HMD^2: Environment-aware Motion Generation from Single Egocentric Head-Mounted Device

Vladimir Guzov*‡[1,2] Yifeng Jiang*†[3] Fangzhou Hong‡[4] Gerard Pons-Moll[1,2]
Richard Newcombe[5] C. Karen Liu[3] Yuting Ye[5] Lingni Ma[5]

[1]Tübingen AI Center, University of Tübingen [2]Max Planck Institute for Informatics, Saarland Informatics Campus
[3]Stanford University [4]Nanyang Technological University [5]Meta Reality Labs Research

https://hmdsquared.github.io

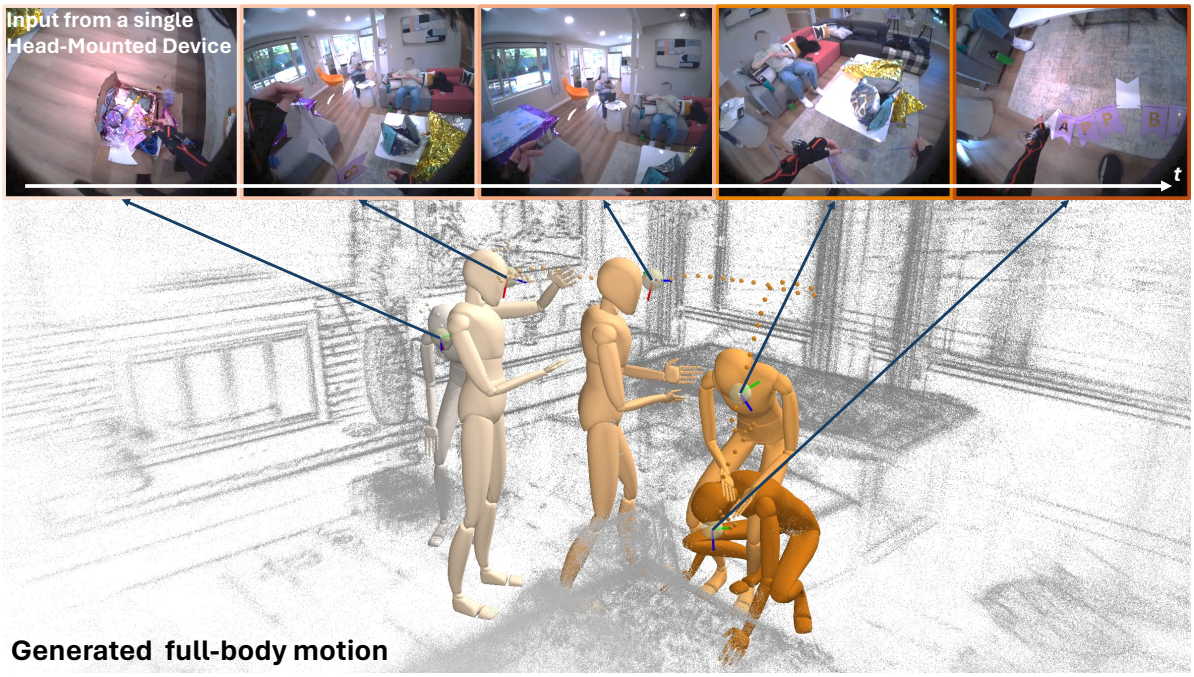

Figure 1. We propose HMD^2, the first system for the online generation of full-body motion using a single head-mounted device (*e.g.* Project Aria Glasses) equipped with an outward-facing camera in complex and diverse environments.

Abstract

This paper investigates the generation of realistic full-body human motion using a single head-mounted device with an outward-facing color camera and the ability to perform visual SLAM. To address the ambiguity of this setup, we present HMD^2, a novel system that balances motion reconstruction and generation. From a reconstruction standpoint, it aims to maximally utilize the camera streams to produce both analytical and learned features, including head motion, SLAM point cloud, and image embeddings. On the generative front, HMD^2 employs a multi-modal conditional motion diffusion model with a Transformer backbone to maintain temporal coherence of generated motions, and utilizes autoregressive inpainting to facilitate online motion inference with minimal latency (0.17 seconds). We show that our system provides an effective and robust solution that scales to a diverse dataset of over 200 hours of motion in complex indoor and outdoor environments.

1. Introduction

Wearable devices such as smart glasses promise to become the cornerstone of next-generation personal computing. A key challenge is accurately interpreting the wearer's motion from the device's limited input signals, taking into account the social and environmental context at the moment. The capability to generate full-body movements solely from a single head-mounted device (HMD) in real-time, outdoors and indoors, will open the door to many downstream applications, including telepresence, fitness and health monitoring, and navigation.

* Equal contribution.
‡Work done during internships at Meta Reality Labs Research.
†Work done partially during internship at Meta Reality Labs Research.

State-of-the-art methods, such as EgoEgo [36], have shown visually impressive results in a similar context. However, these systems operate offline, are optimized for generating short windows of motion, and are mostly trained on a small set of indoor motions. More crucially, they utilize the head-mounted camera only for head pose estimation, missing the opportunity to harness additional image features of the environment and of the wearer's own body.

In this paper, we introduce HMD^2 (Human Motion Diffusion from HMD), the first system, to our knowledge, capable of online generation of full-body movements from a single HMD (Project Aria Glasses [13]), conditioned on outward-facing egocentric camera streams in diverse environments. Given that such devices provide limited observation of the body and surroundings, the critical question is how to maximally utilize the input. Our approach reuses input data to generate features across different modalities, covering independent aspects of the environment and motion. Specifically, from the input streams, we mix and match analytical and learning toolboxes to extract 1. wearer's head motion from off-the-shelf real-time visual SLAM; 2. environment feature points as a by-product of SLAM, important for motion disambiguation in complex scenes; and 3. head camera image embeddings (*e.g.* using CLIP [47]) for additional scene clues and intermittently visible body parts.

However, full recovery of the wearer's motion is still highly under-constrained, given our input. Our system adopts a generative approach with a diffusion-based Transformer backbone to balance motion reconstruction and generation, enabling diverse outcomes, such as varying leg movements, from the same inputs. Additionally, our diffusion model can predict motions with minimal future information (0.17 s), supporting online and real-time use cases.

Contrary to evaluations using large synthetic datasets or small-scale real-world datasets, we train and test our system on the extensive 200-hour real-world Nymeria dataset [41] recorded with publicly available head-mounted device, containing various indoor and outdoor activities performed by over 100 subjects with diverse body sizes and demographics. While most existing research on motion tracking is evaluated solely based on reconstruction accuracy, we acknowledge the inherent ambiguity in our problem and evaluate our system on generation fidelity and diversity as well. Our contributions are summarized as follows:

1. We present a novel application of online full-body motion generation from a single HMD. The multi-modal feature streams extracted from the device serve as a key ingredient for the system's success across a diverse set of environments.
2. We employ a multi-modal conditional motion diffusion backbone, effectively balancing between accurate motion reconstruction and the diversity and fidelity of synthesized movements.
3. We demonstrate the adaption of a time-series motion diffusion model for online autoregressive inference through inpainting, eliminating the dependency on future sensor input and achieving minimal latency.
4. We evaluate the proposed system with large-scale, real-world Nymeria [41] dataset and achieve state-of-the-art performance for single-HMD motion generation.

