification problem, where Binary Cross Entropy (BCE) loss is used for training:

$$L_{contact} = BCELoss(\hat{\boldsymbol{C}}, \boldsymbol{C}). \tag{6}$$

Finally, the contact probability $\hat{\boldsymbol{C}}$ and environment-refined motion representation $\boldsymbol{Z}_{RM}$ are concatenated to regress the final motion with a motion decoder network $F_{out}(\cdot)$:

$$\hat{\boldsymbol{\theta}}_{RM} = F_{out}(concat(\boldsymbol{Z_{RM}}, \hat{\boldsymbol{C}})), \tag{7}$$

where $F_{out}(\cdot)$ consists of two linear layers with a ReLU activation function. We then obtain the final motion loss:

$$L_{M'} = \|\hat{\boldsymbol{\theta}}_{RM} - \boldsymbol{\theta}\|_2. \tag{8}$$

Scene Geometry-Aware Constraint for Motion Refinement. To complement semantic-level constraints, we leverage the geometric structure of the environment to enhance motion estimation accuracy, especially in human-environment interactions (contact scenarios). A straightforward approach is to detect collisions between the estimated human motion and the environment point cloud, thereby refining unrealistic motions. While existing methods [10, 50] often convert 3D environments into Signed Distance Fields (SDF) for collision detection, this approach can be computationally intensive. Inspired by recent advancements [29, 52], we instead employ the COAP model [29] to efficiently compute collisions between the human body and environmental points. With this novel model and the environment point clouds, the COAP-based collision loss [52] can be formulated as:

$$L_{coap} = \frac{1}{N_S}\sum_{i=1}^{N_S} \sigma(f_{\Theta}(V_{S_i}|\mathcal{G}))\mathbb{I}_{f_{\Theta}(V_{S_i}|\mathcal{G})>0}, \tag{9}$$

where $\boldsymbol{V_S}$ denotes the cropped environment point clouds, with N_S representing the number of points. Θ represents the transformation parameters of the human body, encompassing both the motion pose $\boldsymbol{\theta}$ and shape $\boldsymbol{\beta}$. $f_{\Theta}(V_{S_i}|\mathcal{G})$ represents the spatial relationship between the human body model and environmental points. The function $\sigma(\cdot)$ is the $Sigmoid$ function, which maps the value of $f_{\Theta}(V_{S_i}|\mathcal{G})$ to a range between 0 and 1.

Since most motions occur on the ground, we remove ground points from the scene to ensure that the COAP-based collision loss specifically targets interactions with surrounding objects. To account for interactions with the ground, we introduce additional constraints, including foot contact, foot height, and ground penetration losses [55]:

$$\begin{aligned} L_{fc} &= \|(\hat{\boldsymbol{P}}^{feet}_{RM} - \boldsymbol{P}^{feet}) \cdot \boldsymbol{C}\|_1, \\ L_{gfh} &= \|\hat{\boldsymbol{z}}^{feet}_{P_{RM}} - \boldsymbol{z}^{ground}\|_1, \\ L_{gp} &= \|(\hat{\boldsymbol{z}}^{min}_{P_{RM}} - \boldsymbol{z}^{ground}) \cdot \boldsymbol{l}\|_1, \end{aligned} \tag{10}$$

where the $\hat{\boldsymbol{z}}^{feet}_{P_{RM}}$ and $\boldsymbol{z}^{ground}$ means the height of feet joint positions and the floor height, and $\boldsymbol{l}$ denotes whether the joint is lower than the ground.

Final Loss Function. In addition to the first stage loss (Eq.2, Eq.1), COAP-based collision loss (Eq.9), contact loss (Eq.6), foot-ground loss (Eq.10) and final motion loss (Eq.8), we also apply some additional losses to improve model training. With the forward kinematic chain, the joint position $\boldsymbol{P}$, $\hat{\boldsymbol{P}}_{RM}$, and the global hands positions $\boldsymbol{P}^{hand}$, $\hat{\boldsymbol{P}}^{hand}_{RM}$ could be calculated by the motion pose $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$. To penalize the errors accumulating along the kinematic chain, we apply a joint position loss (L_{posi}) and a hand alignment loss(L_{hAL}):

$$L_{posi} = \|\hat{\boldsymbol{P}}_{RM} - \boldsymbol{P}\|_2, L_{hAL} = \|\hat{\boldsymbol{P}}^{hand}_{RM} - \boldsymbol{P}^{hand}\|_1. \tag{11}$$

