EgoMDM: Diffusion-based Human Motion Synthesis from Sparse Egocentric Sensors

Soyong Shin[1]* Anuj Pahuja[2] Alexander Richard[2] Kris Kitani[1,2] Jason Saragih[2]
Yuhua Chen[2] Weipeng Xu[2] Eni Halilaj[1] Timur Bagautdinov[2]

[1]Carnegie Mellon University [2]Meta

{soyongs, ehalilaj, kmkitani}@andrew.cmu.edu

{anuj, yuhuac, jasons, alexr, weipengxu, timurb}@meta.com

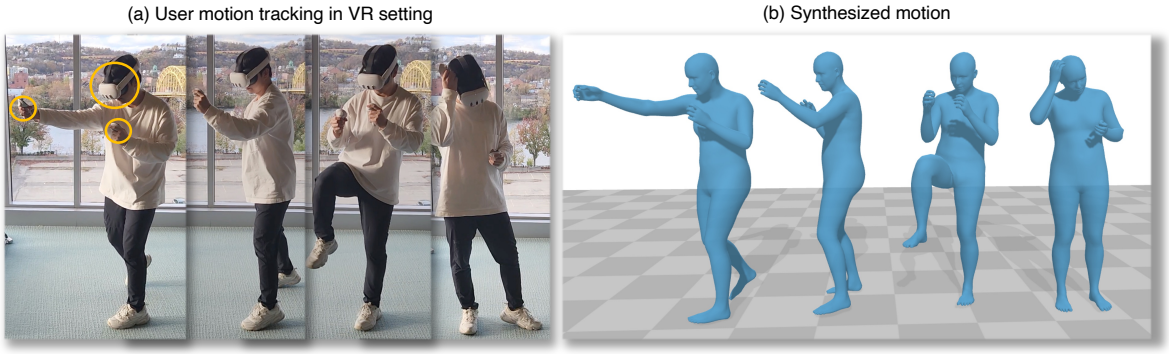

Figure 1. **EgoMDM** : Egocentric Motion Diffusion Model. In this paper, we aim to synthesize the three-dimensional (3D) motion of users wearing a Virtual Reality (VR) headset and two hand controllers. (a) Given the six degree-of-freedom (6-DoF) poses of the VR devices, (b) EgoMDM can synthesize 3D human motion of the full body.

Abstract

Accurate three-dimensional (3D) human motion tracking is essential for immersive augmented reality (AR) and virtual reality (VR) applications, allowing users to engage with virtual environments through realistic full-body avatars. Achieving this level of detail, however, is challenging when the driving signals are sparse, typically coming only from upper-body sensors, such as head-mounted devices and hand controllers. To address this challenge, we propose EgoMDM (Egocentric Motion Diffusion Model), an end-to-end diffusion-based framework designed to reconstruct full-body motion from sparse tracking signals. EgoMDM models human motion in a conditional autoregressive manner using a unidirectional recurrent neural network, making it well-suited for real-time applications. By embedding local-to-global translation, forward and inverse kinematics, and foot-contact detection within the diffusion framework, EgoMDM achieves seamless, end-to-end motion synthesis, effectively reducing artifacts like foot sliding and ground penetration. Additionally, EgoMDM is conditioned on the user's body scale, allowing it to generalize across a diverse population and produce consistent avatar shapes over time. In our extensive experiments on the AMASS motion capture dataset, EgoMDM achieves state-of-the-art performance in both motion tracking accuracy and synthesis quality, demonstrating its robustness and adaptability across various human motion scenarios. Furthermore, EgoMDM significantly outperforms the existing models when tested on real signal inputs, highlighting its robustness and applicability to the real-world data. See the project page at: https://yohanshin.github.io/egomdm.github.io/

1. Introduction

In mixed reality (MR) and virtual reality (VR) applications, accurate tracking of three-dimensional (3D) human motion is essential for enhancing user experience, enabling natural interactions, and improving immersion. In a typical MR system, only the user's head and hands are tracked using head-mounted displays (HMD) and hand controllers. This under-constrained problem of resolving full-body motion

*Work done while the author was an intern at Meta.

when tracking is restricted to the user's head and hands is a grand challenge in realizing truly immersive experiences.

Various data-driven approaches have relied on neural network regressors to learn a direct mapping from sparse observations to full-body human motion [4, 16, 18, 52]. While these methods perform well for simple motions, such as walking or running, where upper- and lower-body dynamics are closely correlated, they struggle with more ambiguous activities, where a particular upper-body observation can correspond to multiple possible lower-body motions. Previous work [3, 6, 7, 9, 45] has attempted to address this challenge through the use of conditional generative networks, including diffusion models. However, these models are generally designed within sequence-to-sequence non-causal frameworks, which have high computational demands and lead to reduced performance in real-time applications.

Existing methods also encounter difficulties with global-space realism, often producing unrealistic effects such as foot sliding, floor penetration, and floating. These failure modes stem from two main limitations. First, prior methods estimate local joint rotations in a pelvis-centered coordinate system and later translate it to global space. This sequential approach makes the model sensitive to sensor noise and error accumulation, and it also prevents the model from learning how the human interacts with the floor at different body poses. Second, most methods assume a mean human body shape, limiting adaptability across individuals of varying body scales. While recent work [4, 18] estimates body shape on a frame-by-frame basis, it often results in inconsistent avatar shapes throughout the sequence. In contrast, our method learns a motion representation tailored for analytical IK, enabling smooth, scale-consistent motion reconstruction.