2. Related Work

Human Motion from Sparse Sensors. Capturing motion with wearable sensors has gained interest across fields like Computer Vision, Graphics, and Health. Self-contained sensors like IMUs [57], electromagnetic sensors [33], and EMGs [9] offer motion reconstruction without the need for costly studios with multiple cameras. The sparse sensor placement reduces user friction, but high noise levels require learning methods to improve reconstruction. Examples include six IMUs configurations [27, 31, 57, 67, 68], head and wrists VR trackers [7, 12, 29, 30, 64, 80], and hybrid approaches with an external RGB camera [65].

Our approach uses a single wearable device to minimize user friction, though this complicates the recovering of motion. However, for many applications like telepresence, visually appealing, realistic, and diverse inferred motions are often more important than precision. Thus, we evaluate our system not just on reconstruction accuracy but also on realism and diversity – metrics often overlooked in this field.

Pose and Motion from Egocentric Cameras. Wearable egocentric cameras are ideal for self-contained motion generation systems, which saw increasing research interest. Two main types of body-mounted cameras – downward-facing (often fish-eye) and outward-facing – have been the focus of research. Most studies on downward cameras [8, 32, 38, 44, 48, 52, 59, 63, 75, 77], directly predict current pose from corresponding images, sacrificing temporal coherence. Wang *et al*. [58] addressed this by adopting a diffusion model for temporal regularization in a separate refinement stage, which inspired us to adopt a diffusion backbone and a single-stage time-window-based learning architecture. Both synthesized [4] and, recently, real-device [3] datasets are used to train and evaluate such methods.

Outward-facing cameras are more common on current devices (e.g. Project Aria [13]), though egocentric motion generation is less explored in this setup. A key challenge identified in early work with chest-mounted cameras [28] is intermittent body visibility, which makes the task under-constrained. Later works [39, 70, 71] explored simulation methods that leveraged physics to address missing motion information. EgoEgo [36] demonstrated the generalizability of single camera systems to large-scale datasets. We build upon EgoEgo, while utilizing additional visual cues beyond head pose inference and enhancing support for non-

flat terrains and low-latency long sequence generation.

There has also been research effort on combining wearable sensors such as IMUs with head-mounted egocentric cameras for accurate motion reconstruction [17, 35, 66]. Our system can be easily adapted to such multi-device setups as well, which could further improve its accuracy.

Learning-based Pose and Motion Generation. Generating controllable and realistic human movements is a long-standing goal in computer graphics and vision. Modern deep learning opens new possibilities for this problem, with earlier attempts exploring both regression-based [23, 24] and generative [20, 37] frameworks. Recently diffusion models demonstrated impressive capabilities in the generative setting across various tasks such as text [49, 81], music [53], and audio [5] conditioned motion generation. While the field starts to see conditional diffusion methods where the control signal is temporally dense [7, 12], frameworks that generate motions in an online fashion with minimal latency [55] are still underexplored. Our work adopts autoregressive inpainting for low-latency inference – this concept of autoregressive diffusion models has been explored in the motion domain albeit in different contexts [18, 50, 69, 76].

The success of diffusion models in motion synthesis has also intrigued researchers to use them for pose reconstruction, *e.g.* from third-person view, especially when ambiguity exists [10, 11, 14, 15, 25, 73, 74]. Our task is highly ambiguous as well, and our system adopts Transformer-based diffusion models to generate temporally coherent motions.

Scene-aware Pose and Motion Modeling. Motion generation and reconstruction satisfying scene and environment constraints is critical for learning-based motion models to become practical. Recent work has looked into various methods and representations to incorporate scene information, such as shape primitives [34, 40], point-cloud-based networks [26, 61, 62, 78], voxel-based networks [19, 51, 60], scene images [6], signed distance fields [79], to name a few. With most methods targeting offline applications and many requiring end-of-motion goal specifications, our scene representation with a per-frame bounding box and autoencoder facilitate online usage and large-scene deployment. The scene points in our method are captured from the same head-mounted device during SLAM without needing additional scanning devices. As a trade-off, the available scene points are sparser and noisier.

3. Method

We introduce a diffusion-based framework for generating full-body motion based on multi-modal signals from an HMD, like the Project Aria Glasses [13]. As shown in Fig. 2, our system uses device with an outward-facing camera, capable of real-time SLAM [1] (which may utilize other sensors) which produces a 6D pose trajectory, and a spatial map of the environment represented by an aggregated point cloud. We extract contextual information from both the environment point clouds and the egocentric video stream, using a CLIP encoder [47] for image embedding and an independently trained point cloud autoencoder for spatial map embedding to supplement the 6D pose.

Given the under-constrained nature of the task, we employ a diffusion model [21] with a time-series Transformer encoder [56] to model the motion distribution. To ensure temporal consistency during streaming, we use autoregressive inpainting during denoising, aligning new body motion with previous predictions.

3.1. Multi-modal Scene and Motion Conditions

Our model is trained to align its output with three modalities of features, all of which are streaming frame by frame to allow infinitely long motion generation. For each frame, the inputs include a head pose $(\boldsymbol{t}, \boldsymbol{R}) \in \mathrm{SE}(3)$ representing the head's position and orientation, a color image $\boldsymbol{I}$ from the camera, and a set of SLAM feature points $\boldsymbol{S} \in \mathbb{R}^{N\times 3}$ of the surrounding scene. We concatenate features per-frame and process the resulting vector with a linear layer (see supplementary). We elaborate on each modality and their respective design considerations below.

Head Pose Trajectory. The device pose provides precise spatial location and movement of the wearer's head. We augment the device pose vector with its linear and angular velocity vector $(\boldsymbol{v}, \boldsymbol{\omega})$ computed from finite differences to form $\boldsymbol{p} = \{\boldsymbol{t}, \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{\omega}\}$. We canonicalize each window of $\{\boldsymbol{p}\}_{0,1,\cdots,T}$ to its first frame $\boldsymbol{p}_0$, allowing the model to function in arbitrarily large spaces and generate infinitely long sequences. This is crucial for navigation in a multistory building or outdoor hiking with large elevation changes.

Camera Image Embeddings. Beyond the head pose trajectory derived from SLAM, the egocentric camera images offer additional valuable information. For example, when a body part becomes visible, the image provides a strong cue of the wearer's pose. However, direct utilization of the image content proves less useful, as it may capture distracting texture details when all we need is high-level semantics such as "the left hand is above the waist." Empirically, we found that CLIP embeddings [47], $E_I(\boldsymbol{I})$, provide significant performance boost to the learning process while avoiding overfitting to superficial image characteristics.