Therefore, with the technical analysis provided above, we can formulate the second stage objective function as follows:

$$\begin{aligned} L_{S-II} &= L_{S-I} + L_{M'} + \lambda_1 L_{posi} + \lambda_2 L_{hAL} + \lambda_3 L_{fc} \\ &+ \lambda_4 L_{contact} + \lambda_5 L_{gfh} + \lambda_6 L_{gp} + \lambda_7 L_{coap}. \end{aligned} \tag{12}$$

Here, $\{\lambda_i\}_{i=1,\ldots,7}$ represents the hyperparameters. Our model is first trained on the AMASS using the Eq. 3 and then fine-tuned on a motion-environment interaction dataset with Eq. 12. The training process and selection of hyperparameters can be found in the supplementary material.

4. Experiments

This section introduces the datasets, metrics, and competing algorithms used in our experiments. We then compare our method EnvPoser with state-of-the-art algorithms, highlighting its advantages in motion estimation. Finally, ablation studies validate the effectiveness of each module. Additional case studies and discussions are provided in the supplementary materials.

Datasets. To comprehensively evaluate the efficacy of our proposed method, we conduct experiments on two challenging public motion-scene interactions datasets: Egobody [51] and GIMO [56] datasets. Detailed information about these datasets can be found in the supplementary document.

Competing methods. We compare the motion estimation performance of the proposed model with the state-of-the-art competing methods, including AvatarPoser [14], AGRoL [6], AvatarJLM [55], and S2Fusion [35] on two public datasets. We re-train all competing methods on the EgoBody and GIMO datasets until convergence. For the results on the GIMO dataset, we reused the values reported in previous studies [35]. For the EgoBody dataset, we tested our trained model under real-time settings and presented the results as follows. It is worth noting that S2Fusion [35] is not fully open-source; therefore, we reproduced S2Fusion based on the details provided in its paper and partial open-source code, marking it as S2Fusion* in this experiment.

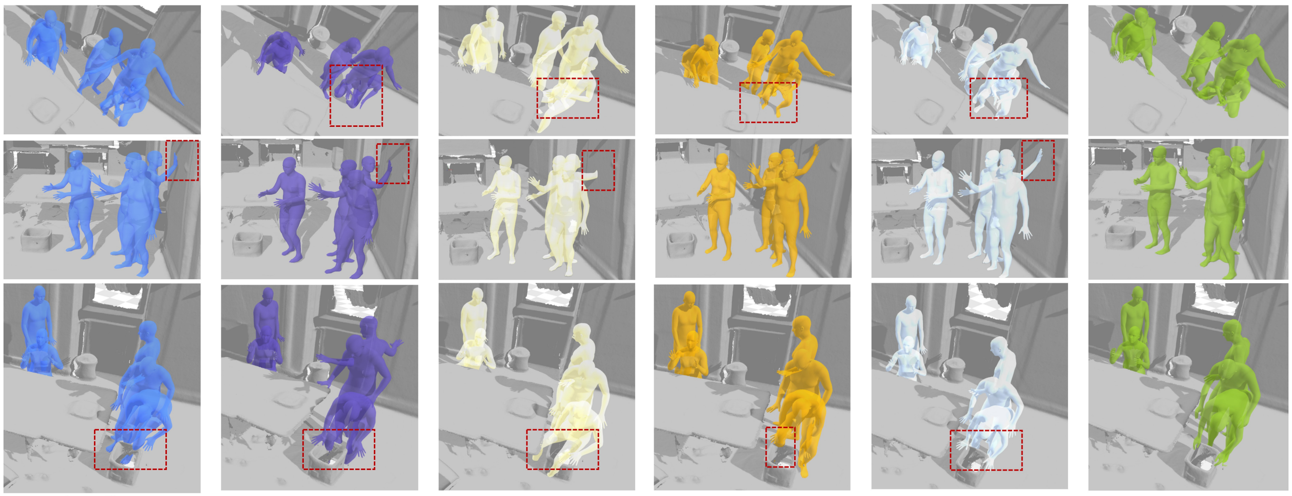
AvatarPoser AGRoL AvatarJLM S2Fusion* EnvPoser GT

Figure 3. Visualization of motion estimation on three test sequences from EgoBody Dataset [51].

