







# Motion Reconstruction via Human Anatomy Diffusion from Sparse Tracking

Zehai Niu<sup>1</sup>, Ke Lu<sup>1,2</sup>, Kun Dong<sup>1</sup>, Jian Xue<sup>\*1</sup>, Xiaoyu Qin<sup>3</sup>, and Jinbao Wang<sup>4</sup>

<sup>1</sup> University of Chinese Academy of Sciences, Beijing, China

<sup>2</sup> Peng Cheng Laboratory, Shenzhen, China  
`{niuzechai18,dongkun22}@mailsucas.ac.cn`  
`{luk,xuejian}@ucas.ac.cn`

<sup>3</sup> Tsinghua University, Beijing, China  
`xyqin@tsinghua.edu.cn`

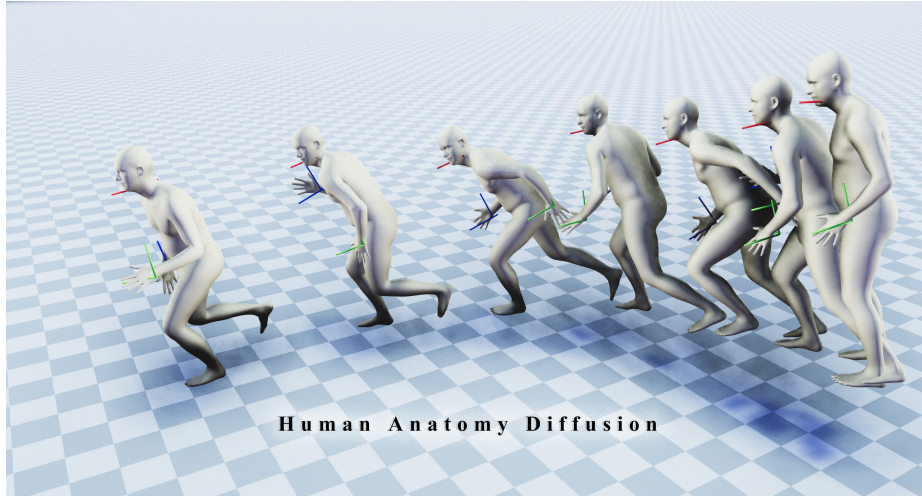
<sup>4</sup> Shenzhen University, Shenzhen, China  
`wangjb@szu.edu.cn`

**Abstract.** In the research field of analysis of people, generating precise full-body human motion from sparse tracking is a significant challenge. It is well known that diffusion techniques excel in generating high-quality two-dimensional (2D) visual content. However, when applied to human motion reconstruction, they might struggle to capture the inherent complexities of human motion, which is characterized by three-dimensional (3D) anatomical features and one-dimensional (1D) temporal dynamics. This heterogeneous structure between human motion and images can lead to accumulated errors at the joints, affecting the accuracy and smoothness of the generated motions. Building on this insight, we propose Human Anatomy Diffusion (HAD), a novel framework that integrates human anatomical features into the denoising process and excels in handling complex motions, accurately capturing body angles and balance, and showing enhanced alignment in motion prediction. HAD remarkably advanced the performance of motion reconstruction, notably enhancing smoothness by 81.29% compared to the previous state-of-the-art works and improving key accuracy metrics like MPJPE, Root PE, and Lower PE by approximately 20% on AMASS. Our method provides a crucial advancement for creating realistic and responsive virtual avatars in real-world applications. The project page is at: <https://niuzechai.github.io/had/>.

**Keywords:** Motion Reconstruction · Human Anatomy · Diffusion · Sparse Tracking

---

\* This work was supported by the National Natural Science Foundation of China (62320106007, 62032022, 62236006), the Guangdong Provincial Key Laboratory (2023B1212060076) and the Scientific Research Program of Beijing Municipal Education Commission (KZ201911417048). (Corresponding author: Jian Xue)



**Fig. 1:** Sequential visualization of full-body pose estimation using the HAD method from sparse tracking inputs. The red coordinate axes represent the Head-Mounted Display (HMD) tracking, while the green and blue coordinate axes correspond to the left and right hand tracking, respectively. The Human Anatomy Diffusion represents a full-body motion capture technology applicable to VR or AR environments.

## 1 Introduction

In the contemporary field of artificial intelligence research, a critical challenge lies in the nuanced understanding and accurate simulation of human motion and behavior. Specifically, in the domain of motion generation, research focusing on deriving human motion from various modalities, such as text-to-motion [17, 20, 34] and audio-to-motion [4, 12, 31], has attracted considerable interest. However, the human motions generated by these methods often fail to meet the user’s expectations, especially regarding accuracy. Methods that utilize RGB cameras [14, 24, 37] or single head-mounted fisheye cameras [27–29] are particularly susceptible to influence from occlusion and truncation, which can compromise the smoothness and authenticity of the generated human motion. As a result, these methods may not fully align with the natural motions people anticipate.

Recently, more researchers have focused on synthesizing full-body motion poses from sparse tracking inputs. In the research field of analysis of people, particularly in cognition and interaction domains, the utilization of Head-Mounted Displays (HMDs) and other spatial computing devices, such as AR and VR, for motion reconstruction is gaining increasing traction [11]. Compared to traditional Inertial Measurement Unit (IMU) motion capture [22] and methods based on six IMUs [32, 33, 36], capturing full-body posture with fewer trackers offers a less invasive and more cost-effective solution. While spatial computing devices inherently track human and hand rotations and positions, they often lack comprehensive capture of full-body motion, particularly in the lower body.

Recent advances in motion reconstruction based on sparse tracking, such as the diffusion-based techniques, such as AGRoL [7], have made significant contributions. However, these methods lack the ability to fully encompass the complex dynamics of human motions. In particular, they often overlook the anatomical structure of the human body and the causal, one-dimensional temporal nature of human movement. This oversight can lead to cumulative errors at joints and diminish the accuracy and smoothness of the generated motions. Therefore, there is a pressing need for more advanced methods that can accurately capture both the anatomical and temporal characteristics of human motion.