It is crucial to note that embeddings from human-related backbones, such as those trained for pose reconstruction from monocular videos, do not perform well in our case. Figure 3 shows a typical input camera sequence when only a few parts of the body (hands in this case) are visible. This differs significantly from downward-facing egocentric cameras, which observe most of the body. This discrepancy leads to failures in existing network backbones for full-body

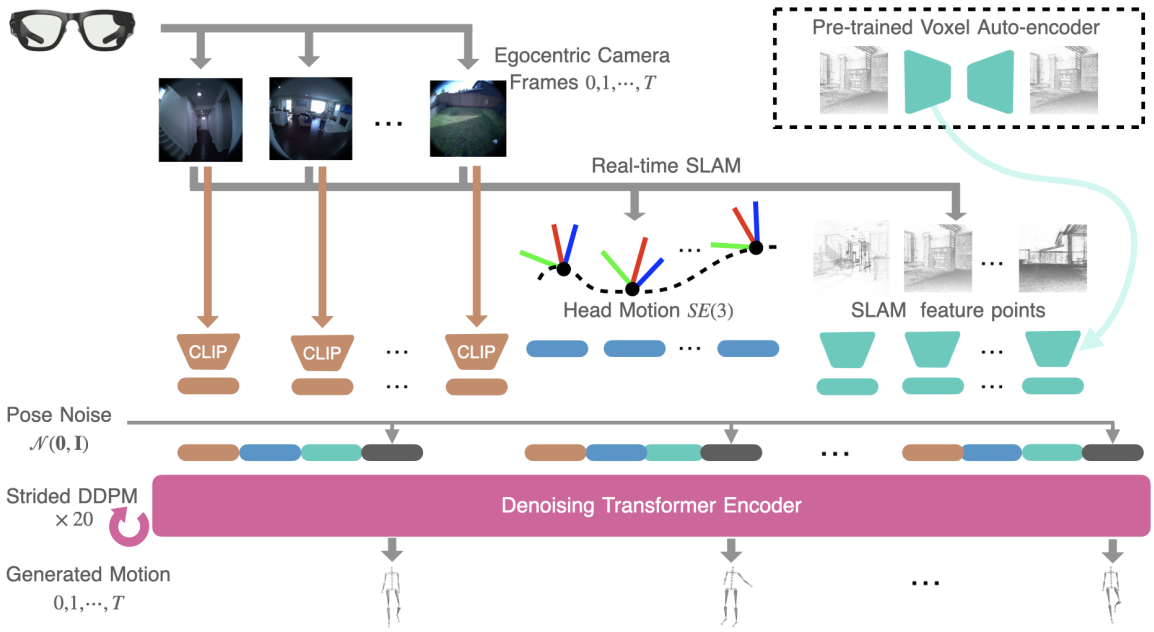

Figure 2. Overview: HMD^2 generates realistic full-body motion that aligns with the signals from a single head-mounted device. Using the image streams from the egocentric camera and head trajectory with the feature cloud from the onboard SLAM system, we employ a diffusion-based framework to generate the wearer's full-body motion.

Figure 3. A typical input sequence from egocentric camera with only few body parts of the wearer intermittently visible, rendering standard full-body reconstruction network backbones ineffective.

motion, and it may be tempting to assume that such input might not be useful for full-body motion reconstruction. However, high-level descriptions of the images that contain scene information, such as "hand reaching to the sink (which is typically at a standard height)" or "a person kicking a football (implicitly indicating that the wearer might also soon interact with the ball)", are actually quite useful for spatial reasoning of the wearer's end effectors. We hypothesize that this observation explains why CLIP embeddings are advantageous in our unique problem setting.

SLAM Point Cloud Embeddings. Visual SLAM algorithms identify static feature points in the environment (*e.g.* corners and edges of furniture) and aggregate them over time to build 3D maps. These points offer crucial environment features to constrain motion generation, akin to pre-scanned scenes utilized in prior work [17, 35]. At each frame, we only consider the SLAM feature points $\boldsymbol{S}$ within a 2m x 2m x 2m volume. The center of the volume is the current device position offset downwards by one meter, similar to prior works [51]. This ensures the model focuses only on relevant spatial information as the wearer moves around. To better handle the noisy and often incomplete SLAM point clouds, we pre-train an autoencoder on the voxelized SLAM point clouds $V(\boldsymbol{S})$ within the bounding volume on all frames in our training dataset and use its encoder $E_S(\cdot)$ to generate point cloud embeddings $E_S(V(\boldsymbol{S}))$. While a new map may not offer much information right away, rich point cloud features could quickly build up if the wearer stays in the same environment for a prolonged period (*e.g.* 15 min) or if they have access to a prebuilt map.

3.2. Conditional Motion Diffusion Model

Given all input signals from the device, $\boldsymbol{c} = \{\boldsymbol{p}, E_I(\boldsymbol{I}), E_S(V(\boldsymbol{S}))\}_{0,1,\cdots,T}$, diffusion models such as DDPM [21] can model the distribution of all motions conditioned on $\boldsymbol{c}$ by progressively introducing distortions (Gaussian diffusion noises) into the motion sequence and learning a neural network model D to reverse these distortions. The sequence of forward distortions can be described by the following equation:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0, \boldsymbol{c}) = \mathcal{N}(\sqrt{\alpha_t}\boldsymbol{x}_0, (1-\alpha_t)\mathbb{I}) = \sqrt{\alpha_t}\boldsymbol{x}_0 + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}, \tag{1}$$

where the motion $\boldsymbol{x} \in \mathbb{R}^{T\times F}$ is represented as a time series with window length T and motion feature dimension denoted as F. Here, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$ denotes the unit Gaussian noise, and $t \in \{0, 1, \cdots, S\}$ signifies the level of distortion, with $t = 0$ indicating no distortion and $t = S$ representing maximum distortion such that $\alpha_S = 0$ and $\boldsymbol{x}_S \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$.

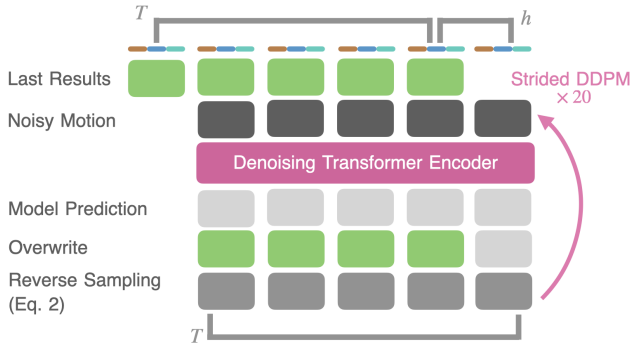

Figure 4. Autoregressive inpainting is performed at each reverse diffusion step to allow long sequence generations both in high- and low-latency settings.

The parameter α_t is a monotonically decreasing scalar that governs the noise schedule. The reverse diffusion process is derived using Bayes' rule and can be expressed as:

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0, \boldsymbol{c}) = \mathcal{N}(\sqrt{\alpha_{t-1}}\boldsymbol{x}_0 + c_t \frac{(\boldsymbol{x}_t - \sqrt{\alpha_t}\boldsymbol{x}_0)}{\sqrt{1-\alpha_t}}, \sigma_t^2 \mathbb{I}), \quad (2)$$

$$c_t = \sqrt{1 - \alpha_{t-1} - \sigma_t^2}, \quad \sigma_t^2 = (1 - \frac{\alpha_t}{\alpha_{t-1}}) \frac{1-\alpha_{t-1}}{1-\alpha_t}. \quad (3)$$

With $\boldsymbol{x}_0$ in Eq. 2 estimated by the neural net module $\hat{\boldsymbol{x}}_0 = D(\boldsymbol{x}_t, \boldsymbol{c}, t)$, we can iteratively generate a sequence of samples $(\boldsymbol{x}_S, \boldsymbol{x}_{S-1}, \ldots, \boldsymbol{x}_1, \boldsymbol{x}_0)$, initiating from $\boldsymbol{x}_S \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$ and progressing towards the desired motion distribution $q(\boldsymbol{x}_0|\boldsymbol{c})$ over S reverse diffusion steps. During model training, we randomly sample t from a uniform distribution $U(0, S)$ for every training data. At inference time, we apply $\bar{S} = 20$ evenly spaced strided reverse diffusion steps [43]. Note that no Gaussian noise is applied to the condition vector $\boldsymbol{c}$. Training loss is defined as:

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{x}_0 \times t \sim U(0,S)} ||D(\boldsymbol{x}_t, \boldsymbol{c}, t) - \boldsymbol{x}_0||^2, \quad (4)$$

We did not find it necessary to include auxiliary loss terms to refine output quality.