	EgoBody			
	MPJRE(°)	MPJPE(mm)	MPJVE(mm/s)	Jitter
AvatarPoser[14]	6.80	97.8	237.3	11.0
AGRoL[6]	7.21	100.9	370.3	21.1
AvatarJLM[55]	6.42	91.7	177.4	7.2
S2Fusion*[35]	6.65	89.2	219.4	12.5
EnvPoser	**6.00**	**74.7**	**174.0**	**6.6**
	GIMO			
	MPJRE(°)	MPJPE(mm)	MPJVE(mm/s)	Jitter
AvatarPoser[14]	7.02	91.3	324.0	16.4
AGRoL[6]	6.58	88.6	269.4	12.5
AvatarJLM[55]	4.95	70.7	258.1	10.7
S2Fusion[35]	4.65	57.8	235.7	10.1
EnvPoser	**4.38**	**57.6**	**234.6**	**8.9**

Table 1. The performance comparison with SOTAs.

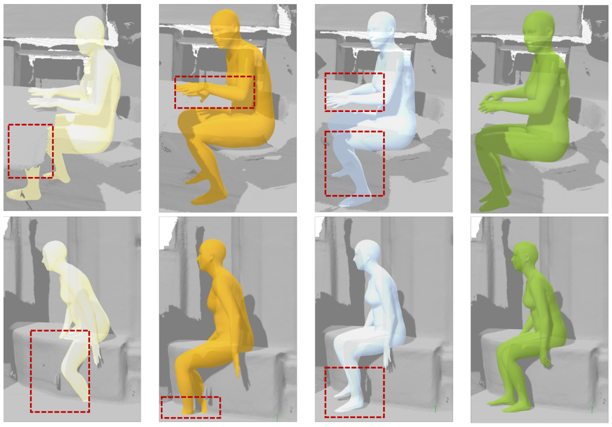
AvatarJLM S2Fusion* EnvPoser GT

Figure 4. Qualitative Comparison of Interaction Details.

Metrics. Following [35, 55], we use these metrics for evaluation: *1) MPJRE [°]*: The mean joint rotational error across all body joints, measured in degrees; *2) MPJPE [cm]*: The mean positional error per joint, measured in centimeters. *3) MPJVE [cm/s]*: Represents the average error in joint velocities, expressed in centimeters per second; *4) Jitter* : Measures motion smoothness by calculating the average jerk (rate of change of acceleration) across all joints, with lower values indicating smoother motion.

4.1. Comparison with the State-of-the-Art

Quantitative Experimental Analysis. We compare our proposed method with the competing methods on the two public motion-environment interaction datasets. Among these, AvatarPoser [14], AGRoL [6], and AvatarJLM [55] are the SOTAs that rely solely on sparse tracking signals. As shown in Tab. 1, our model, which incorporates pre-scanned environmental point cloud information, effectively constrains the uncertainty in motion reconstruction caused by sparse tracking signals. Compared to AvatarJLM [55], which does not utilize environmental information, our method achieves significant improvements on two key metrics—MPJRE and MPJPE—by 0.42 (6.5%)/17.0 (18.5%) and 0.57 (11.5%)/13.1 (18.5%) on the EgoBody and GIMO datasets, respectively. These results underscore the increased accuracy in motion estimation achieved through the integration of environmental information.

Additionally, we benchmark our model against the latest algorithm S2Fusion [35], which introduces environmental information as a constraint for human motion estimation. On the GIMO dataset, our method reduces MPJRE errors by 5.8% compared to S2Fusion [35], and it outperforms S2Fusion across all evaluation metrics, demonstrating the superior effectiveness of our approach in leveraging environmental information. Moreover, the motion estimation performance of our method also outperforms S2Fusion* on the EgoBody dataset, which reduces MPJRE and MPJPE errors by 9.8% and 16.3%. Furthermore, our method shows a marked improvement in motion smoothness, with an 11.9%

	EgoBody		GIMO	
	MPJRE	MPJPE	MPJRE	MPJPE
①Baseline	7.45	87.1	5.72	74.6
②w/o env-semantic	6.04	78.0	4.55	62.5
③w/o env-geometry	6.03	77.5	4.63	63.2
EnvPoser	**6.00**	**74.7**	**4.38**	**57.6**

Table 2. The ablation study on environment refinement module.

reduction in Jitter over S2Fusion and a 16.8% improvement over AvatarJLM on the GIMO dataset.