To address these challenges, we propose EgoMDM (Egocentric Motion Diffusion Model), the first framework to learn a human motion representation that couples 3D joint positions with limb twist angles, enabling analytical IK without jitter and across diverse body scales. Built on an autoregressive diffusion process with unidirectional RNNs, EgoMDM is well suited for real-time application. Unlike methods that assume a mean body shape or estimate it frame-by-frame, our model is conditioned on body shape—calibrated from a T-pose or user measurements like height and wingspan—to decode motion that accurately aligns with each user's unique body scale. Our model learns the distribution of human motion decomposed into foot-ground contact probability, position and twist angle of limb joints, and torso joint rotation. The full-body motion is then analytically computed through a differentiable inverse kinematics (IK) solver. This motion representation allows diverse synthesis of human motion, while maintaining controllability to minimize foot sliding artifacts. While an analytical IK often introduces jittery motion when enforcing limb-length constraints, EgoMDM iteratively denoises motion representation, resulting in smooth motion while satisfying the constraints. In addition, by embedding local-to-global coordinate transformation into a diffusion framework, our model efficiently learns global space human motion.

Our main contributions are as follows. (1) We present an autoregressive conditional diffusion framework that synthesizes 3D human motion from sparse signals, achieving accurate and realistic full-body motion in real-time. (2) By learning human motion in global coordinates and incorporating foot contact detection, EgoMDM attains state-of-the-art performance in both tracking accuracy and synthesis quality. (3) Because EgoMDM is conditioned on individual body shape, it generalizes seamlessly across users with diverse body scales, ensuring consistent and personalized motion. (4) We introduce a novel motion representation, decoupling full-body kinematics into limb joint positions, twist angle, and torso angles to enable seamless motion reconstruction with analytical IK. (5) Despite being fully-trained on synthetic data, EgoMDM demonstrates remarkable robustness to real-world data, underscoring its practical effectiveness for real-life scenarios.

2. Related Work

Human motion tracking from wearables. Human motion tracking from sparse wearable sensors has garnered attention in recent years [2, 4, 6, 7, 14, 16–19, 21, 22, 40, 41, 44, 45, 47, 48, 52]. Previous work [14, 19, 40, 41, 47, 48] used six body-worn IMUs attached to the head, pelvis, arms, and legs to capture full-body motion. Following initial optimization-based approaches [41], DIP [14] introduced a training framework that used synthetic IMU data generated from mocap datasets [27]. Follow-up learning-based approaches have used RNN [46] and Transformer [19] architectures, and some even incorporated physics-based simulations [47, 48].

More recently, tremendous progress [2, 4, 6, 7, 9, 16–18, 22, 44, 45, 52] has been made on tracking full-body movements from AR/VR equipment: HMD and two hand controllers. The initial body of work in this area used physics simulators to leverage physics priors for human motion tracking. For example, QuestSim [44] and QuestEnvSim [22] utilized NVIDIA's *IsaacGym* [28] as the physics engine and reinforcement learning to train the model. Physics simulators, however, require significant computational resources and therefore can be challenging to deploy on mobile hardware. In contrast, another line of work directly learns a regressor that maps sparse signals to the full-body human motion. For example, AvatarPoser [16] uses a Transformer architecture to extract the temporal correlation between the three-point signals and full-body human motion, whereas AvatarJLM [52] uses a spatio-temporal Transformer to explicitly model joint-level features. HMD_Poser [4], on the other hand, introduced a unidirectional RNN-based architecture and built a lightweight system suitable for deployment.

Recently, EgoPoser [18] modeled tracking errors of hand controllers using the field of view of the HMD to improve the generalizability of the system to real-world data capture settings. MANIKIN [17] introduces an inverse kinematics operation that reconstructs 3D joint rotations while maintaining end-effector positions. However, these methods typically perceive the task as a one-to-one mapping problem, assuming a deterministic relationship between sparse input signals and full-body motion. This assumption limits their ability to overcome the inherent ill-posed nature of the problem.

Conditional human motion synthesis. Recent advancements in generative models have enabled the synthesis of human motion across a wide range of conditions. One body of work focuses on generating full-body human motion from text descriptions [11, 15, 32, 38, 49] or action labels [10, 31], aiming to produce movements that align closely with the provided prompts. Another line of research [5, 42, 43, 51] synthesizes motion based on scene environments or object interactions, with the goal of creating natural, context-aware motions that seamlessly integrate with the surrounding environment.

Similarly, synthesizing human motion from sparse tracking signals presents challenges that generative models are well-suited to address. Early work leverages flow-based architectures [2], a variational autoencoder (VAE) framework [6], or latent space codebook matching [37] to learn a conditional distribution of human motion. More recent studies [3, 7] capture the conditional probabilistic distribution of full-body motion using the diffusion model [12, 36]. SAGE [9] decomposes human motion into upper and lower body components, reconstructing full-body motion from upper-body sensor observations using a hierarchical approach. Most of these methods rely on a sequence-to-sequence framework for human motion modeling, which can impose significant computational demands or lead to reduced performance in real-time scenarios. In contrast, we develop an autoregressive conditional diffusion model with a causal structure, optimized for efficiency and real-time use.

3. Preliminaries

SMPL [25] is a differentiable function $\mathcal{M}(\theta, \beta, \gamma)$, representing a human body mesh M with 6,890 vertices through a set of low-dimensional parameters. Pose parameter θ is a set of 3D rotations of 23 body joints and the root joint's global orientation. The shape parameter β represents the 16 principle components of human-body shape learned from thousands of body scan data [20]. Finally, the translation γ is the root-joint position in the world coordinate system.

Conditional diffusion model. Following prior work [33, 39, 50] in human motion synthesis, we adopt the Denoising Diffusion Probabilistic Models (DDPMs) formulation [12]. Let $\mathbf{x}_0 = \{x_0^{(n)}\}_{n=1:N}$ be the full-body human motion with N frames, that follows the distribution $\mathbf{x}_0 \sim q(x_0)$. The forward diffusion process is a Markov chain adding Gaussian noise with the variance $\beta_t \in (0, 1)$ at each diffusion step $t \in \{1, ..., T\}$ according to a pre-defined schedule:

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

where $\mathbf{I}$ is the identity matrix. Ho *et al*. [12] show that we can directly sample $\mathbf{x}_t$ from x_0 using the properties of Gaussian distribution:

$$\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{1-\alpha_t}\epsilon,\ \text{where } \epsilon \sim \mathcal{N}(0, \mathbf{I}),$$

where $\alpha_t = 1 - \beta_t$. The reverse diffusion process is an iterative denoising to reconstruct $\mathbf{x}_0$ from Gaussian noise $\mathbf{x}_T$ over T diffusion steps. In practice, we follow prior work [39, 50] to train a denoiser neural network D that removes the added Gaussian noise based on a condition signal, sparse tracking signals $\mathbf{s}$ in our case, and a diffusion step t: $\hat{\mathbf{x}}_{t-1} = D(\mathbf{x}_t, t, \mathbf{s})$, which iteratively produces $\hat{\mathbf{x}}_0$ from random noise $\hat{\mathbf{x}}_T$.