Online Inference of Long Sequences. Our motion diffusion model generates up to 4 seconds of motion ($T = 240$ frames). To extend this for longer, coherent motions, previous research [7, 49, 72] suggests generating overlapping windows and enforcing consistency at overlaps during denoising. However, for online generation, we need to remove the dependency on future windows by using an autoregressive approach [22], where each window depends only on the previous one. Specifically, when two windows overlap by $T - h$ frames (i.e., the current window advances by a stride of h), we enforce consistency during each of the $\bar{S}$ denoising steps. After each model evaluation $\hat{\boldsymbol{x}}_0 = D(\boldsymbol{x}_t, \boldsymbol{c}, \boldsymbol{\tau}_i)$, the prediction $\hat{\boldsymbol{x}}_0$ is overwritten by the overlapping prediction from the preceding window:

$$\hat{\boldsymbol{x}}_0 = \hat{\boldsymbol{x}}_0 \odot \boldsymbol{m} + \hat{\boldsymbol{x}}_{s0} \odot (1 - \boldsymbol{m}), \quad (5)$$

where $\boldsymbol{m} \in \mathbb{R}^{T \times F}$ is a constant mask that is zero for the initial $T - h$ frames and one for the last h frames. $\hat{\boldsymbol{x}}_{s0} = \text{cat}(\boldsymbol{x}_0^-[h{:}T], \mathbf{0}^{h \times F})$ denotes the prediction from the previous window, shifted by h frames. $\odot$ denotes element-wise multiplication. Following this inpainting operation, we move to denoising step with the updated $\hat{\boldsymbol{x}}_0$ using Eq. 2. We report the main results of our system with stride $h = 180$.

However, eliminating the need for future windows is insufficient for online inference with minimal latency since a new window of motion is generated only every h frames, resulting in a latency of $(h - 1) \times \delta t$, where $1/\delta t$ is the frame rate. We additionally report our results with $h = 10$, indicating a latency of just 0.17 seconds, close to online requirements. Nonetheless, a smaller h compromises motion quality, as it limits the use of future information. In general, h can be a tunable parameter to trade off quality and latency.

4. Experiments

We conducted a set of experiments to support these claims:

- Our multi-modal conditioning improves motion quality.
- Our system achieved high reconstruction accuracy, motion diversity, and physical realism.
- Our online (low-latency) variant minimally degrades motion quality compared to high-latency inference.
- Our system achieved improved results over state-of-the-art baselines on a large-scale dataset.

Datasets and Experiment setup. To address the limitation of evaluating on synthetic or smaller real-world datasets, we train and evaluate our system on a large-scale, first-of-its-kind real-device dataset Nymeria [41]. This dataset contains paired multi-modal HMD input signals (captured by Project Aria Glasses [13]) and ground-truth full-body motions (with Xsens inertial motion capture system [45]). The dataset covers a diverse range of daily activities and is around 300 hours in size. After initial filtering, we split the data into train, validation, and test split with 202, 3, and 56 hours of data correspondingly. We make sure all subjects and environment scenes in the test split are **unseen** during training, and distributions of subjects' body sizes and activity scripts are roughly unbiased across the test split. We trained our models with a context window of 240 frames (4 sec) for 20 epochs or 3.5 days with 4 GPUs. The inference is done on a single Nvidia A100 GPU and achieves better than real-time throughput of $>$ 70 FPS with an online 0.17s-latency ($h = 10$) model and $>$ 1350 FPS with high-latency (3s, $h = 180$) model.

Baselines. We benchmark our low- and high-latency systems against EgoEgo [36] and AvatarPoser [30], retraining both models on our dataset. For EgoEgo, we bypass its first stage, using Aria Glasses' SLAM for accurate head motion tracking, and test with its long-sequence inference code. For AvatarPoser, we only provide head motion, masking out

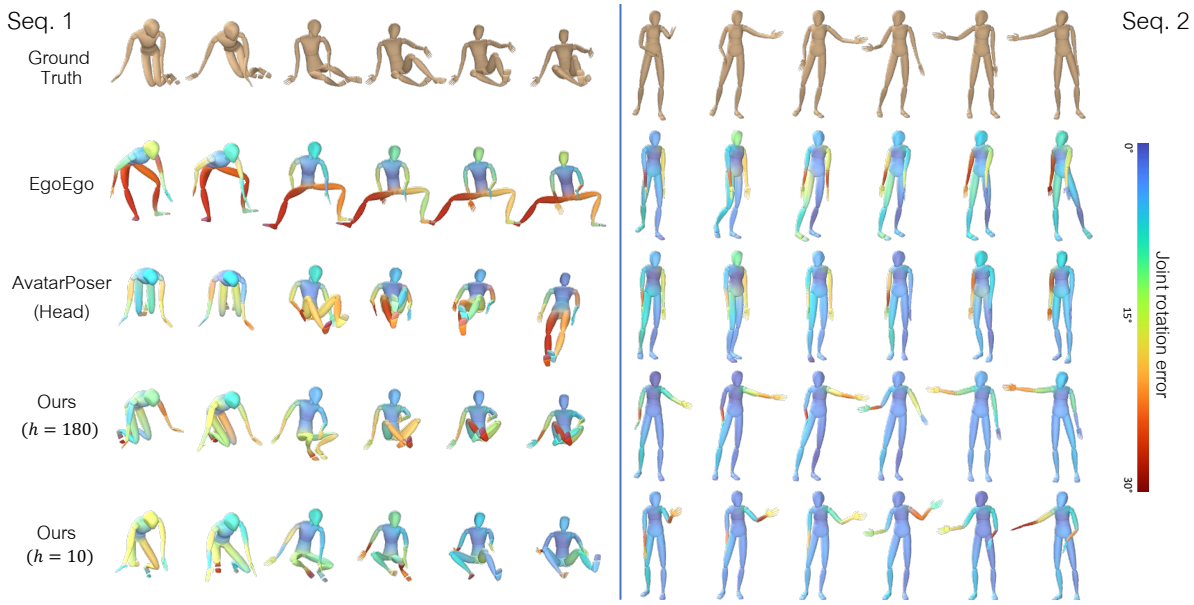

Figure 5. Qualitative comparison between HMD^2 (Ours) and baseline methods.

	MPJPE ↓	Hand PE ↓	FID ↓	Diversity →	Physicality →	Floor Pen. ↓
Ground-truth	0	0	0	16.13	0.56	0
EgoEgo	$16.61^{\pm 1.49}$	$34.64^{\pm 1.64}$	$35.69^{\pm 0.54}$	$20.15^{\pm 0.21}$	$3.68^{\pm 0.74}$	$2.43^{\pm 1.54}$
AvatarPoser (Head)	10.64	21.51	27.61	12.99	1.69	4.21
Ours ($h = 180$)	$\mathbf{8.36}^{\pm 0.08}$	$\mathbf{16.64}^{\pm 0.21}$	$\mathbf{2.16}^{\pm 0.02}$	$\mathbf{15.74}^{\pm 0.29}$	$\mathbf{1.03}^{\pm 0.01}$	$\mathbf{1.03}^{\pm 0.06}$
Ours ($h = 10$)	$9.19^{\pm 0.05}$	$17.67^{\pm 0.06}$	$5.00^{\pm 0.02}$	$15.23^{\pm 0.02}$	$1.30^{\pm 0.10}$	$1.19^{\pm 0.04}$

Table 1. Quantitative results comparing our system with EgoEgo and AvatarPoser.

wrist device input during training and testing. Unlike the Nymeria paper's short-segment evaluations [41], we test all methods with full motion sequences (each around 15min) in an online, autoregressive setting, reflecting real-world use.