Distinct from previous methods, our approach overcomes the aforementioned issues of joint errors and unnatural motions seen in previous methodologies. By focusing on the anatomical and temporal aspects of human movement, the proposed method significantly improves upon the current state-of-the-art methods in both accuracy and smoothness of motions. Building on these foundations, our contributions to the field are threefold:

- A powerful motion reconstruction network named Human Anatomy Network (HAN) is proposed, which uses a hierarchical structure to process various body segments, enhancing motion dynamics representation. It leverages techniques like Latent Space Mapping, Iterative Feature Enhancement, and Temporal Feature Pyramid, coupled with Hierarchical Motion Refinement, to significantly improve motion prediction and refinement.
- Based on HAN, we further introduce a novel diffusion-based framework, Human Anatomy Diffusion (HAD). It integrates human anatomical features into the denoising process of motion generation. This novel method effectively captures the complex dynamics and structural intricacies of human motions.
- Extensive experiments have been conducted on the widely-used AMASS benchmark. The results show that our proposed method can boost motion smoothness by 81.29% and accuracy in key metrics by around 20%, showing potential for applications in realistic, interactive VR and AR scenarios.

## 2 Related Work

**Full-Body Pose Tracking from Sparse Motion Sensing.** Generating full-body posture from sparse tracking signals of body joints has become a field of considerable interest in the research community. Lots of previous work on this field such as [10, 26, 32] has used up to 6 body-worn inertial sensors, which are commonly distributed over head, arms, pelvis and legs, making motion capture inflexible and clumsy. To overcome this limitation, CoolMoves [1] was first to estimate full body posture using only 3 tracking signals from headphones and handheld controllers. However, the proposed KNN-based method interpolates poses from a smaller dataset with only specific motion, resulting in high errors when applied to larger benchmarks with diverse subjects and activities. Recently, AvatarPoser [11] used a transformer-based architecture to solve the

3-point problem. DAP [5] proposed a dual-path attention scheme to extract features of the sparse signals. Other methods considered tracking the full body from sparse inputs as a conditional generation problem. For instance, Dittadi et al. [6] proposed a Variational Autoencoder (VAE) method, which encodes all joints relative to the pelvis. However, it implicitly takes knowledge of the pelvis as the fourth input position, leaving the highly ill-posed problem with only three inputs unsolved.

**Denoising diffusion models.** Starting from the field of image generation, denoising diffusion models [2, 9, 16, 19, 23] have gained significant attention due to their ability to produce high-quality results and is better suited for handling large amounts of data. They learn a probabilistic model over a denoising process on inputs, which is supervised to gradually denoise a Gaussian noise to a target output. Moreover, diffusion models can support conditional generation. For instance, ILVR [3] guided the generative process in DDPM [9] to generate high-quality images based on a given reference image. GLIDE [18] explored diffusion models for the problem of text-conditional image synthesis. Recent advance has extended diffusion models to motion synthesis [13, 25, 35]. However, these models particularly focused on the text-to-motion task, with little attention paid to tracking the full body from sparse inputs. AGRoL [7] presented the first diffusion model solely purposed for solving motion reconstruction from sparse inputs. However, it did not fully consider the complex hierarchical structure of human anatomy. Our method, Human Anatomy Diffusion, innovatively overcomes this issue and thus generates more accurate and smoother human motions.

### 3 Methodology

In this paper, we focus on the task of full-body motion reconstruction from sparse tracking inputs, i.e. the positional and rotational data from a headset and two handheld controllers. This section starts by introducing the formulation briefly in Section 3.1 and then describes the detailed architecture of the Human Anatomy Network (HAN) for motion reconstruction in Section 3.2, based on which we further introduce the proposed Human Anatomy Diffusion (HAD) in Section 3.3. Finally, we introduce the adopted loss function of our method in Section 3.4.

#### 3.1 Formulation

A sequence of sparse human motion containing  $N$  frames can be represented by  $\mathbf{c} = \{\mathbf{c}^i\}_i^N \in \mathbb{R}^{N \times C}$ , where  $C$  represents the dimensions of the observed joint features. The goal is to predict the corresponding full-body poses  $\mathbf{x} = \{\mathbf{x}^i\}_i^N \in \mathbb{R}^{N \times K}$ , where  $K$  represents the dimensions of the full-body joint features. We adopt the SMPL [15] model to represent human poses and follow the practice of disregarding the joints on the hands and face in [6, 7, 11].

In our method, we consider the motion prediction task as a conditional generation problem, where the sparse tracking serves as the conditioner. Specifically,

a novel diffusion model detailed in Section 3.2 and Section 3.3 is employed to generate full-body poses. Assume  $p_{motion}$  to be the distribution of the full-body poses in the dataset, in the forward diffusion process, we gradually add Gaussian noise into the clean data distribution  $p_{motion}$  until the output distribution is close to an isotropic Gaussian distribution. Taking a sample data pair  $(\mathbf{x}_0, \mathbf{c})$  as an example, the diffusion process from clean data  $\mathbf{x}_0$  to  $\mathbf{x}_T$  is defined as

$$q(\mathbf{x}_1, \dots, \mathbf{x}_T | \mathbf{x}_0) = \prod_{t=1}^T q(\mathbf{x}_t | \mathbf{x}_{t-1}), \quad (1)$$

where  $q(\mathbf{x}_t | \mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t} \mathbf{x}_{t-1}, (1 - \alpha_t) \mathbf{I})$ , the hyperparameters  $\alpha_t$  are predefined positive constants, and  $\mathbf{x}_T$  tends to an isotropic Gaussian distribution when  $T \rightarrow \infty$ . The reverse diffusion process is conditioned on the sparse tracking inputs  $\mathbf{c}$ . We train a diffusion model  $p_\theta$  to predict and eliminate the noise added in the diffusion process. Let  $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$  be a latent variable. The reverse process from latent  $\mathbf{x}_T$  to clean data  $\mathbf{x}_0$  is defined as

$$p_\theta(\mathbf{x}_0, \dots, \mathbf{x}_{T-1} | \mathbf{x}_T, \mathbf{c}) = \prod_{t=1}^T p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{c}), \quad (2)$$