Qualitative Experimental Analysis. Fig. 3 showcases the motion estimation results on the EgoBody dataset, where we present three representative sequences and visualize the human motion estimation results at key moments. For state-of-the-art algorithms that do not incorporate environmental information, the visualization results in Fig. 3 show multiple instances of environment penetration during interactions. For example, AvatarJLM exhibits errors such as hand intersections with walls (second row) and leg penetrations through boxes (third row). Furthermore, in scenarios without chairs, squatting motions are misinterpreted as sitting (first row) in AvatarJLM due to the absence of environment understanding. In contrast, EnvPoser effectively leverages environmental information to reduce penetration errors and produce more realistic motion estimates.

In more challenging interaction scenarios, EnvPoser shows certain advantages over S2Fusion, which also incorporates environmental information. Notably, S2Fusion fails to prevent interpenetration between the lower body and surrounding objects (third row in Fig. 3). Moreover, Fig. 4 provides a comparison between EnvPoser and S2Fusion, highlighting the advantages of our approach in handling upper-body interactions with objects and complex lower-body motions through the integration of uncertainty modeling and environmental semantic/geometric constraints. Additionally, compared to AvatarJLM, both EnvPoser and S2Fusion demonstrate the effectiveness of incorporating environmental information to improve human motion estimation.

4.2. Ablation Study

Effectiveness of the environment-aware refinement module. We conducted ablation studies comparing our method with three variants: ① Baseline: an auto-regressive transformer-based network with joint uncertainty estimation; ② w/o env-semantic: incorporating only geometric constraints; ③ w/o env-geometry: incorporating only semantic constraints. Tab. 2 presents the motion estimation performance for each variant. Compared to variant ①, our method achieves substantial improvements by integrating both semantic and geometric environmental information. Further analysis shows that each environment-aware refinement module independently enhances performance, with the combination yielding the best overall results.

	EgoBody			GIMO		
	MPJRE	MPJPE	MPJVE	MPJRE	MPJPE	MPJVE
1)Baseline w/o UNC	7.63	90.8	333.0	5.90	77.5	270.2
2)Baseline	7.45	87.1	342.9	5.72	74.6	266.7
3)EnvPoser w/o UNC	6.56	80.1	**172.7**	4.45	60.2	278.6
EnvPoser	**6.00**	**74.7**	174.0	**4.38**	**57.6**	**234.6**

Table 3. The effectiveness of the uncertainty estimation.

Effectiveness of the uncertainty estimation. As shown in Tab. 3, we compare two variants: 1) Baseline w/o UNC: a baseline without uncertainty estimation and 2) Baseline: a baseline with uncertainty estimation. Results show that integrating the uncertainty estimation module not only quantifies joint uncertainty across motions but also improves full-body motion estimation through additional supervision. We further incorporated both baseline methods into the environment-aware refinement module. As shown in Tab. 3, our method achieves significant improvements over variant 3) EnvPoser w/o UNC: our method without uncertainty estimation, reducing MPJRE and MPJPE errors by 8.11% and 5.68%, respectively, on the EgoBody dataset. These results indicate that the uncertainty estimation and resampling methods effectively capture the multi-hypothesis nature of initial motion estimates and, when combined with environmental refinement, adaptively align motion estimation with sparse tracking signals and environmental context.

5. Conclusions

This paper presents a novel two-stage framework for reliable human motion estimation leveraging sparse tracking signals and environmental context. In the first stage, we model the multi-hypothesis nature of motion through uncertainty estimation, followed by a refinement phase that applies environment-aware constraints from both semantic and geometric perspectives. Extensive experiments show that EnvPoser outperforms existing methods, particularly in complex interaction scenarios, highlighting the significance of environmental context in wearable motion estimation and paving the way for future approaches to harness this valuable contextual information. This work has the potential to impact a wide range of interactive applications that benefit from more accurate and context-aware motion estimation.

Limitations and future works. The model assumes a static environment, lacking consideration for dynamic multi-user interactions or object motion, which may limit its effectiveness. This limitation could be addressed by incorporating a third-person perspective to estimate the movement of other people and objects, along with additional semantic and geometric constraints. Furthermore, real-time mesh quality in complex environments is often limited, which may affect performance. Future work will explore using raw images to infer contacts in complex scenes via 2D semantic understanding and introducing additional constraints to improve overall reconstruction quality.