Metrics. An ideal solution must balance reconstruction accuracy, motion diversity, and physical realism. For instance, when arms are visible to the HMD camera, generated motions should reflect that. When multiple motions are equally valid, *e.g.* sitting, squatting, or kneeling, predictions should cover all possibilities. Finally, any output motion should be visually realistic and within the distribution of physically plausible human movements. We choose metrics that evaluate a system's capability to balance these three goals.

- **Reconstruction:** we report joint position errors (Mean Per Joint Position Errors, MPJPE, in cm) for all methods. As we use the head frame from Aria as the body reference frame for all methods, we assume zero error on head positions or orientations. Instead, we report position errors of the wrist joints (Hand PE, in cm).
- **Diversity:** Following prior work [16, 46], we report the diversity metric as the mean distance between two same-size randomly sampled subsets from predicted and ground-truth motions in the same latent space as used for FID computation [16].
- **Realism:** we report FID scores measuring the distances in distributions between predicted and ground-truth motions. This is done through training an auto-encoder to construct a motion latent space, following the protocol in Guo *et al.* 2020 [16]. We also report the physicality of motions, following the metric proposed in EDGE [54], which correlates with foot sliding. Lastly, we report the mean floor penetration depth (in cm). Since the floor level varies across time and is non-trivial to estimate for outdoor and complex indoor environments (e.g. the "floor" height for lying in bed should sensibly be the bed height), we adopt a conservative proxy using the lowest joint position of the ground-truth motion across the neighboring 20 seconds.

4.1. Main Results

We evaluated high- ($h = 180$) and low-latency ($h = 10$) variants of our system on the 56-hour (224 sequences) test split, averaging 15 minutes per sequence. These test sequences are **not** cut into short segments to fit the temporal horizon T of the model – all models are tasked to generate the entire sequence coherently, which is closer to practical application setup. Unlike EgoEgo, where statistics are re-

ported using the best among 200 repetitions, we report the mean and standard deviation of all repetitions. As our test set is very large (e.g. the AMASS [42] testing subset used in AvatarPoser contains just two hours of motion), we only run eight repetitions for each of the 224 sequences.

Quantitative Results. The main quantitative results are summarized in Table 1, with a finer-grained analysis provided in the supplementary. Our system achieved superior performance across all three metric axes of reconstruction, diversity, and realism. As expected, the online variant of our system degrades performance slightly, given inaccessibility of future sensor information, but still outperforms baselines.

Our adapted version of AvatarPoser (referred to as AvatarPoser (Head) in Table 1) performs well, but its frame-by-frame prediction lacks temporal coherence, reducing realism. As a regression model, it captures only the average trend in training data, leading to lower diversity scores. Unlike our multi-modal approach, it lacks environmental awareness, impacting performance (Fig. 5). EgoEgo generates plausible motions but has two key issues. First, it produces discontinuities during long motion inference, which affect realism metrics. Second, EgoEgo tends to produce overly dynamic arm movements, similar to how some image diffusion models create stylized rather than naturalistic outputs. This leads to higher Hand Position Errors and contributes to increased MPJPE and Diversity scores compared to ground truth. While all the metrics in Table 1 are measured as mean across all runs, we additionally report MPJPE of the best-case run: 8.246, 14.678, for HMD^2 and EgoEgo (AvatarPoser stays the same). Compared to Table 1, errors for EgoEgo are noticeably lower but are still behind Ours.

In summary, our system uniquely balances the accuracy of motion reconstruction and fidelity and diversity of motion generation, surpassing baseline methods. The online variant of our system achieves 0.17-second latency with only a slight degradation in terms of performance, though the gap leaves room for future research and improvement.

Qualitative Examples. Fig. 5 visually compares all methods on two motion subsequences from the test set. *Sequence 1* shows a complex transition from kneeling to sitting. Regression models like AvatarPoser struggle in under-constrained scenarios, either abruptly switching between poses or averaging them into unnatural ones (e.g., a floating avatar in the last frame). EgoEgo, as a generative model, produces plausible motions but lacks the context to match the ground truth given only head motion. *Sequence 2* demonstrates another important advantage of our model – making use of the semantic features from color images. In this ground truth motion, the hands are raised and visible in the camera alternately. We successfully reproduce similar arm movements by conditioning on the CLIP embeddings while both baselines have the arms down.

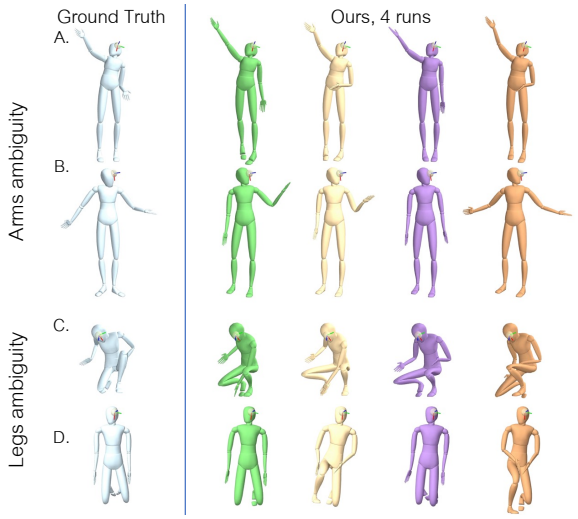

Figure 6. Our system can predict diverse outcomes from identical input (head pose marked as a sphere with coordinate system).

The generative nature of our model also allows us to produce diverse motions in case of ambiguities. Fig. 6 shows several examples: in the left column, our model generates various plausible states when hands are not visible, such as different poses for the non-visible left hand (seq. A). The right column shows cases with equally possible leg positions, like kneeling vs. squatting (seq. C).

4.2. Additional Analysis

Ablations. We ablated our system by removing the point cloud encoder branch (w/o PC) and/or the raw egocentric video branch (w/o CLIP). The results are summarized in Table 2, demonstrating the importance of multi-modal scene and motion conditions in our system.

Even without point cloud and CLIP embeddings, our system generates temporally coherent and realistic full-body motions, capturing diverse motion distributions. However, ambiguity arises with head movement alone, such as distinguishing between standing and sitting. Without environmental context, the system might randomly generate or switch between these actions, affecting realism metrics (FID & Floor Penetration Depth). Table 2 shows that point cloud embeddings help align motions with ground truth and reduce environment interpenetration, improving realism. The image encoder also enhances reconstruction accuracy by using semantic clues, particularly when hands are visible. This reduces MPJPE by encouraging specific poses, however it also mildly affects the realism of motion, hence Physicality metric slightly degrades. Fig. 7 illustrates that PC embeddings enable correct sitting motion detection, while image embeddings improve hand motion accuracy. Together, they produce more accurate and realistic results.