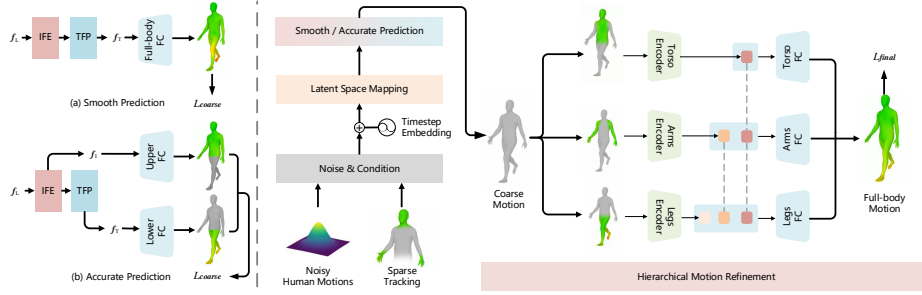
where  $p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{c}) := \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, \mathbf{c}, t), \sigma_t^2 \mathbf{I})$ . The mean  $\mu_\theta(\mathbf{x}_t, \mathbf{c}, t)$  is a neural network parameterized by  $\theta$  and the variance  $\sigma_t^2$  is a time-step dependent constant. The parameterization is  $\sigma_t^2 = \frac{1 - \bar{\alpha}_t}{1 - \alpha_t} (1 - \alpha_t)$ ,  $\mu_\theta(\mathbf{x}_t, \mathbf{c}, t) = \frac{1}{\sqrt{\alpha_t}} (\mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, \mathbf{c}, t))$ , where  $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$ . Following [21], we directly predict the clean body poses  $\mathbf{x}_0$  instead of predicting the residual noise  $\epsilon_\theta(\mathbf{x}_t, \mathbf{c}, t)$ . Therefore, the objective function in training can be formulated as

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{x}_0 \sim p_{motion}} \left[ \|\mathbf{x}_0 - \hat{\mathbf{x}}_0\|_2^2 \right], \quad (3)$$

where the  $\hat{\mathbf{x}}_0$  denotes the output of the proposed diffusion model.

### 3.2 Human Anatomy Network

As illustrated in Figure 2, the proposed Human Anatomy Network(HAN) consists of four components: Latent Space Mapping (LSM), Iterative Feature Enhancement (IFE), Temporal Feature Pyramid (TFP) and Hierarchical Motion Refinement (HMR). In particular, LSM maps the input features into a unified latent space for subsequent processing. Then IFE takes the coarse features from LSM as input and enhances them for better representation. To further improve the smoothness and coherence of the generated motions, MTP is applied to integrate multi-scale temporal motion features and output coarse predictions. Finally, inspired by the human body structure, HMR adopts a hierarchical architecture to refine the obtained predictions for more natural and accurate motion. The detailed architectures of each module are as follows.



**Fig. 2:** The architecture of Human Anatomy Network (HAN). Taking noisy human motions  $\mathbf{x}_t$  and sparse tracking  $\mathbf{c}$  as input, we adopt a dual-path approach for motion generation. In particular, smooth Prediction (SP) focuses on motion smoothness, whereas Accuracy Prediction (AP) emphasizes spatial accuracy, with both paths converging to refine full-body motion hierarchically.

**Latent Space Mapping:** At time step  $t$ , fully connected layers are first applied to map the noisy motion poses  $\mathbf{x}_t \in \mathbb{R}^{N \times K}$  and the observed joint features  $\mathbf{c} \in \mathbb{R}^{N \times C}$  to a unified latent space. Then we can obtain the corresponding latent features  $\mathbf{f}_x \in \mathbb{R}^{N \times D}$  and  $\mathbf{f}_c \in \mathbb{R}^{N \times D}$ , where  $D$  denotes the dimension of latent features.

Subsequently, the diffusion step  $t$  is transformed into a step embedding vector  $\mathbf{f}_t$ , which is concatenated with  $\mathbf{f}_x$  and  $\mathbf{f}_c$  and input into the BasicBlock for further processing. Note that  $\mathbf{f}_t$  will be injected repetitively into each BasicBlock to prevent the information loss of time step embedding [7]. We denote the output of this step as  $\mathbf{f}_L$ . The architecture of BasicBlock is described in detail in the supplemental material.

**Iterative Feature Enhancement:** Due to the insufficient interaction of features for each frame in LSM, we employ IFE to enhance the motion features  $\mathbf{f}_L$  to obtain more robust representation, noted as  $\mathbf{f}_I$ . In particular, we adopt an iterative approach to gradually enhance the motion features. The number of iterations is noted as  $M$  and the output of the  $j$ -th iteration is noted as  $\mathbf{f}_I^j$ . Taking the  $j$ -th iteration as an example, the refinement in this iteration can be mathematically expressed as:

$$\mathbf{f}_I^j = \mathbf{f}_I^{j-1} + \text{BasicBlock}(\mathbf{f}_I^{j-1}), \quad (4)$$

where  $\mathbf{f}_I^0$  is initialized as  $\mathbf{f}_L$ .

Upon completing all iterations, the final output is  $\mathbf{f}_I = \mathbf{f}_I^M$ , which represents the enhanced motion features.

This iterative refinement process, facilitated through the BasicBlock, is essential for achieving precise and reliable full-body motion predictions. Experiments in Section 4.5 have demonstrated the vital importance of this refinement.

**Temporal Feature Pyramid:** Considering that distinct actions correspond to different time scales, we apply TFP to further improve the smoothness of the predictions.

Let  $\mathcal{S} = \{s_1, \dots, s_L\}$  be the set of  $L$  downsampling multiples. For the  $i$ -th scale factor  $s_i \in \mathbb{N}$ ,  $f_I$  is first downsampled by a scale factor of  $s_i$ . Then a BasicBlock is applied to process these downsampled features and output intermediate features  $\tilde{f}_T^i \in \mathbb{R}^{\frac{N}{s_i} \times D}$ . Subsequently, we utilize MLP to increase the sequence length of  $\tilde{f}_T^i$  back to  $N$  and the corresponding features are noted as  $f_T^i$ .