References

[1] Sadegh Aliakbarian, Pashmina Cameron, Federica Bogo, Andrew Fitzgibbon, and Thomas J Cashman. Flag: Flow-based 3d avatar generation from sparse observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13253–13262, 2022. 2

[2] Sadegh Aliakbarian, Fatemeh Saleh, David Collier, Pashmina Cameron, and Darren Cosker. Hmd-nemo: Online 3d avatar motion generation from sparse observations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9622–9631, 2023. 1, 2

[3] Rongyu Chen, Linlin Yang, and Angela Yao. Mhentropy: Entropy meets multiple hypotheses for pose and shape recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14840–14849, 2023. 2

[4] 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. 2

[5] Andrea Dittadi, Sebastian Dziadzio, Darren Cosker, Ben Lundell, Thomas J Cashman, and Jamie Shotton. Full-body motion from a single head-mounted device: Generating smpl poses from partial observations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11687–11697, 2021. 2

[6] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 481–490, 2023. 1, 2, 6, 7

[7] Zhen Fan, Peng Dai, Zhuo Su, Xu Gao, Zheng Lv, Jiarui Zhang, Tianyuan Du, Guidong Wang, and Yang Zhang. Emhi: A multimodal egocentric human motion dataset with hmd and body-worn imus. *arXiv preprint arXiv:2408.17168*, 2024. 2

[8] Han Feng, Wenchao Ma, Quankai Gao, Xianwei Zheng, Nan Xue, and Huijuan Xu. Stratified avatar generation from sparse observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 153–163, 2024. 1, 2

[9] Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Diffpose: Toward more reliable 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13041–13051, 2023. 2

[10] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2282–2292, 2019. 3, 6

[11] Chun-Hao P Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J Black. Capturing and inferring dense full-body human-scene contact. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13274–13285, 2022. 3

[12] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16750–16761, 2023. 3

[13] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser: learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Trans. Graph. (TOG)*, 37(6):1–15, 2018. 1, 2

[14] 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. 1, 2, 6, 7

[15] 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, 2025. 2

[16] Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1737–1747, 2024. 3

[17] 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)*, pages 1–9, 2022. 2, 4, 5

[18] Manuel Kaufmann, Yi Zhao, Chengcheng Tang, Lingling Tao, Christopher Twigg, Jie Song, Robert Wang, and Otmar Hilliges. Em-pose: 3d human pose estimation from sparse electromagnetic trackers. In *Proceedings of the IEEE/CVF international conference on computer vision(ICCV)*, pages 11510–11520, 2021. 1

[19] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. 3, 4

[20] Jiye Lee and Hanbyul Joo. Mocap everyone everywhere: Lightweight motion capture with smartwatches and a head-mounted camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1091–1100, 2024. 5

[21] Sunmin Lee, Sebastian Starke, Yuting Ye, Jungdam Won, and Alexander Winkler. Questenvsim: Environment-aware simulated motion tracking from sparse sensors. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–9, 2023. 3

[22] Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9887–9895, 2019. 2

[23] Han Li, Bowen Shi, Wenrui Dai, Hongwei Zheng, Botao Wang, Yu Sun, Min Guo, Chenglin Li, Junni Zou, and Hongkai Xiong. Pose-oriented transformer with uncertainty-guided refinement for 2d-to-3d human pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1296–1304, 2023. 3, 4

[24] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13147–13156, 2022. 2

[25] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 5

[26] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6), 2015. 3

[27] Zhenyu Lou, Qiongjie Cui, Haofan Wang, Xu Tang, and Hong Zhou. Multimodal sense-informed forecasting of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2144–2154, 2024. 3, 5

[28] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 1, 4

[29] Marko Mihajlovic, Shunsuke Saito, Aayush Bansal, Michael Zollhoefer, and Siyu Tang. Coap: Compositional articulated occupancy of people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13201–13210, 2022. 6

[30] Jose Luis Ponton, Haoran Yun, Andreas Aristidou, Carlos Andujar, and Nuria Pelechano. Sparseposer: Real-time full-body motion reconstruction from sparse data. *ACM Transactions on Graphics*, 43(1):1–14, 2023. 2

[31] Jose Luis Ponton, Eduard Pujol, Andreas Aristidou, Carlos Andujar, and Nuria Pelechano. Dragposer: Motion reconstruction from variable sparse tracking signals via latent space optimization. *arXiv preprint arXiv:2406.14567*, 2024. 2

[32] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 5

[33] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao. Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14761–14771, 2023. 2

[34] Zehong Shen, Zhi Cen, Sida Peng, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Learning human mesh recovery in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17038–17047, 2023. 3, 5