	MPJPE ↓	Hand PE ↓	FID ↓	Diversity →	Physicality →	Floor Pen. ↓
Ground-truth	0	0	0	16.13	0.56	0
Ours, w/o PC, w/o CLIP	$9.28^{\pm 0.23}$	$19.47^{\pm 0.36}$	$6.75^{\pm 0.08}$	$14.44^{\pm 0.30}$	$0.90^{\pm 0.01}$	$3.29^{\pm 0.31}$
Ours, w/ PC, w/o CLIP	$8.97^{\pm 0.10}$	$20.38^{\pm 0.28}$	$3.68^{\pm 0.03}$	$15.29^{\pm 0.42}$	$\mathbf{0.86}^{\pm 0.00}$	$\mathbf{0.99}^{\pm 0.07}$
Ours, w/o PC, w/ CLIP	$8.57^{\pm 0.11}$	$\mathbf{16.32}^{\pm 0.22}$	$6.17^{\pm 0.02}$	$14.79^{\pm 0.22}$	$1.01^{\pm 0.01}$	$2.15^{\pm 0.15}$
Ours, w/ PC, w/ CLIP	$\mathbf{8.36}^{\pm 0.08}$	$16.64^{\pm 0.21}$	$\mathbf{2.16}^{\pm 0.02}$	$\mathbf{15.74}^{\pm 0.29}$	$1.03^{\pm 0.01}$	$1.03^{\pm 0.06}$

Table 2. **Ablation study.** HMD^2 leverages both point cloud (PC) and egocentric video information (CLIP) to reduce per-joint error while keeping the realism and physical plausibility of the motions.

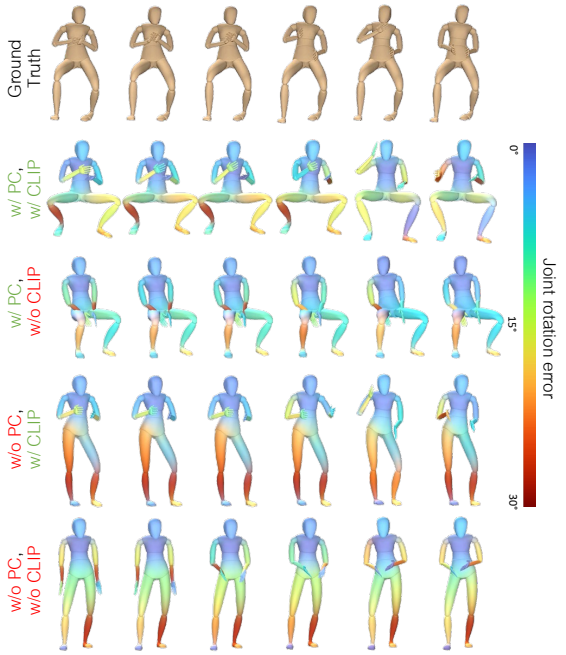

Figure 7. Example motion when ablating the point cloud (PC) or video (CLIP) branches.

Error Distribution. As we evaluate on a large scale dataset of realistic daily activities, the metric statistics could be skewed and dominated by mundane actions such as sitting or standing still, or walking from A to B. The more interesting and challenging scenarios that highlight core issues may fall into a long-tail distribution and be obscured by the mean error. To this end, we also report the top 5% errors in Table 3, which is more representative of improvements we expect from our approach.

	MPJPE ↓	Hand PE ↓	Floor Pen. ↓
Ground-truth	0	0	0
Ours, w/o PC, w/o CLIP	$18.31^{\pm 0.89}$	$40.15^{\pm 1.17}$	$12.91^{\pm 1.75}$
Ours, w PC, w/o CLIP	$16.65^{\pm 0.44}$	$41.68^{\pm 1.05}$	$\mathbf{3.97}^{\pm 0.32}$
Ours, w/o PC, w/ CLIP	$\underline{16.30}^{\pm 0.55}$	$\mathbf{34.25}^{\pm 0.90}$	$8.28^{\pm 0.78}$
Ours, w/ PC, w/ CLIP	$\mathbf{15.49}^{\pm 0.38}$	$\underline{34.86}^{\pm 0.92}$	$\underline{4.22}^{\pm 0.28}$

Table 3. Ablation study on the top 5% per-frame errors (95% performance), showing significant reduction of peak errors by our multi-modal conditioning.

5. Conclusions and Discussions

We introduced HMD^2, a diffusion-based framework for online motion generation from a single head-mounted device. By combining camera-based image embeddings with SLAM-derived head trajectories and semi-dense point clouds, HMD^2 produces diverse, natural motions aligned with the environment. Our evaluation across various settings and activities shows that HMD^2 outperforms state-of-the-art methods in accuracy, diversity, and realism.

Our insight of leveraging egocentric image features and the capability of modern SLAM systems opens up many new opportunities. For instance, we could incorporate more comprehensive contextual information available from recent advancements in image understanding, including depth estimation from monocular videos, panoptic segmentation, or scene reconstruction through neural radiance fields or 3D Gaussian Splats. Additionally, we envision leveraging video embeddings over extended context windows, potentially from visual language models (VLMs) [2], to refine context conditions further.

Currently, the performance of our system is still limited by available context information. For example, the CLIP embeddings cannot provide precise spatial information, so they fall short of constraining the precise pose of the hands even when they are visible. The noisy and sparse point clouds are less ideal than dense depth maps for accurate environment contact information; the errors from the SLAM reconstruction can also propagate to our system. On the other hand, incorporating more and denser input streams pose challenge in runtime performance. We will further elaborate on the above limitations in the supplementary.

Acknowledgments: We thank the Surreal team, especially Svetoslav Kolev, Hyo Jin Kim, Rowan Postyeni, and Renzo De Nardi, for their valuable discussions and help in the project. Gerard Pons-Moll is supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A, by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 409792180 (Emmy Noether Programme, project: Real Virtual Humans). Gerard Pons-Moll is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645 and is supported by the Carl Zeiss Foundation. Yifeng Jiang is partially supported by the Wu Tsai Human Performance Alliance at Stanford University.

References

[1] Project aria machine perception services (accessed January 7, 2025), https://facebookresearch.github.io/projectaria_tools/docs/ARK/mps 3

[2] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023) 8

[3] Akada, H., Wang, J., Golyanik, V., Theobalt, C.: 3d human pose perception from egocentric stereo videos. In: Computer Vision and Pattern Recognition (CVPR) (2024) 2

[4] Akada, H., Wang, J., Shimada, S., Takahashi, M., Theobalt, C., Golyanik, V.: UnrealEgo: A new dataset for robust egocentric 3d human motion capture. In: European Conference on Computer Vision (ECCV) (2022) 2

[5] Alexanderson, S., Nagy, R., Beskow, J., Henter, G.E.: Listen, denoise, action! audio-driven motion synthesis with diffusion models. ACM Transactions on Graphics (TOG) **42**(4), 1–20 (2023) 3

[6] Cao, Z., Gao, H., Mangalam, K., Cai, Q., Vo, M., Malik, J.: Long-term human motion prediction with scene context. In: ECCV (2020) 3

[7] Castillo, A., Escobar, M., Jeanneret, G., Pumarola, A., Arbeláez, P., Thabet, A., Sanakoyeu, A.: Bodiffusion: Diffusing sparse observations for full-body human motion synthesis. ICCV (2023) 2, 3, 5