4. Methods

4.1. Problem formulation

Given the sequence of sparse tracking signals $\mathcal{S} = \{s^{(n)}\}_{n=0}^{N}$ of N frames from HMD and hand controllers, we aim to synthesize the 3D full-body motion $\mathbf{X} = \{X^{(n)}\}_{i=0}^{N}$ of the subject with body scale b. The input signal $s^{(n)}$ at each time contains the sensor orientation s_o, position s_p, angular velocity $s_{\dot{o}}$, and linear velocity $s_{\dot{p}}$ of the head and hand controllers. Note that we use 6D representation for the sensor orientation s_o to avoid discontinuity in the input signal. Our model's final output, $\mathbf{X}$, is the set of full-body joint rotations θ and root translation γ. However, the denoising network learns the distribution of motion representation $\mathbf{x}$ that contains 1) pelvis-centered limb joint position J_{limb}, 2) torso joint rotation θ_{torso}, 3) joint twist angles θ_{twist}, and 4) foot-ground contact probability f:

$$\begin{aligned} \mathbf{s} \in \mathbf{R}^{45} &= \{s_o, s_p, s_{\dot{o}}, s_{\dot{p}}\}, \\ \mathbf{x} \in \mathbf{R}^{120} &= \{J_{limb}, \theta_{torso}, \theta_{twist}, f\}. \end{aligned}$$

Then, the full-body motion $\mathbf{X}$ can be computed deterministically from J_{limb}, θ_{torso}, and θ_{twist} using an analytical IK solver: $\mathbf{X} = IK(J_{limb}, \theta_{torso}, \theta_{twist})$. Leveraging the one-to-one mapping property of the IK solver, the diffusion model can effectively learn an equivalent representation space as $\mathbf{X}$, while maintaining a more controllable and interpretable motion decomposition.

4.2. Network overview

The overview of the proposed diffusion model architecture is shown in Fig. 2. At each time n, we denoise the motion representation $x_t^{(n)}$ conditioned on the subject body scale

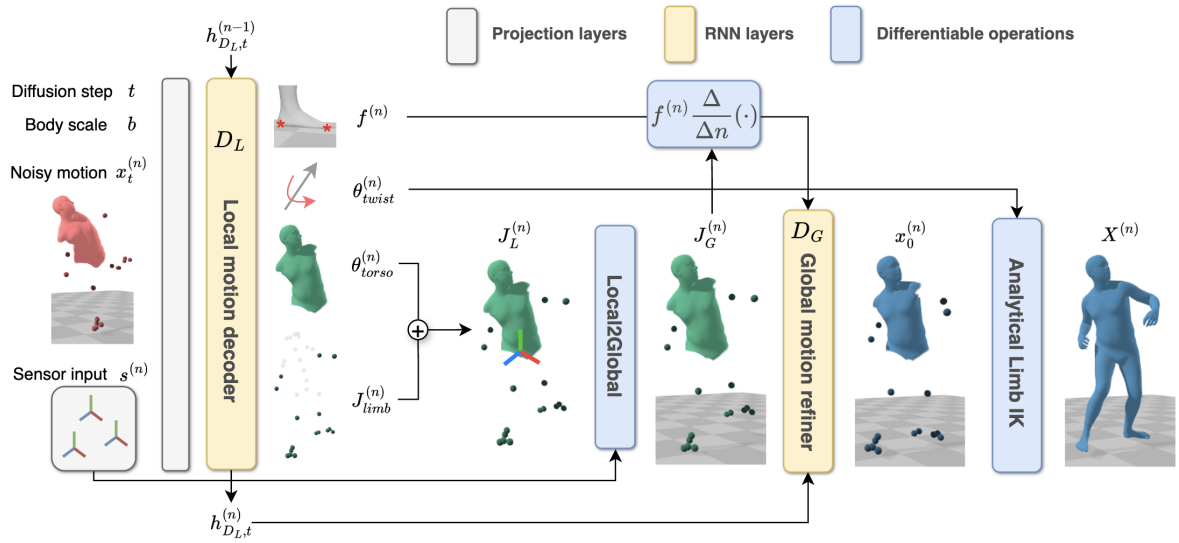

Figure 2. **Framework Overview.** Given the tracking signals of the headset and two hand controllers, we first denoise the partially represented motion and construct the full-body motion using an analytical IK solver. The denoiser network first estimates the foot-ground contact probability, limb joint twists and positions, and torso joint angles. Followed by local-to-global translation, we construct the initial global-space human motion the use a residual refinement network to update motion. Finally, full-body mesh motion is analytically computed.

b, diffusion timestep t, input signal $s^{(n)}$, and the motion context propagated from the previous frame $h_{D_L,t}^{(n-1)}$. Then, the full-body SMPL parameters are computed from a differentiable IK solver. In the following subsections, we explain the forward step of the diffusion model in three subsequent steps: local motion estimation (Section 4.3), contact-aware global motion refinement (Section 4.4), and analytical limb IK (Section 4.5).