To aggregate these multi-scale outputs, the final output  $f_T$  is the normalized sum of the features from each scale:

$$f_T = \frac{1}{L} \sum_i^L f_T^i. \quad (5)$$

By aid of this multi-scale refinement process, we transition  $f_I$  into more robust motion features  $f_T$ , which effectively capture motion dynamics across different temporal resolutions and enhance the smoothness of full body. We denote the approach of directly using  $f_T$  for prediction as **Smooth Prediction** (SP). However, in the experiment we found that SP tends to cause excessive smoothing of upper body pose, leading to suboptimal accuracy. Therefore, we propose **Accurate Prediction** (AP) to ensure the accuracy of predictions. As illustrated in Figure 2(b),  $f_I$  and  $f_T$  are utilized to predict the upper body pose  $\hat{\mathbf{x}}_{t-1}^{upper}$  and lower body pose  $\hat{\mathbf{x}}_{t-1}^{lower}$ , respectively. Then we merge  $\hat{\mathbf{x}}_{t-1}^{upper}$  and  $\hat{\mathbf{x}}_{t-1}^{lower}$  to get the full body posture  $\hat{\mathbf{x}}_{t-1}$ . Given that  $f_I$  contains accurate upper body information while  $f_T$  captures information from different time scales, the separated prediction method ensures accuracy while also considering smoothness. The influence of two approaches will be discussed in Section 4.5.

**Hierarchical Motion Refinement:** Inspired by the human body structure, we adopt a hierarchical approach for motion refinement. As shown in Figure 2, we begin by partitioning the coarse predictions  $\hat{\mathbf{x}}_{t-1} \in \mathbb{R}^{N \times K}$  into three distinct parts:  $\hat{\mathbf{x}}_{t-1}^{torso} \in \mathbb{R}^{N \times K_t}$ ,  $\hat{\mathbf{x}}_{t-1}^{arms} \in \mathbb{R}^{N \times K_a}$  and  $\hat{\mathbf{x}}_{t-1}^{legs} \in \mathbb{R}^{N \times K_l}$ , where  $K = K_t + K_a + K_l$ . Each part corresponds to different positions, i.e. torso, arms and legs. Then we encode the three parts separately:

$$f_{t-1}^{torso} = \text{TorsoEncoder}(\hat{\mathbf{x}}_{t-1}^{torso}), \quad (6)$$

$$f_{t-1}^{arms} = \text{ArmsEncoder}(\hat{\mathbf{x}}_{t-1}^{arms}), \quad (7)$$

$$f_{t-1}^{legs} = \text{LegsEncoder}(\hat{\mathbf{x}}_{t-1}^{legs}). \quad (8)$$

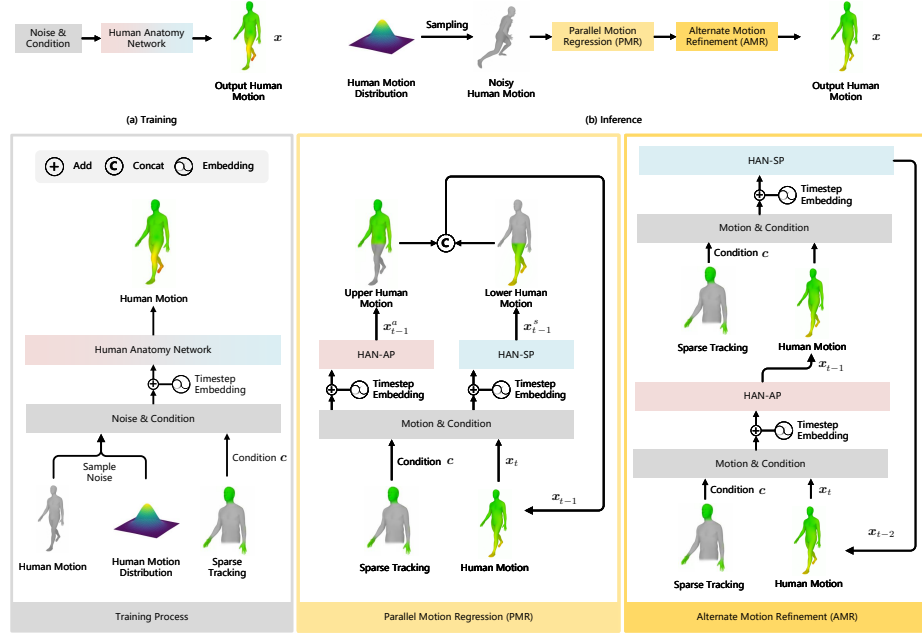
The architectures of the three encoders are the same. In particular, we use linear layer, BasicBlock and linear layer in turn to extract the motion features. Subsequently, we apply a hierarchical approach to refine the three body parts, respectively. The process is represented as:

$$\mathbf{x}_{t-1}^{torso} = \text{FC}(f_{t-1}^{torso}), \quad (9)$$

$$\mathbf{x}_{t-1}^{arms} = \text{FC}(\text{Concat}(f_{t-1}^{torso}, f_{t-1}^{arms})), \quad (10)$$

$$\mathbf{x}_{t-1}^{legs} = \text{FC}(\text{Concat}(f_{t-1}^{torso}, f_{t-1}^{arms}, f_{t-1}^{legs})), \quad (11)$$

where FC means fully connected layers. Finally, we merge  $\mathbf{x}_{t-1}^{torso}$ ,  $\mathbf{x}_{t-1}^{arms}$  and  $\mathbf{x}_{t-1}^{legs}$  to obtain the well-refined full-body poses  $\mathbf{x}_{t-1}$ .



**Fig. 3:** Schematic representation of Human Anatomy Diffusion. This framework integrates Human Anatomy Network with diffusion-based motion prediction, employing Parallel Motion Regression (PMR) and Alternate Motion Refinement (AMR) for generating precise human motion.

This hierarchical and anatomically aware approach allows for a more nuanced representation of human motion. By treating each body part separately and then recombining them, we achieve a more comprehensive and realistic depiction of the full-body motion.