[8] Cha, Y.W., Price, T., Wei, Z., Lu, X., Rewkowski, N., Chabra, R., Qin, Z., Kim, H., Su, Z., Liu, Y., et al.: Towards fully mobile 3d face, body, and environment capture using only head-worn cameras. IEEE transactions on visualization and computer graphics **24**(11), 2993–3004 (2018) 2

[9] Chiquier, M., Vondrick, C.: Muscles in action. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22091–22101 (2023) 2

[10] Choi, J., Shim, D., Kim, H.J.: Diffupose: Monocular 3d human pose estimation via denoising diffusion probabilistic model. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3773–3780. IEEE (2023) 3

[11] Ci, H., Wu, M., Zhu, W., Ma, X., Dong, H., Zhong, F., Wang, Y.: Gfpose: Learning 3d human pose prior with gradient fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4800–4810 (2023) 3

[12] 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) 2, 3

[13] Engel, J., Somasundaram, K., Goesele, M., Sun, A., Gamino, A., Turner, A., Talattof, A., Yuan, A., Souti, B., Meredith, B., Peng, C., Sweeney, C., Wilson, C., Barnes, D., DeTone, D., Caruso, D., Valleroy, D., Ginjupalli, D., Frost, D., Miller, E., Mueggler, E., Oleinik, E., Zhang, F., Somasundaram, G., Solaira, G., Lanaras, H., Howard-Jenkins, H., Tang, H., Kim, H.J., Rivera, J., Luo, J., Dong, J., Straub, J., Bailey, K., Eckenhoff, K., Ma, L., Pesqueira, L., Schwesinger, M., Monge, M., Yang, N., Charron, N., Raina, N., Parkhi, O., Borschowa, P., Moulon, P., Gupta, P., Mur-Artal, R., Pennington, R., Kulkarni, S., Miglani, S., Gondi, S., Solanki, S., Diener, S., Cheng, S., Green, S., Saarinen, S., Patra, S., Mourikis, T., Whelan, T., Singh, T., Balntas, V., Baiyya, V., Dreewes, W., Pan, X., Lou, Y., Zhao, Y., Mansour, Y., Zou, Y., Lv, Z., Wang, Z., Yan, M., Ren, C., Nardi, R.D., Newcombe, R.: Project Aria: A new tool for egocentric multi-modal AI research (2023) 2, 3, 5

[14] Foo, L.G., Gong, J., Rahmani, H., Liu, J.: Distribution-aligned diffusion for human mesh recovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9221–9232 (2023) 3

[15] Gong, J., Foo, L.G., Fan, Z., Ke, Q., Rahmani, H., Liu, J.: Diffpose: Toward more reliable 3d pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13041–13051 (2023) 3

[16] Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M., Cheng, L.: Action2motion: Conditioned generation of 3d human motions. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2021–2029 (2020) 6

[17] Guzov, V., Mir, A., Sattler, T., Pons-Moll, G.: Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2021) 3, 4

[18] Han, B., Peng, H., Dong, M., Xu, C., Ren, Y., Shen, Y., Li, Y.: Amd autoregressive motion diffusion. AAAI (2024) 3

[19] Hassan, M., Ceylan, D., Villegas, R., Saito, J., Yang, J., Zhou, Y., Black, M.J.: Stochastic scene-aware motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11374–11384 (2021) 3

[20] Henter, G.E., Alexanderson, S., Beskow, J.: Moglow: Probabilistic and controllable motion synthesis using

normalising flows. ACM Transactions on Graphics (TOG) **39**(6), 1–14 (2020) 3

[21] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models (2020) 3, 4

[22] Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. arXiv:2204.03458 (2022) 5

[23] Holden, D., Kanoun, O., Perepichka, M., Popa, T.: Learned motion matching. ACM Transactions on Graphics (TOG) **39**(4), 53–1 (2020) 3

[24] Holden, D., Komura, T., Saito, J.: Phase-functioned neural networks for character control. ACM Transactions on Graphics (TOG) **36**(4), 1–13 (2017) 3

[25] Holmquist, K., Wandt, B.: Diffpose: Multi-hypothesis human pose estimation using diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15977–15987 (2023) 3

[26] Huang, S., Wang, Z., Li, P., Jia, B., Liu, T., Zhu, Y., Liang, W., Zhu, S.C.: Diffusion-based generation, optimization, and planning in 3d scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16750–16761 (2023) 3

[27] 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, (Proc. SIGGRAPH Asia) **37**(6), 185:1–185:15 (nov 2018) 2

[28] Jiang, H., Grauman, K.: Seeing invisible poses: Estimating 3d body pose from egocentric video. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3501–3509. IEEE (2017) 2

[29] Jiang, J., Streli, P., Meier, M., Fender, A., Holz, C.: Egoposer: Robust real-time ego-body pose estimation in large scenes. arXiv preprint arXiv:2308.06493 (2023) 2

[30] 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) 2, 5

[31] Jiang, Y., Ye, Y., Gopinath, D., Won, J., Winkler, A.W., Liu, C.K.: Transformer inertial poser: Real-time human motion reconstruction from sparse imus with simultaneous terrain generation. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–9 (2022) 2

[32] Kang, T., Lee, K., Zhang, J., Lee, Y.: Ego3dpose: Capturing 3d cues from binocular egocentric views. In: SIGGRAPH Asia 2023 Conference Papers. pp. 1–10 (2023) 2

[33] Kaufmann, M., Zhao, Y., Tang, C., Tao, L., Twigg, C., Song, J., Wang, R., Hilliges, O.: Em-pose: 3d human pose estimation from sparse electromagnetic trackers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11510–11520 (2021) 2

[34] Lee, J., Joo, H.: Locomotion-action-manipulation: Synthesizing human-scene interactions in complex 3d environments. ICCV (2023) 3

[35] Lee, J., Joo, H.: Mocap everyone everywhere: Lightweight motion capture with smartwatches and a head-mounted camera. Computer Vision and Pattern Recognition (CVPR) (2024) 3, 4

[36] Li, J., Liu, K., Wu, J.: Ego-body pose estimation via ego-head pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17142–17151 (2023) 2, 5

[37] Ling, H.Y., Zinno, F., Cheng, G., Van De Panne, M.: Character controllers using motion vaes. ACM Transactions on Graphics (TOG) **39**(4), 40–1 (2020) 3

[38] Liu, Y., Yang, J., Gu, X., Guo, Y., Yang, G.Z.: Egohmr: Egocentric human mesh recovery via hierarchical latent diffusion model. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 9807–9813. IEEE (2023) 2

[39] Luo, Z., Hachiuma, R., Yuan, Y., Kitani, K.: Dynamics-regulated kinematic policy for egocentric pose estimation. Advances in Neural Information Processing Systems **34**, 25019–25032 (2021) 2

[40] Luo, Z., Iwase, S., Yuan, Y., Kitani, K.: Embodied scene-aware human pose estimation. In: Advances in Neural Information Processing Systems (2022) 3

[41] Ma, L., Ye, Y., Hong, F., Guzov, V., Jiang, Y., Postyeni, R., Pesqueira, L., Gamino, A., Baiyya, V., Kim, H.J., Bailey, K., Fosas, D.S., Liu, C.K., Liu, Z., Engel, J., De Nardi, R., Newcombe, R.: Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In: European Conference on Computer Vision (ECCV) (2024) 2, 5, 6