4.3. Local motion estimation

In this stage, given the human body scale b, noisy motion $x_t^{(n)}$ at diffusion timestep t, and sensor signal $s^{(n)}$ at time n, we synthesize human motion in the pelvis-centered coordinates. The body scale is defined as the joint position configuration of the subject in T-Pose. Local coordinate denoiser D_L is combined with the projection layers that embed each input into the latent space and the RNN-based local motion decoder. The outputs of this module include the foot-ground contact probability $f^{(n)}$, torso joint rotation $\theta_{torso}^{(n)}$, twist angle $\theta_{twist}^{(n)}$, and the positions of limb joints $J_{limb}^{(n)}$:

$$f^{(n)}, \theta_{torso}^{(n)}, \theta_{twist}^{(n)}, J_{limb}^{(n)} = D_L\left(x_t^{(n)}, s^{(n)}, t, b | h_{D_L}^{(n-1)}\right).$$

Here, $h_{D_L,t}^{(n-1)}$ is the RNN hidden state propagated from the previous time $n-1$. We construct local motion $J_L^{(n)}$ by concatenating the limb joints with the torso (Fig. 2). The network is designed autoregressively that the prediction at frame $n-1$ is fed into the network at frame n. At the first frame, where we do not have the previous frame prediction, we estimate $h_{D_L}^{(0)}$ and the pseudo 0th frame prediction using the first frame sensor signal [47].

Human body decomposition into limb joints and torso is designed to address foot skating, a common artifact in MR systems. Representing motion in the joint-rotation space poses challenges in controlling this artifact, since foot positions are computed as the cumulative $SO(3)$ matrix multiplication along the kinematic tree, combined with root joint translation. By contrast, our approach directly represents human motion with the limb joint positions, enabling more precise control of limb positions and effectively mitigating foot skating artifacts. While the recently introduced MANIKIN [17] also decomposes the torso and limb joints, our approach diverges in two significant ways. First, MANIKIN assumes a clean input signal and defines hand position based on the direct hand controller measurement, making it susceptible to noise; our method predicts hand position, improving robustness. Second, MANIKIN works under the mean body-shape assumption, limiting scalability across individuals; in contrast, we adapt to diverse body sizes by conditioning the network on each subject's body scale using joint configurations from a neutral T-pose.

4.4. Global motion refinement

The pelvis-centered motion J_L can subsequently be translated to the global space using the tracking signals measured in the global coordinate system [4, 7, 9, 16, 52]. However, this approach often leads to artifacts, such as foot skating, ground penetration, and floating, as the local motion network D_L is limited in its ability to learn human motion distributions in a global coordinate system. To address theses issues, we propose a straightforward, yet effective, refinement

scheme. First, similar to prior approaches, we translate the predicted local motion J_L into the global coordinate system using the HMD position, $s_{p,head}$:

$$J_G = J_L + s_{p,head} - P2H(J_L),$$

where $P2H(\cdot)$ is an operation to compute the head position relative to the pelvis. We then introduce a refinement network D_G that processes this roughly translated global motion J_G and outputs a residual correction. Inspired by recent work [34], we further condition this refinement on the weighted foot velocity v_f computed from global motion J_G and contact probability f:

$$v_f^{(n)} = f^{(n)} \otimes (\Delta J_{G,f}/\Delta n)\,,$$

$$\Delta J_G^{(n)} = D_G\left(J_G^{(n)}, v_f^{(n)}, h_{D_L,t}^{(n)} \mid h_{D_G}^{(n-1)}\right),$$

where $\otimes$ denotes element-wise product, $J_{G,f}$ is the global foot position. We initialize $\Delta J_G^{(0)}$ and $h_{D_G}^{(0)}$ with zeros. This refinement step enables the network to efficiently address artifacts by directly refining the global-space motion, improving stability and realism in the generated motion. We obtain the motion representation $\mathbf{x}$ by integrating f, θ_{twist} and $\hat{J}_G = J_G + \Delta J_G$.

4.5. Analytical limb IK

Following HybrIK [23, 24], we reconstruct joint angles analytically from the global joint positions ($\hat{J}_G$) and twist angles (θ_{twist}). However, directly solving IK from the pelvis outward, as in HybrIK, can cause cumulative positional errors due to minor limb-length mismatches. To mitigate this, we first refine mid-joint positions (knees, elbows) using a bone-length constraint. Specifically, with fixed parent and child joint positions (e.g., hips and ankles), mid-joint positions must lie on an orbit defined by these fixed joint positions and the corresponding bone lengths. We select the refined position closest to the initial prediction while maintaining the constraints. After refinement, limb segment rotations are computed analytically following HybrIK's approach, decomposing the rotation into swing and twist angles. For full derivation, please refer to the *Sup. Mat.*

4.6. Losses

We train our diffusion denoiser network using the integrated loss defined as:

$$\begin{aligned}\mathcal{L} \quad = \quad & \mathcal{L}_{simple} + \lambda_{J3D}\mathcal{L}_{J3D} + \lambda_{\theta}\mathcal{L}_{\theta} + \\ & \lambda_{\text{skate}}\mathcal{L}_{\text{skate}} + \lambda_{\text{recon}}\mathcal{L}_{\text{recon}}.\end{aligned}$$

The simple diffusion objective $\mathcal{L}_{simple}$ is defined by comparing ground-truth motion $\mathbf{x}_0$ and denoised motion $\hat{\mathbf{x}}_0$:

$$\mathcal{L}_{simple} = \mathrm{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0), t \sim [1,T]}\left[||\mathbf{x}_0 - \hat{\mathbf{x}}_0||_2^2\right].$$

$\mathcal{L}_{J3D}$ and θ are the losses enforcing consistency of estimated 3D joint position $\hat{J}_{3D}$ and joint angles $\hat{\theta}$ with the ground truth data:

$$\mathcal{L}_{J3D} = ||J - \hat{J}||_2^2, \quad \mathcal{L}_{\theta} = ||\theta - \hat{\theta}||_2^2.$$

$\mathcal{L}_{\text{skate}}$ penalizes the displacement of the foot keypoints when the estimated contact probability f is high:

$$\mathcal{L}_{\text{skate}} = ||f_{0.5} \circ v_f||_2^2,$$

where $f_{0.5}$ denotes the binary contact mask based on threshold of 0.5, and $\circ$ is the masking operation. Finally, we use a tracking-signal reconstruction loss $\mathcal{L}_{\text{recon}}$ to enforce the synthesized motion to generate a virtual-tracking signal that is consistent with the input:

$$\mathcal{L}_{\text{recon}} = ||s - \mathcal{S}(\hat{\mathbf{x}})||_2^2,$$

where $\mathcal{S}(\cdot)$ is the function that synthesizes virtual tracking signals from the full-body motion.