### 3.3 Human Anatomy Diffusion

As mentioned in the previous section, Smooth Prediction (SP) is beneficial for smooth motion prediction but will lead to suboptimal accuracy for the upper body. On the contrary, Accurate Prediction (AP) can generate more accurate predictions but cannot ensure the smoothness of motion, especially in the lower body.

To address the limitations of using a single approach, we propose a novel architecture, called Human Anatomy Diffusion (HAD), which is based on the Human Anatomy Network (HAN). As shown in Figure 3, it utilizes AP and SP simultaneously to generate both accurate and smooth predictions. Specifically, the proposed architecture consists of two phase: Parallel Motion Regression (PMR) and Alternate Motion Refinement (AMR). Following are the detailed architectures of these two phases.



**Parallel Motion Regression:** We denote the HAN using SP and AP for prediction as HAN-SP ( $m_s$ ) and HAN-AP ( $m_a$ ), respectively. Considering the advantages of each method, we adopt a dual path approach to predict the upper and lower parts of the human body separately.

In particular, the noisy motion poses  $\mathbf{x}_t$ , the observed joint features  $\mathbf{c}$  and the time step  $t$  are input into  $m_s$  and  $m_a$ , respectively. Then we can get two corresponding outputs  $\mathbf{x}_{t-1}^s$  and  $\mathbf{x}_{t-1}^a$ , which are formulated as:

$$p(\mathbf{x}_{t-1}^s | \mathbf{x}_t, \mathbf{c}) := \mathcal{N}(\mathbf{x}_{t-1}^s; m_s(\mathbf{x}_t, \mathbf{c}, t), \sigma_t^2 I), \quad (12)$$

$$p(\mathbf{x}_{t-1}^a | \mathbf{x}_t, \mathbf{c}) := \mathcal{N}(\mathbf{x}_{t-1}^a; m_a(\mathbf{x}_t, \mathbf{c}, t), \sigma_t^2 I), \quad (13)$$

where  $\mathbf{x}_{t-1}^s$  and  $\mathbf{x}_{t-1}^a$  are both full-body poses. Then we extract poses  $\mathbf{x}_{t-1}^{lower}$  and  $\mathbf{x}_{t-1}^{upper}$  of the upper and lower bodies separately, which are subsequently merged to form the full-body poses  $\mathbf{x}_{t-1}$ .

This stage is crucial for generating accurate and smooth predictions, paving the way for a more nuanced and realistic synthesis of human movement.

**Alternate Motion Refinement:** To further improve the performance of the network, we employ a sequential method to refine  $\mathbf{x}_t$ .

Distinct from the parallel approach adopted by PMR, at this phase, we gradually refine motion predictions by applying  $m_a$  and  $m_s$  in sequence.

Specifically, the sequential refinement process is defined as follows:

$$p(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{c}) := \mathcal{N}(\mathbf{x}_{t-1}; m_a(\mathbf{x}_t, \mathbf{c}, t), \sigma_t^2 I), \quad (14)$$

$$p(\mathbf{x}_{t-2} | \mathbf{x}_{t-1}, \mathbf{c}) := \mathcal{N}(\mathbf{x}_{t-2}; m_s(\mathbf{x}_{t-1}, \mathbf{c}, t-1), \sigma_{t-1}^2 I). \quad (15)$$

AMR meticulously refines the motion predictions through a serialization process. In the experiment, this phase is of vital importance for enhancing the accuracy and naturalness of human motion synthesis.

### 3.4 Loss Function

In training, the loss function adopted for HAN consists of two components: Intermediate Supervision (IS) and Final Supervision (FS), both of which are calculated in the same way as Equation 3.

To be concrete, the IS is utilized to minimize the difference between the coarse motion predictions and ground truth. Therefore, the total loss can be computed as

$$\mathcal{L}_{total} = \mathcal{L}_{coarse} + \mathcal{L}_{final}, \quad (16)$$

where  $\mathcal{L}_{coarse}$  is the introduced IS, while  $\mathcal{L}_{final}$  is the  $L_2$  distance between final predictions and ground truth.

For inference, two pretrained HANs ( $m_s$  and  $m_a$ ) are applied to construct HAD. It is worth noting that we did not retrain HAD but directly used the parameters of  $m_s$  and  $m_a$ . Assuming  $R$  steps are used in the reverse diffusion process, we assign  $R_p$  and  $R_a$  steps to PMR and AMR, respectively. Note that  $R = R_p + R_a$ .

## 4 Experiments

### 4.1 Datasets

Our research primarily employs the AMASS dataset for training and evaluation, in line with current standards in the field. Following [11], we concentrate on the CMU, BMLr, and HDM05 subsets. Human poses are represented using the SMPL model [15], with a focus on the root joint’s global orientation and the relative rotations of other joints.

### 4.2 Metrics

The key metrics for evaluation can be divided into three categories: 1) Rotation-Oriented Metrics, with Mean Per Joint Rotation Error (MPJRE) in degrees for rotational accuracy; 2) Velocity-Oriented Metrics, featuring Mean Per Joint Velocity Error (MPJVE) in  $cm/s$  [11] and Jitter [8] in  $10^2 m/s^3$  for joint velocity and motion smoothness; and 3) Position-Oriented Metrics, assessing spatial accuracy through Mean Per Joint Position Error (MPJPE) in centimeters, and specific errors like Root PE, Hand PE, and Upper and Lower PE for different body parts [7].

### 4.3 Implementation Details

We represent the joint rotations by the 6D reparametrization due to its simplicity and continuity. Therefore, for the sequences of body poses  $\mathbf{x} \in \mathbb{R}^{N \times K}$ ,  $K = 22 \times 6$ . The observed joint features  $\mathbf{c} \in \mathbb{R}^{N \times C}$  consist of the orientation, translation, orientation velocity and translation velocity of the head and hands in global coordinate system. Additionally, we adopt 6D reparametrization for the orientation and orientation velocity, thus  $C = 18 \times 3$ . Unless otherwise stated, we set the frame number  $N$  to 196. We used PyTorch as the deep learning framework and trained our model on a computer with an Intel i9-9900K CPU and an NVIDIA GTX3090 GPU. In addition, our network used Adam optimizer with a base learning rate of 0.0003.

