UniEgoMotion: A Unified Model for Egocentric Motion Reconstruction, Forecasting, and Generation

Chaitanya Patel[1] Hiroki Nakamura[2] Yuta Kyuragi[1,3]
Kazuki Kozuka[2] Juan Carlos Niebles[1] Ehsan Adeli[1]

[1]Stanford University [2]Panasonic Holdings Corporation [3]Panasonic R&D Company of America

https://chaitanya100100.github.io/UniEgoMotion/

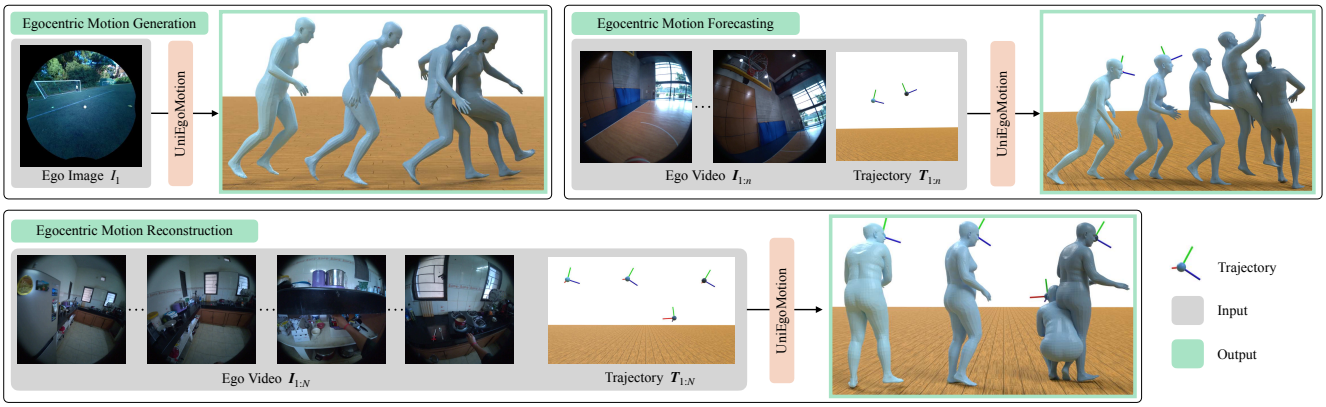

Figure 1. UniEgoMotion is a unified, scene-aware motion model designed for egocentric settings: (1) It generates plausible future motion from a single egocentric image – for example, predicting how you might take your shot on goal. (2) It forecasts upcoming motion using past egocentric video and ego-device trajectory, showing how you could complete your run-up to score. (3) It reconstructs accurate 3D motion from past egocentric observations, showing how you squatted down to reach the lower cabinet.

Abstract

Egocentric human motion generation and forecasting with scene-context is crucial for enhancing AR/VR experiences, improving human-robot interaction, advancing assistive technologies, and enabling adaptive healthcare solutions by accurately predicting and simulating movement from a first-person perspective. However, existing methods primarily focus on third-person motion synthesis with structured 3D scene contexts, limiting their effectiveness in real-world egocentric settings where limited field of view, frequent occlusions, and dynamic cameras hinder scene perception. To bridge this gap, we introduce Egocentric Motion Generation and Egocentric Motion Forecasting, two novel tasks that utilize first-person images for scene-aware motion synthesis without relying on explicit 3D scene. We propose UniEgoMotion, a unified conditional motion diffusion model with a novel head-centric motion representation tailored for egocentric devices. UniEgoMotion's simple yet effective design supports egocentric motion reconstruction, forecasting, and generation from first-person visual inputs in a unified framework. Unlike previous works that overlook scene semantics, our model effectively extracts image-based scene context to infer plausible 3D motion. To facilitate training, we introduce EE4D-Motion, a large-scale dataset derived from EgoExo4D, augmented with pseudo-ground-truth 3D motion annotations. UniEgoMotion achieves state-of-the-art performance in egocentric motion reconstruction and is the first to generate motion from a single egocentric image. Extensive evaluations demonstrate the effectiveness of our unified framework, setting a new benchmark for egocentric motion modeling and unlocking new possibilities for egocentric applications.