4.7. Implementation details

We implement the temporal network using a unidirectional Long Short-Term Memory (LSTM) network [13]. The local motion decoder consists of three LSTM layers, while the global motion refinement model uses two layers. Following the autoregressive design of recent video-based human motion estimation models [34], EgoMDM's LSTM layers take in the previous frame's estimation, with the initial RNN hidden state and the prediction for the zeroth frame derived from the sensor signals in the first frame. For body scale representation, we use the pelvis-centered resting pose joint positions. The input signals are split into three components: left controller position, right controller position, and the remaining signals (head position along with the sensors' orientation, linear velocity, and angular velocity), which are then independently projected into the embedding space. This design choice allows for random masking of the left and right hand controller position signals, enhancing the system's robustness to sensor noise. Please see *Sup. Mat.* for more detail.

During training, we crop the motion sequences to 81 frames and apply random-rotation augmentation around the vertical axis to the ground plane. We use $T = 1000$ and uniformly sample diffusion timestep $t \in [1, \ldots, 1000]$. At inference, we start with the pure Gaussian noise and follow 5 DDIM sampling [36] steps. We used the AdamW optimizer [26] with a learning rate of $3e-4$, batch size of 1024, and then reduced the learning rate to $1/10$ after 225,000 and 350,000 steps. Experiments were performed on a single NVIDIA A100 GPU, using PyTorch framework [30].

5. Experiments

Datasets. We train and evaluate our method on AMASS [27], a large-scale public dataset that integrates multiple

Models	Protocol 1							Protocol 2						
	MPJPE	MPJVE	Jitter	UPE	LPE	HPE	RPE	MPJPE	MPJVE	Jitter	UPE	LPE	HPE	RPE
Ground Truth	0	0	1.28	0	0	0	0	0	0	1.15	0	0	0	0
AGRoL† [7]	3.71	19.08	1.86	1.55	6.84	1.31	3.36	6.17	24.14	1.79	2.42	12.36	1.69	5.61
AvatarJLM† [52]	3.35	20.91	2.45	1.54	6.56	0.66	2.96	4.93	27.50	2.46	**2.09**	9.86	**0.93**	4.46
SAGE† [9]	3.28	20.62	1.81	**1.39**	6.01	1.18	2.95	5.86	33.54	2.82	2.40	12.07	1.99	5.00
MANIKIN† [17]	3.19	20.10	–	1.43	6.27	**0.01**	–	–	–	–	–	–	–	–
EgoMDM† (Ours)	**3.13**	**15.18**	**1.46**	1.70	**5.63**	1.55	**2.87**	**4.62**	**20.73**	**1.52**	2.24	**8.79**	2.04	**4.09**
HMD-Poser [4]	**3.19**	17.47	1.83	**1.67**	6.27	1.65	**3.21**	5.44	30.15	2.59	**2.44**	9.77	2.56	4.83
EgoMDM (Ours)	3.63	**15.41**	**1.46**	2.24	**6.05**	**1.51**	3.34	**4.89**	**20.74**	**1.51**	2.51	**9.05**	**1.86**	**4.23**

Table 1. Quantitative comparison of geometric accuracies with state-of-the-art models on AMASS dataset. The best results are shown in **bold**. † denote the model assumes the known body shape of the subjects.

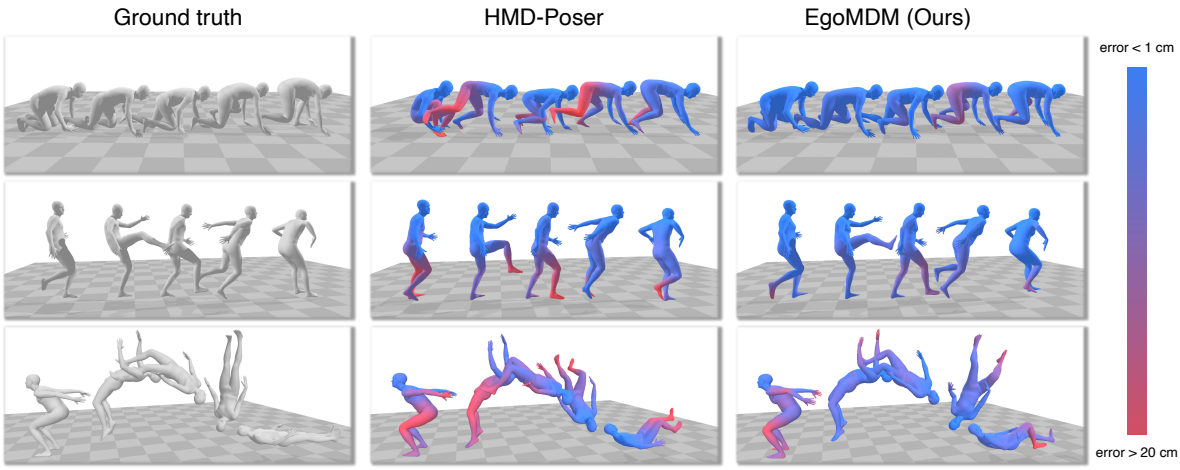

Figure 3. **Qualitative Assessment of Overall Performance**. A comparison of motion-tracking accuracy between HMD-Poser [4] (second column) and EgoMDM (ours, third column). Ground-truth is derived from the AMASS mocap dataset (first column). Vertices are colored differently based on the per-vertex distance to the ground-truth motion (red indicates worse performance). The motion synthesized by EgoMDM shows larger similarity to the reference ground-truth motion than the state-of-the-art method in various movement scenarios.