### 4.4 Comparison with State-of-the-Art Methods

As shown in Table 1, the proposed HAD demonstrates remarkable advancements over the current state-of-the-art technique, AGRoL [7]. Key performance metrics from the AMASS dataset show HAD’s superiority: a reduction in MPJRE by 14.29%, an improvement in MPJPE by 18.87%, and a 16.46% betterment in MPJVE. It is worth noting that HAD has improved the motion accuracy of the hands, upper body, and lower body by 8.4%, 15.48%, and 19.88%, respectively. The improvement in overall jitter compared to the real motion (GT) and the previous state-of-the-art (SOTA) is approximately 81.29%, while the improvement in upper body jitter is about 90.58%.

**Table 1:** Performance comparison of our methods with SOTA approaches on AMASS.

Method	MPJRE ↓	MPJPE ↓	MPJVE ↓	Hand PE ↓	Upper PE ↓	Lower PE ↓	Root PE ↓	Jitter ↓	Upper Jitter ↓	Lower Jitter ↓
LoBSTr [30]	10.69	9.02	44.97	-	-	-	-	-	-	-
CoolMoves [1]	5.20	7.83	100.54	-	-	-	-	-	-	-
VAE-HMD [6]	4.11	6.83	37.99	-	-	-	-	-	-	-
AvatarPoser [11]	3.08	4.18	27.70	2.12	1.81	7.59	3.34	14.49	7.36	24.81
DAP [5]	2.69	3.68	24.03	-	-	-	-	-	-	-
AGRoL-MLP [7]	2.69	3.93	22.85	2.62	1.89	6.88	3.35	13.01	9.13	18.61
AGRoL [7]	2.66	3.71	18.59	1.31	1.55	6.84	3.36	7.26	5.88	9.27
HAN-SP (Ours)	2.41	3.31	16.59	1.75	1.50	5.91	2.87	4.69	3.93	5.78
HAN-AP (Ours)	2.40	3.18	16.42	1.15	1.37	5.79	2.90	7.35	5.66	9.79
HAD (Ours)	<b>2.29</b>	<b>3.03</b>	<b>15.45</b>	<b>1.15</b>	<b>1.30</b>	<b>5.52</b>	<b>2.71</b>	<b>4.61</b>	<b>3.86</b>	<b>5.70</b>
GT	0	0	0	0	0	0	0	4.00	3.65	4.52

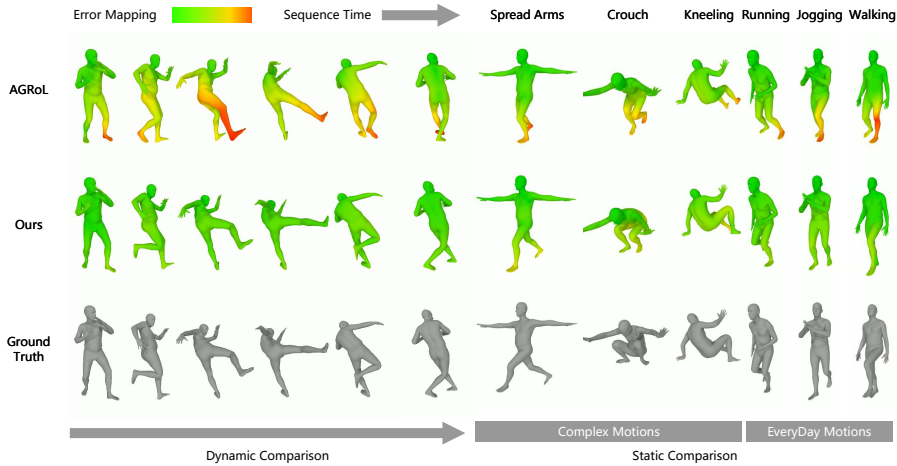
**Fig. 4:** Visual comparison of motion prediction accuracy between the AGRoL method, our method, and the ground truth for dynamic and static motions. The error mapping colors range from green, indicating high accuracy, to red, signifying greater errors.

Figure 4 presents a visual comparison across different methods. Our qualitative assessment emphasizes two critical aspects: dynamic and static motion comparisons. Through a series of colors ranging from green to red, we illustrate the varying precision of predictions, with redder indicating the larger error. In dynamic comparisons, our HAD achieves remarkably precise full-body predictions from merely three sparse tracking points, even during vigorous movements such as kicks. The static comparison encompasses complex actions like Spread Arms, Crouch, and Kneeling, as well as common activities such as Running, Jogging, and Walking. HAD can capture subtle body angles and balance nuances, especially in complex stances and rapid movements, which gives it the excellent ability for realistic motion capture. The dynamic analysis, as depicted in Figure 5, encompasses various dynamic activities, including complex motions, physical exercises, swimming, and soccer. Our method distinguishes itself by capturing the intricacies of these movements more accurately, achieving a closer resemblance to ground truth than the AGRoL method.



**Fig. 5:** Comparative visualization of dynamic motion prediction between the AGRoL method (top), our proposed approach (middle), and the ground truth data (bottom). Our method exhibits more accurate pose estimation, particularly in dynamic motions, as evidenced by its closer resemblance to the ground truth.

In addition to high-quality predictions, the proposed HAD (running on GTX 3090 GPU) also demonstrates extraordinary efficiency. It owns an average processing time of just 0.33 ms per pose, translating to a frame rate of 3014 FPS. We believe that further leaps in speed can be achieved with model optimization. This high performance suggests HAD has the potential for real-time HMD-driven avatar animation in realistic applications.

#### 4.5 Ablation Studies

In this section, we conduct extensive ablation studies for our Human Anatomy Diffusion (HAD) architecture, which is structured into three core components: the Motion Initialization Network, the Human Anatomy Network, and the Human Anatomy Diffusion process.