[35] Jiangnan Tang, Jingya Wang, Kaiyang Ji, Lan Xu, Jingyi Yu, and Ye Shi. A unified diffusion framework for scene-aware human motion estimation from sparse signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21251–21262, 2024. 2, 3, 6, 7

[36] Tom Van Wouwe, Seunghwan Lee, Antoine Falisse, Scott Delp, and C Karen Liu. Diffusionposer: Real-time human motion reconstruction from arbitrary sparse sensors using autoregressive diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2513–2523, 2024. 2

[37] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 4

[38] Tom Wehrbein, Marco Rudolph, Bodo Rosenhahn, and Bastian Wandt. Probabilistic monocular 3d human pose estimation with normalizing flows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11199–11208, 2021. 2

[39] Alexander Winkler, Jungdam Won, and Yuting Ye. Questsim: Human motion tracking from sparse sensors with simulated avatars. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–8, 2022. 1

[40] Yinghao Wu, chaoran wang, Lu Yin, Shihui Guo, and Yipeng Qin. Accurate and steady inertial pose estimation through sequence structure learning and modulation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2

[41] Songpengcheng Xia, Lei Chu, Ling Pei, Zixuan Zhang, Wenxian Yu, and Robert C Qiu. Learning disentangled representation for mixed-reality human activity recognition with a single imu sensor. *IEEE Transactions on Instrumentation and Measurement*, 70:1–14, 2021. 2

[42] Songpengcheng Xia, Lei Chu, Ling Pei, Jiarui Yang, Wenxian Yu, and Robert C Qiu. Timestamp-supervised wearable-based activity segmentation and recognition with contrastive learning and order-preserving optimal transport. *IEEE Transactions on Mobile Computing*, 2024. 1

[43] Yuan Xu, Xiaoxuan Ma, Jiajun Su, Wentao Zhu, Yu Qiao, and Yizhou Wang. Scorehypo: Probabilistic human mesh estimation with hypothesis scoring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 979–989, 2024. 2

[44] Haibiao Xuan, Xiongzheng Li, Jinsong Zhang, Hongwen Zhang, Yebin Liu, and Kun Li. Narrator: Towards natural control of human-scene interaction generation via relationship reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22268–22278, 2023. 3, 5

[45] Hongdi Yang, Chengyang Li, Zhenxuan Wu, Gaozheng Li, Jingya Wang, Jingyi Yu, Zhuo Su, and Lan Xu. Smgdiff: Soccer motion generation using diffusion probabilistic models. *arXiv preprint arXiv:2411.16216*, 2024. 3

[46] Xueyuan Yang, Chao Yao, and Xiaojuan Ban. Spatial-related sensors matters: 3d human motion reconstruction assisted with textual semantics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10225–10233, 2024. 3

[47] Hongwei Yi, Chun-Hao P Huang, Shashank Tripathi, Lea Hering, Justus Thies, and Michael J Black. Mime: Human-aware 3d scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12965–12976, 2023. 3, 5

[48] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions On Graphics (TOG)*, 40(4):1–13, 2021. 2

[49] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 13167–13178, 2022. 1, 2

[50] Siwei Zhang, Yan Zhang, Federica Bogo, Marc Pollefeys, and Siyu Tang. Learning motion priors for 4d human body capture in 3d scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11343–11353, 2021. 6

[51] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *European conference on computer vision*, pages 180–200. Springer, 2022. 6, 7

[52] Siwei Zhang, Qianli Ma, Yan Zhang, Sadegh Aliakbarian, Darren Cosker, and Siyu Tang. Probabilistic human mesh recovery in 3d scenes from egocentric views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7989–8000, 2023. 3, 6

[53] Yufei Zhang, Hanjing Wang, Jeffrey O Kephart, and Qiang Ji. Body knowledge and uncertainty modeling for monocular 3d human body reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9020–9032, 2023. 3

[54] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1889–1899, 2024. 2

[55] Xiaozheng Zheng, Zhuo Su, Chao Wen, Zhou Xue, and Xiaojie Jin. Realistic full-body tracking from sparse observations via joint-level modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14678–14688, 2023. 1, 2, 6, 7

[56] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, C Karen Liu, and Leonidas J Guibas. Gimo: Gaze-informed human motion prediction in context. In *European Conference on Computer Vision*, pages 676–694. Springer, 2022. 6