Mocap datasets and represents human motion in terms of SMPL [25] body model parameters. Following prior work [4, 7, 9, 52], we split training and testing sets using two distinct protocols. The first protocol (P1) uses 3 subsets within the AMASS dataset: specifically, CMU [1], BMLr [8], and HDM05 [29]. In this protocol, we randomly split the data into 90% training and 10% testing data. The second protocol (P2) uses 12 entire subsets for training and 2 left-over subsets for evaluation. Transitions [27] and HumanEva [35] are the evaluation subsets. As a result, while P1 includes more testing subjects, it has overlap between subjects in the training and testing sets, whereas P2 ensures no such overlap. We use the original body shape parameters of the subjects to construct the ground-truth motion and input signals.

Additionally, we assess the real-world applicability of our method by evaluating it on the recently published PICO-FreeDancing dataset [4]. Unlike P1 and P2, which are synthetically generated, PICO-FreeDancing contains real-world HMD signals alongside corresponding ground-truth motions captured through Mocap systems.

Evaluation metrics. We compare our model against state-of-the-art methods in two categories. First, we evaluate the tracking accuracy of each method using Position Error (*PE*, in cm) across whole body (*MPJPE*), upper–body (*UPE*), lower-body (*LPE*), hands (*HPE*), and root joint (*RPE*). We also assess the temporal coherence to the ground truth motion using joint velocity (*MPJVE* in cm/s) and the jitter (*Jitter* in $10^2\ m/s$) that measures the third-order derivative of joint positions. Second, we evaluate the feasibility and realism of the generated motion. We quantify the average displacement of the foot joints during contact (*Skate*) and the mean of absolute ground penetration and floating (*Ground*). In addition, we compute the *FID* score that measures the dissimilarity between the synthesized and ground truth motion

	Models	Skate	Ground	FID	Diversity
Protocol 1	AGRoL† [7]	0.21	2.07	0.29	7.30
	AvatarJLM† [52]	0.22	1.74	0.27	–
	SAGE† [9]	0.28	1.81	0.26	0.03
	EgoMDM† (Ours)	**0.10**	**1.48**	**0.10**	**11.66**
	HMD-Poser [4]	0.24	1.51	0.26	–
	EgoMDM (Ours)	**0.10**	**1.43**	**0.25**	**11.99**
Protocol 2	AGRoL† [7]	0.32	2.58	0.95	19.26
	AvatarJLM† [52]	0.34	1.53	0.41	–
	SAGE† [9]	0.51	1.74	0.55	0.03
	EgoMDM† (Ours)	**0.19**	**1.29**	**0.40**	**20.98**
	HMD-Poser [4]	0.24	1.51	0.61	–
	EgoMDM (Ours)	**0.19**	**1.26**	**0.37**	**20.93**

Table 2. Quantitative comparison of motion synthesis quality. † denote the model assumes the known body shape of the subjects.

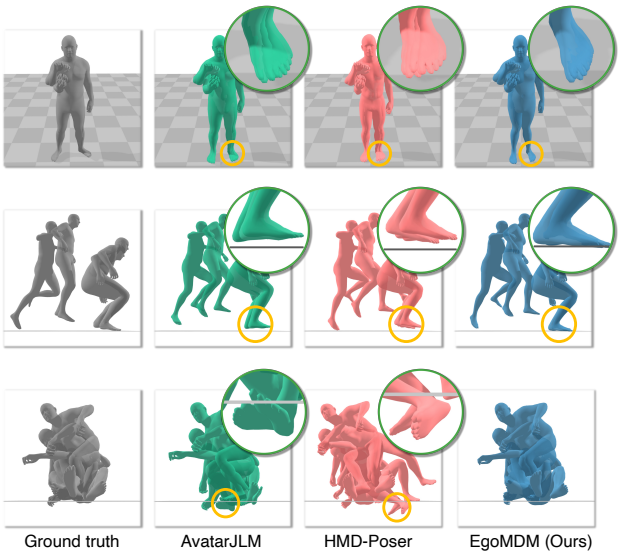

Figure 4. **Qualitative Assessment of Realism**. A comparison of motion-synthesis quality between AvatarJLM [52] (second column), HMD-Poser [4] (third column), and EgoMDM (fourth). Ground-truth is based on mocap AMASS data (first column). EgoMDM shows less foot skating (**first row**), floating (**second row**), and floor penetration (**third row**) compared to the other methods.

distribution in latent space. Last, we use *Diversity* metric to quantify the variation in lower-body joint positions between motion samples synthesized from the same sensor input.

During the evaluation, we compare our model with HMD-Poser, which predicts body shapes, by using the predicted shape parameters derived from the T-pose (*i.e.* height and wingspan measurements). In contrast, methods such as AGRoL [7], AvatarJLM [52], SAGE [9], and MANIKIN [17] assume a mean body shape and are trained and tested on shape-normalized data, where the model directly leverages the known body scale of test subjects. To ensure a fair comparison in terms of available body-scale information, we additionally evaluate our model with the ground-truth body shape parameters of test subjects, thus aligning the experimental conditions with those of prior methods. Except for the *Diversity* assessment, we generate 16 distinct noise samples for each input tracking signal and average the motion at every denoising step. Additional details on the evaluation protocols can be found in the *Sup. Mat.*

5.1. 3D Human Motion Synthesis.

Motion tracking accuracy. To demonstrate the effectiveness of our method in synthesizing accurate and temporally coherent motion from partial input data, Table 1 compares EgoMDM to state-of-the-art approaches [4, 7, 9, 52] under two distinct evaluation protocols (P1 and P2). Notably, Protocol 2 involves separating training and testing data by subject, with no subject-level overlap, making it a strong indicator of each method's ability to generalize to new subjects. Under both known/unknown shape scenarios, EgoMDM outperforms existing methods in the tracking accuracy, particularly in *MPJVE*, *Jitter* and *LPE*, showing that EgoMDM can accurately synthesize lower body from upper-body sensors while preserving temporal coherence. Moreover, the proposed model demonstrates a significant improvement over state-of-the-art methods under Protocol 2, highlighting EgoMDM's enhanced generalizability to unseen subjects. Figure 3 shows qualitative examples of how EgoMDM outperforms the current state-of-the-art method [4] in diverse and challenging motions.