[42] Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: International Conference on Computer Vision. pp. 5442–5451 (Oct 2019) 7

[43] Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. pp. 8162–8171. PMLR (2021) 5

[44] Park, J., Kaai, K., Hossain, S., Sumi, N., Rambhatla, S., Fieguth, P.: Domain-guided spatio-temporal self-attention for egocentric 3d pose estimation. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 1837–1849 (2023) 2

[45] Paulich, M., Schepers, M., Rudigkeit, N., Bellusci, G.: Xsens MTw Awinda: Miniature Wireless Inertial-Magnetic Motion Tracker for Highly

Accurate 3D Kinematic Applications (05 2018). https://doi.org/10.13140/RG.2.2.23576.49929 5

[46] Raab, S., Leibovitch, I., Li, P., Aberman, K., Sorkine-Hornung, O., Cohen-Or, D.: Modi: Unconditional motion synthesis from diverse data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13873–13883 (2023) 6

[47] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021) 2, 3

[48] Rhodin, H., Richardt, C., Casas, D., Insafutdinov, E., Shafiei, M., Seidel, H.P., Schiele, B., Theobalt, C.: Egocap: egocentric marker-less motion capture with two fisheye cameras. ACM Transactions on Graphics (TOG) **35**(6), 1–11 (2016) 2

[49] Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior. ICLR (2023) 3, 5

[50] Shi, Y., Wang, J., Jiang, X., Dai, B.: Controllable motion diffusion model. arXiv preprint arXiv:2306.00416 (2023) 3

[51] Starke, S., Zhang, H., Komura, T., Saito, J.: Neural state machine for character-scene interactions. ACM Trans. Graph. **38**(6), 209–1 (2019) 3, 4

[52] Tome, D., Alldieck, T., Peluse, P., Pons-Moll, G., Agapito, L., Badino, H., De la Torre, F.: Selfpose: 3d egocentric pose estimation from a headset mounted camera. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020) 2

[53] Tseng, J., Castellon, R., Liu, K.: Edge: Editable dance generation from music. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 448–458 (June 2023) 3

[54] Tseng, J., Castellon, R., Liu, K.: Edge: Editable dance generation from music. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 448–458 (2023) 6

[55] Van Wouwe, T., Lee, S., Falisse, A., Delp, S., Liu, C.K.: Diffusion inertial poser: Human motion reconstruction from arbitrary sparse imu configurations. arXiv preprint arXiv:2308.16682 (2023) 3

[56] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 3

[57] von Marcard, T., Rosenhahn, B., Black, M., Pons-Moll, G.: Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. Computer Graphics Forum 36(2), Proceedings of the 38th Annual Conference of the European Association for Computer Graphics (Eurographics) pp. 349–360 (2017) 2

[58] Wang, J., Cao, Z., Luvizon, D., Liu, L., Sarkar, K., Tang, D., Beeler, T., Theobalt, C.: Egocentric whole-body motion capture with fisheyevit and diffusion-based motion refinement. arXiv preprint arXiv:2311.16495 (2023) 2

[59] 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) 2

[60] 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) 3

[61] Wang, J., Rong, Y., Liu, J., Yan, S., Lin, D., Dai, B.: Towards diverse and natural scene-aware 3d human motion synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20460–20469 (2022) 3

[62] Wang, Z., Chen, Y., Liu, T., Zhu, Y., Liang, W., Huang, S.: Humanise: Language-conditioned human motion generation in 3d scenes. Advances in Neural Information Processing Systems **35**, 14959–14971 (2022) 3

[63] Xu, W., Chatterjee, A., Zollhoefer, M., Rhodin, H., Fua, P., Seidel, H.P., Theobalt, C.: Mo 2 cap 2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera. IEEE transactions on visualization and computer graphics **25**(5), 2093–2101 (2019) 2

[64] Yang, D., Kang, J., Ma, L., Greer, J., Ye, Y., Lee, S.H.: Divatrack: Diverse bodies and motions from acceleration-enhanced three-point trackers. Eurographics (2024) 2

[65] Yang, J., Chen, T., Qin, F., Lam, M.S., Landay, J.A.: Hybridtrak: Adding full-body tracking to vr using an off-the-shelf webcam. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. pp. 1–13 (2022) 2

[66] Yi, X., Zhou, Y., Habermann, M., Golyanik, V., Pan, S., Theobalt, C., Xu, F.: Egolocate: Real-time motion capture, localization, and mapping with sparse body-mounted sensors. ACM Transactions on Graphics (TOG) **42**(4) (2023) 3

[67] 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) 2

[68] 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) 2

[69] Yin, W., Tu, R., Yin, H., Kragic, D., Kjellström, H., Björkman, M.: Controllable motion synthesis and reconstruction with autoregressive diffusion models. arXiv preprint arXiv:2304.04681 (2023) 3

[70] Yuan, Y., Kitani, K.: 3d ego-pose estimation via imitation learning. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 735–750 (2018) 2

[71] Yuan, Y., Kitani, K.: Ego-pose estimation and forecasting as real-time pd control. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019) 2

[72] Zhang, Q., Song, J., Huang, X., Chen, Y., Liu, M.Y.: Diffcollage: Parallel generation of large content with diffusion models. arXiv preprint arXiv:2303.17076 (2023) 5

[73] Zhang, S., Bhatnagar, B.L., Xu, Y., Winkler, A., Kadlecek, P., Tang, S., Bogo, F.: Rohm: Robust human motion reconstruction via diffusion. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024) 3

[74] Zhang, S., Ma, Q., Zhang, Y., Aliakbarian, S., Cosker, D., Tang, S.: Probabilistic human mesh recovery in 3d scenes from egocentric views. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). p. 7955–7966. IEEE, Paris, France (Oct 2023). https://doi.org/10.1109/ICCV51070.2023.00734 3

[75] Zhang, Y., You, S., Gevers, T.: Automatic calibration of the fisheye camera for egocentric 3d human pose estimation from a single image. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1772–1781 (2021) 2

[76] Zhang, Z., Liu, R., Aberman, K., Hanocka, R.: Tedi: Temporally-entangled diffusion for long-term motion synthesis. arXiv preprint arXiv:2307.15042 (2023) 3

[77] Zhao, D., Wei, Z., Mahmud, J., Frahm, J.M.: Egoglass: Egocentric-view human pose estimation from an eyeglass frame. In: 2021 International Conference on 3D Vision (3DV). pp. 32–41. IEEE (2021) 2

[78] Zhao, K., Wang, S., Zhang, Y., Beeler, T., , Tang, S.: Compositional human-scene interaction synthesis with semantic control. In: European conference on computer vision (ECCV) (Oct 2022) 3

[79] Zhao, K., Zhang, Y., Wang, S., Beeler, T., Tang, S.: DIMOS: Synthesizing diverse human motions in 3d indoor scenes. In: International Conference on Computer Vision (ICCV) (2023) 3

[80] Zheng, X., Su, Z., Wen, C., Xue, Z., Jin, X.: Realistic full-body tracking from sparse observations via joint-level modeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14678–14688 (2023) 2

[81] Zhou, W., Dou, Z., Cao, Z., Liao, Z., Wang, J., Wang, W., Liu, Y., Komura, T., Wang, W., Liu, L.: Emdm: Efficient motion diffusion model for fast, high-quality motion generation. arXiv preprint arXiv:2312.02256 (2023) 3