**Table 2:** Ablation study on the Human Anatomy Network’s architecture and the influence of individual loss terms, with Intermediate Supervision (w/ IS) and without Intermediate Supervision (w/o IS). The study examines the effect of LSM, IFE, TFP, and HMR components on joint position (MPJPE) and motion smoothness (Jitter).

Method	LSM	IFE	TFP	HMR	MPJPE ↓	Jitter ↓
LSM	✓				3.75	8.62
LSM + IFE	✓	✓			3.42	14.75
LSM + TFP	✓		✓		3.59	5.88
LSM + IFE + TFP (SP)	✓	✓	✓		3.33	5.88
LSM + IFE + TFP (AP)	✓	✓	✓		3.30	8.51
HAN-SP (w/o IS)	✓	✓	✓	✓	3.48	6.66
HAN-SP (w IS)	✓	✓	✓	✓	3.31	<b>4.69</b>
HAN-AP (w/o IS)	✓	✓	✓	✓	3.31	7.99
HAN-AP (w IS)	✓	✓	✓	✓	<b>3.18</b>	7.35

**Table 3:** Ablation results for Human Anatomy Diffusion components, detailing their impact on MPJRE, MPJPE, MPJVE, and Jitter metrics.

Method	PMR	AMR	MPJRE ↓	MPJPE ↓	MPJVE ↓	Jitter ↓
PMR	✓		2.40	3.19	17.48	8.42
AMR		✓	2.32	3.10	16.50	8.10
HAD	✓	✓	<b>2.29</b>	<b>3.03</b>	<b>15.45</b>	<b>4.61</b>

**Architecture of the Human Anatomy Network.** As show in Table 2, we evaluate the impact of each component on HAN. LSM serves as a key encoding strategy, setting the stage for further enhancements. Adding IFE to LSM improves MPJPE, but increases motion jitter. Incorporating the TFP with LSM significantly enhances motion smoothness, reducing jitter effectively. To balance accuracy and smoothness, we develop two specialized predictions: LSM + IFE + TFP (SP) for smoothness and LSM + IFE + TFP (AP) for spatial accuracy. Integrating the Hierarchical Motion Refinement (HMR) module into these predictions results in HAN-SP and HAN-AP. With intermediate supervision (w/ IS), HAN-SP achieves the lowest jitter (4.69), and HAN-AP attains the best MPJPE (3.18), demonstrating the effectiveness of structured refinements. For a comprehensive overview of all metrics and their detailed analysis, readers are directed to the Supplemental Material.

**Two phases of Human Anatomy Diffusion.** In the ablation study highlighted in Table 3, PMR alone achieved an MPJRE reduction to 2.40 and a Jitter decrease to 8.42, showcasing its effectiveness in improving motion accuracy and smoothness. When AMR was applied, it further reduced MPJRE to 2.32 and Jitter to 8.10, enhancing the model’s performance. The integration of both PMR and AMR in the HAD framework resulted in the most significant improvements, with the lowest MPJRE of 2.29 and Jitter of 4.62, indicating a marked enhancement in both accuracy and the smoothness of generated motions.

**Number of Sampling Steps during Human Anatomy Diffusion Inference.** As summarized in Table 4, using 25% PMR in 8 sampling steps can

**Table 4:** Ablation results for Percentage of PMR and Sampling Step, detailing their impact on MPJRE, MPJPE, MPJVE, and Jitter metrics.

Percentage of PMR Sampling Steps		$R$	MPJRE ↓	MPJPE ↓	MPJVE ↓	Jitter ↓
25%	8		<b>2.29</b>	<b>3.03</b>	<b>15.45</b>	<b>4.61</b>
50%	8		2.29	3.03	15.49	4.63
75%	8		2.31	3.08	15.73	4.66
25%	2		2.58	3.68	17.22	7.04
25%	5		2.31	3.07	15.55	4.61
25%	8		<b>2.29</b>	<b>3.03</b>	<b>15.45</b>	<b>4.61</b>
25%	10		2.30	3.04	15.84	7.19
25%	100		2.33	3.05	16.48	7.32
25%	1000		2.36	3.10	16.58	4.98

**Table 5:** Results of cross-dataset evaluation between different methods.

Method	Dataset	MPJRE ↓	MPJPE ↓	MPJVE ↓	Hand PE ↓	Upper PE ↓	Lower PE ↓	Root PE ↓	Jitter ↓
AGRoL [7]	BMLrub	2.33	2.99	17.95	1.07	1.29	5.45	2.65	7.25
	CMU	3.11	4.71	19.31	1.67	1.92	8.73	4.38	7.45
	HDM05	2.96	4.43	20.92	1.98	1.84	8.17	3.87	7.66
HAD	BMLrub	<b>2.04</b>	<b>2.60</b>	<b>15.49</b>	<b>0.95</b>	<b>1.11</b>	<b>4.75</b>	<b>2.26</b>	<b>5.18</b>
	CMU	<b>2.67</b>	<b>3.65</b>	<b>15.39</b>	<b>1.40</b>	<b>1.57</b>	<b>6.65</b>	<b>3.37</b>	<b>3.91</b>
	HDM05	<b>2.26</b>	<b>3.06</b>	<b>15.39</b>	<b>1.40</b>	<b>1.34</b>	<b>5.54</b>	<b>2.69</b>	<b>3.40</b>

achieve the best performance, with the lowest MPJRE and jitter of 2.29 and 4.61, respectively. This configuration also yields the best MPJPE and MPJVE. Additionally, too many sampling steps can’t improve accuracy continuously and will result in higher computational load.

**Cross-dataset evaluation.** In the cross-dataset evaluation detailed in Table 5, HAD outshines AGRoL, showing marked improvements in complex motion datasets HDM05 and CMU, with MPJPE and MPJVE reductions of 30.93% and 26.40% on HDM05, and 22.51% and 20.48% on CMU, respectively. The quantitative analysis shown in Figure 4 and our comprehensive cross-dataset evaluations compellingly validate our method’s effectiveness in accurately predicting complex motions.