Motion synthesis quality. Beyond tracking accuracy, realistic motion synthesis should be physically plausible, distributionally similar to real motion data, and sufficiently diverse. Thus, we evaluate our method using three categories of metrics: (1) physical plausibility (*Skate*, *Ground*), (2) distributional similarity (*FID*), and (3) motion diversity (*Diversity*) (Table 2). EgoMDM achieves significantly lower foot skating, ground penetration, and floating artifacts compared to existing methods, highlighting our model's effective integration of contact detection for superior foot positioning control. Figure 4 demonstrates the superior performance of our model over the state-of-the-art methods [4, 52] in synthesizing human motion with less foot skating, floating, and ground penetration. Additionally, despite employing a simple and computationally efficient network architecture, EgoMDM produces more realistic results (*FID*) than competitors that rely on computationally intensive Transformer architectures. Finally, our model generates more diverse motion samples than existing conditional generative approaches [7, 9], without sacrificing accuracy or plausibility.

Real-world data evaluation. Since the previous experiments leveraged synthetic datasets with idealized input signals, they do not reflect the noisy conditions typical of real-world sensor data. To verify our method's robustness and

Models	PICO-FreeDancing					
	MPJPE	HPE	MPJVE	Jitter	Skate	Ground
Ground Truth	0	0	0	1.48	0.07	0.82
AGRoL [7]	10.49	12.91	30.60	2.88	0.48	1.65
AvatarJLM [52]	9.02	8.71	32.72	4.55	0.51	1.62
SAGE [9]	10.04	10.07	41.61	4.08	0.87	1.85
HMD-Poser [4]	9.71	8.53	38.69	3.79	0.76	1.93
EgoMDM (Ours)	**7.88**	**6.20**	**25.15**	**2.39**	**0.28**	**1.01**

Table 3. Quantitative results on real data, PICO-FreeDancing [4].

Models	Protocol 2				
	MPJPE	MPJVE	Jitter	Skate	Ground
w/o Analytical IK	5.89	28.00	**1.24**	0.23	1.45
w/o Shape cond.	5.21	23.51	1.77	<u>0.22</u>	**1.26**
w/o Refinement	**4.88**	21.74	1.62	0.24	1.28
w/o Diffusion	5.20	24.16	1.67	0.28	1.66
w/o $\mathcal{L}_{skate}$	4.90	<u>21.29</u>	1.52	0.24	1.59
EgoMDM (Ours)	<u>4.89</u>	**20.74**	<u>1.51</u>	**0.19**	<u>1.27</u>

Table 4. Ablation experiments under the AMASS protocol 2. The best and second-best results are in **bold** and <u>underline</u>.

applicability in practical scenarios, we further evaluate performance on the real-world PICO-FreeDancing dataset (Table 3). We observed consistent superiority of EgoMDM across all metrics, confirming its robustness and suitability for practical applications. A particularly noteworthy improvement is observed in hand tracking accuracy (*HPE*). Unlike synthetic datasets where controllers perfectly align with hand positions—conditions under which prior methods excel—real-world signals often contain noise and misalignments between controllers and hands. In these realistic conditions, EgoMDM achieves substantial improvements, with over 15% higher accuracy in motion tracking and approximately 60% fewer artifacts in motion feasibility compared to previous state-of-the-art methods. These results highlight EgoMDM's superior robustness and ability to handle noisy sensor inputs effectively.

5.2. Ablation Study

Our entire system outperforms the different variants of EgoMDM that ablate each component (Table 4). Specifically, we observe that the direct full-body joint angle regression (vs. w/o Analytical IK) enhances both tracking accuracy and motion synthesis quality, though it comes at the cost of reduced smoothness in the motion. The ablation of shape conditioning exhibits not only higher tracking error but also larger *Jitter*, inferring the Analytical IK can provide smooth motion when the body scale of the subject is provided. In addition, EgoMDM significantly outperforms the version without global motion refinement network D_G (w/o Refinement) in *Skate* and *Jitter* metrics. This indicates that the refinement process conditioned on the feet velocity effectively reduces the foot skating artifacts, as well as smooth the motion in the world coordinate system. We further observe that the proposed diffusion framework provides more accurate and feasible motion synthesis performance throughout iterative diffuse-denoise process compared to the one used deterministic regressor (w/o Diffusion). Last, the addition of foot skating error $\mathcal{L}_{skate}$ allows EgoMDM to synthesize human motion with less foot skating (*Skate*) and ground penetration and floating (*Ground*).

5.3. Inference Speed

To evaluate real-time feasibility, we conducted an inference speed analysis on both a GPU (NVIDIA RTX 3090) and a CPU (Macbook M1 Pro). Our method achieved 123.4 FPS on the GPU and 80.8 FPS on the CPU, utilizing *5-step DDIM* sampling for the diffusion process. These results demonstrate that our model runs effectively in real-time on standard consumer hardware, highlighting its potential for practical, on-device deployment.

6. Conclusion

We introduced EgoMDM, an efficient diffusion-based framework for synthesizing 3D human motion from sparse tracking inputs in VR settings. By leveraging conditional diffusion on body shape and integrating foot-ground-contact detection with global motion refinement, EgoMDM overcomes common challenges such as foot sliding and ground penetration, producing realistic and stable full-body motion. The proposed framework achieves state-of-the-art results in both geometric accuracy and feasibility, even with a lightweight, unidirectional RNN-based architecture suitable for real-time applications. Our evaluations on standard benchmarks highlight EgoMDM's robust generalizability to unseen subjects and significant improvements over existing methods. This work makes it possible to use avatars beyond coarse motion-tracking by more realistically modeling contact dynamics, improving human-scene interaction, and expanding the application of VR systems to rehabilitation, among other applications where contact dynamics are of interest.

Limitations and future directions: Our diffusion-based framework, trained with random hand-controller masking, demonstrates robustness to real-world data. However, as the network is trained exclusively on synthetic data, it may be susceptible to sensor noise that falls outside the training distribution. Additionally, our model only addresses human interaction with a flat ground plane, which may limit its generalizability to more complex human-scene interactions. Future work could focus on integrating physics-based priors and enhancing adaptability to improve motion realism across a wider range of VR environments.

References

[1] Carnegie mellon university. cmu mocap dataset. 6

[2] Sadegh Aliakbarian, Pashmina Cameron, Federica Bogo, Andrew Fitzgibbon, and Thomas J. Cashman. Flag: Flow-based 3d avatar generation from sparse observations. In *CVPR*, pages 13243–13252, 2022. 2, 3

[3] Angela Castillo, Maria Escobar, Guillaume Jeannerete, Albert Pumarola, Pablo Arbeláez, Ali Thabet, and Artsiom Sanakoyeu. Bodiffusion: Diffusing sparse observations for full-body human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2, 3

[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*, 2024. 2, 4, 6, 7, 8

[5] Markos Diomataris, Nikos Athanasiou, Omid Taheri, Xi Wang, Otmar Hilliges, and Michael J. Black. WANDR: Intention-guided human motion generation. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[6] 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 (ICCV)*, pages 11687–11697, 2021. 2, 3

[7] 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 *CVPR*, 2023. 2, 3, 4, 6, 7, 8

[8] Nikolaus F. Troje. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns, 2002. 6

[9] Han Feng, Wenchao Ma, Quankai Gao, Xianwei Zheng, Nan Xue, and Huijuan Xu. Stratified avatar generation from sparse observations. In *CVPR*, 2024. 2, 3, 4, 6, 7, 8

[10] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020. 3

[11] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161, 2022. 3

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020. 3

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 5

[14] 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 Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 37:185:1–185:15, 2018. First two authors contributed equally. 2

[15] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[16] 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 *Proceedings of European Conference on Computer Vision*. Springer, 2022. 2, 4

[17] Jiaxi Jiang, Paul Streli, Xuejing Luo, Christoph Gebhardt, and Christian Holz. Manikin biomechanically accurate neural inverse kinematics for human motion estimation. In *ECCV*, 2024. 3, 4, 6, 7

[18] Jiaxi Jiang, Paul Streli, Manuel Meier, and Christian Holz. Egoposer: Robust real-time egocentric pose estimation from sparse and intermittent observations everywhere. In *ECCV*, 2024. 2, 3

[19] 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*, 2022. 2

[20] Robinette Kathleen, Blackwell Sherri, Daanen Hein, Boehmer Mark, Fleming Scott, Brill Tina, Hoeferlin David, and Burnsides Dennis. Civilian american and european surface anthropometry resource (caesar) final report. *Tech. Rep. AFRL-HEWP-TR-2002-0169*, 2002. 3

[21] 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 (CVPR)*, 2024. 2

[22] 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*, New York, NY, USA, 2023. Association for Computing Machinery. 2

[23] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3383–3393, 2021. 5

[24] Jiefeng Li, Siyuan Bian, Chao Xu, Zhicun Chen, Lixin Yang, and Cewu Lu. Hybrik-x: Hybrid analytical-neural inverse kinematics for whole-body mesh recovery. *arXiv preprint arXiv:2304.05690*, 2023. 5

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 3, 6

[26] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam, 2018. 5

[27] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of

motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, 2019. 2, 5, 6

[28] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. Isaac gym: High performance gpu-based physics simulation for robot learning, 2021. 2

[29] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. Documentation mocap database hdm05. Technical Report CG-2007-2, Universität Bonn, 2007. 6

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. *PyTorch: an imperative style, high-performance deep learning library*. Curran Associates Inc., Red Hook, NY, USA, 2019. 5

[31] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[32] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[33] Yoni Shafir, Guy Tevet, Roy Kapon, and Amit Haim Bermano. Human motion diffusion as a generative prior. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[34] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 5

[35] L. Sigal, A. Balan, and M. J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87(1):4–27, 2010. 6

[36] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 3, 5

[37] Sebastian Starke, Paul Starke, Nicky He, Taku Komura, and Yuting Ye. Categorical codebook matching for embodied character controllers. *ACM Trans. Graph.*, 43(4), 2024. 3

[38] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 358–374. Springer, 2022. 3

[39] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[40] 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 (CVPR)*, pages 2513–2523, 2024. 2

[41] T. von Marcard, B. Rosenhahn, M. J. Black, and G. Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. *Comput. Graph. Forum*, 36(2):349–360, 2017. 2

[42] Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. Scene-aware generative network for human motion synthesis. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[43] Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. Towards diverse and natural scene-aware 3d human motion synthesis. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[44] Alexander Winkler, Jungdam Won, and Yuting Ye. Questsim: Human motion tracking from sparse sensors with simulated avatars. In *SIGGRAPH Asia 2022 Conference Papers*. Association for Computing Machinery, 2022. 2

[45] Ming Yan, Xin Wang, Yudi Dai, Siqi Shen, Chenglu Wen, Lan Xu, Yuexin Ma, and Cheng Wang. Cimi4d: A large multimodal climbing motion dataset under human-scene interactions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12977–12988, 2023. 2

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

[47] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 4

[48] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Physical non-inertial poser (pnp): Modeling non-inertial effects in sparse-inertial human motion capture. In *SIGGRAPH 2024 Conference Papers*, 2024. 2

[49] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. 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. 3

[50] Siwei Zhang, Bharat Lal Bhatnagar, Yuanlu Xu, Alexander Winkler, Petr Kadlecek, Siyu Tang, and Federica Bogo. Rohm: Robust human motion reconstruction via diffusion. In *CVPR*, 2024. 3

[51] Kaifeng Zhao, Yan Zhang, Shaofei Wang, Thabo Beeler, and Siyu Tang. Synthesizing diverse human motions in 3d indoor scenes. In *International conference on computer vision (ICCV)*, 2023. 3

[52] 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 (ICCV)*, 2023. 2, 4, 6, 7, 8