## 5 Conclusion and Limitation

This paper introduces Human Anatomy Diffusion (HAD), an innovative approach for generating human motion from sparse tracking. HAD significantly advances the state-of-the-art by integrating human anatomy into the denoising diffusion process. Our method fully considers the complex dynamics and structural intricacies of human motion, leading to substantial improvements in smoothness and accuracy metrics. The proposed architecture enables HAD to generate more accurate and natural human motions, demonstrating its potential in virtual reality and interactive applications. Despite the significant progress, our method still faces challenges with tracking loss, sequence delay, and complex samples, which will be crucial directions for future research.

## References

1. Ahuja, K., Ofek, E., Gonzalez-Franco, M., Holz, C., Wilson, A.D.: Coolmoves: User motion accentuation in virtual reality. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **5**(2), 1–23 (2021)
2. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 18000–18010 (2023)
3. Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938* (2021)
4. Dabral, R., Mughal, M.H., Golyanik, V., Theobalt, C.: Mofusion: A framework for denoising-diffusion-based motion synthesis. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 9760–9770 (2023)
5. Di, X., Dai, X., Zhang, X., Chen, X.: Dual attention poser: Dual path body tracking based on attention. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. pp. 2795–2804. IEEE (2023)
6. Dittadi, A., Dziadzio, S., Cosker, D., Lundell, B., Cashman, T.J., Shotton, J.: 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*. pp. 11687–11697 (2021)
7. Du, Y., Kips, R., Pumarola, A., Starke, S., Thabet, A., Sanakoyeu, A.: Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 481–490 (2023)
8. Flash, T., Hogan, N.: The coordination of arm movements: an experimentally confirmed mathematical model. *The Journal of Neuroscience* (1985)
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020)
10. Huang, Y., Kaufmann, M., Aksan, E., Black, M.J., Hilliges, O., Pons-Moll, G.: Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics (TOG)* **37**(6), 1–15 (2018)
11. Jiang, J., Streli, P., Qiu, H., Fender, A., Laich, L., Snape, P., Holz, C.: Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In: *European Conference on Computer Vision*. pp. 443–460. Springer (2022)
12. Jin, Z., Wang, Z., Wang, Q., Jia, J., Bai, Y., Zhao, Y., Li, H., Wang, X.: Holosinger: Semantics and music driven motion generation with octahedral holographic projection. In: *Proceedings of the 31st ACM International Conference on Multimedia*. pp. 9393–9395 (2023)
13. Kim, J., Kim, J., Choi, S.: Flame: Free-form language-based motion synthesis & editing. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 37, pp. 8255–8263 (2023)
14. Lin, J., Zeng, A., Wang, H., Zhang, L., Li, Y.: One-stage 3d whole-body mesh recovery with component aware transformer. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 21159–21168 (2023)
15. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. In: *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. pp. 851–866 (2023)
16. Lou, Y., Zhu, L., Wang, Y., Wang, X., Yang, Y.: Diversemotion: Towards diverse human motion generation via discrete diffusion. *arXiv preprint arXiv:2309.01372* (2023)

17. Lu, S., Chen, L.H., Zeng, A., Lin, J., Zhang, R., Zhang, L., Shum, H.Y.: Humantomato: Text-aligned whole-body motion generation. arXiv preprint arXiv:2310.12978 (2023)
18. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
19. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. pp. 8162–8171. PMLR (2021)
20. Petrovich, M., Black, M.J., Varol, G.: TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In: International Conference on Computer Vision (ICCV) (2023)
21. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 1(2), 3 (2022)
22. Roetenberg, D., Luinge, H., Slycke, P.J.: Xsens mvn: Full 6dof human motion tracking using miniature inertial sensors (2009)
23. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015)
24. Sun, Y., Bao, Q., Liu, W., Mei, T., Black, M.J.: Trace: 5d temporal regression of avatars with dynamic cameras in 3d environments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8856–8866 (2023)
25. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022)
26. Von Marcard, T., Rosenhahn, B., Black, M.J., Pons-Moll, G.: Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In: Computer graphics forum. vol. 36, pp. 349–360. Wiley Online Library (2017)
27. Wang, J., Liu, L., Xu, W., Sarkar, K., Luvizon, D., Theobalt, C.: Estimating egocentric 3d human pose in the wild with external weak supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13157–13166 (2022)
28. Wang, J., Liu, L., Xu, W., Sarkar, K., Theobalt, C.: Estimating egocentric 3d human pose in global space. international conference on computer vision (2021)
29. Wang, J., Luvizon, D., Xu, W., Liu, L., Sarkar, K., Theobalt, C.: Scene-aware egocentric 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13031–13040 (2023)
30. Yang, D., Kim, D., Lee, S.H.: Lobstr: Real-time lower-body pose prediction from sparse upper-body tracking signals. In: Computer Graphics Forum. vol. 40, pp. 265–275. Wiley Online Library (2021)
31. Yi, H., Liang, H., Liu, Y., Cao, Q., Wen, Y., Bolkart, T., Tao, D., Black, M.J.: Generating holistic 3d human motion from speech. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 469–480 (June 2023)
32. Yi, X., Zhou, Y., Habermann, M., Shimada, S., Golyanik, V., Theobalt, C., Xu, F.: Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13167–13178 (2022)
33. Yi, X., Zhou, Y., Xu, F.: Transpose: real-time 3d human translation and pose estimation with six inertial sensors. ACM Transactions on Graphics (TOG) 40(4), 1–13 (2021)



34. Zhang, J., Zhang, Y., Cun, X., Huang, S., Zhang, Y., Zhao, H., Lu, H., Shen, X.: T2m-gpt: Generating human motion from textual descriptions with discrete representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
35. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. arXiv preprint arXiv:2208.15001 (2022)
36. Zhang, Y., Xia, S., Chu, L., Yang, J., Wu, Q., Pei, L.: Dynamic inertial poser (dynaip): Part-based motion dynamics learning for enhanced human pose estimation with sparse inertial sensors. arXiv preprint arXiv:2312.02196 (2023)
37. Zhu, W., Ma, X., Liu, Z., Liu, L., Wu, W., Wang, Y.: Motionbert: A unified perspective on learning human motion representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)