1. Introduction

Egocentric human motion reconstruction, forecasting, and generation are fundamental for AR/VR, assistive technologies, and healthcare applications. The egocentric camera provides a personalized first-person perspective, enabling interactive and adaptive experiences. Imagine you are learning to play soccer while wearing smart glasses. Egocen-

tric motion **reconstruction** can help analyze and refine your kicking technique. Egocentric motion **forecasting** (prediction) can show how you can continue your current run-up to execute a strike. Egocentric motion **generation** can simulate how you might score a goal from your current position and angle. These capabilities also extend to movement analysis in healthcare, aiding in gait assessment, early detection of neurological and vestibular disorders, and fall prediction.

Despite its potential, egocentric motion understanding remains a challenging problem. The front-facing egocentric camera provides only partial visibility of the user's body, forcing models to infer motion from a dynamic, first-person viewpoint with frequent occlusions and motion blur. As a result, research on egocentric motion forecasting has been limited [95], and to our knowledge, no prior studies have explored egocentric motion generation. Most context-aware motion forecasting and generation works assume explicit 3D scene context in the form of point cloud [59, 73, 102], mesh [83], voxel grid [10, 33], signed distance field [90, 98], or object geometry [12, 46, 49]. Such 3D scene context is not available in many real-world egocentric applications. While RGB videos are the most accessible modality to acquire scene context, very few works use image-based context for motion synthesis. The most relevant prior works use a third-person scene-wide RGB image to forecast [8] or generate [84] human motion *within* that scene. However, these methods cannot generalize to egocentric settings where the wider scene context is unavailable due to the limited field of view and future motion may extend beyond the visible region, requiring strong motion priors.

To this end, we establish two novel tasks of scene-aware egocentric motion generation and forecasting only from egocentric images without requiring 3D scene context. We introduce UniEgoMotion, a unified model for egocentric motion reconstruction, forecasting, and generation (see Figure 1). In particular, it can (1) reconstruct motion using an input ego video and ego-device's inertial SLAM trajectory, (2) forecast motion based on past egocentric inputs, and (3) generate motion from a single egocentric image. Unlike prior works [42, 48, 92, 95] that discard scene semantics, UniEgoMotion leverages egocentric image-based scene context to predict plausible and accurate 3D motion from an egocentric viewpoint. To train UniEgoMotion, we present EE4D-Motion, a new dataset derived from the large-scale EgoExo4D [23] dataset. We augment EgoExo4D videos with paired pseudo-ground-truth 3D motion annotations using a comprehensive motion fitting pipeline, enabling training on in-context egocentric video-motion pairs.

At its core, UniEgoMotion is a transformer-based [80] conditional motion diffusion model that enables flexible conditioning via cross-attention. To establish scene context, it leverages a robust image encoder, initialized with strong pretraining [62], to extract fine-grained visual features, enabling precise mapping of the visible environment while constructing a comprehensive prior of the unseen areas for holistic motion synthesis. During training, we strategically mask conditioning inputs (ego images and 3D device trajectory) such that they support both egocentric reconstruction and generation during inference. Egocentric forecasting is achieved through diffusion inpainting [55] during inference, which utilizes the learned egocentric motion diffusion prior to predict future motion based on past motion reconstruction. Unlike pelvis-centric motion representations used in motion synthesis literature [26, 77], UniEgoMotion adopts a head-centric representation, making it more aligned with egocentric devices. In addition to generation and forecasting, UniEgoMotion also outperforms state-of-the-art methods on egocentric motion reconstruction [42, 48, 92] task.

In summary, we make the following contributions:

1. We introduce two novel tasks—Egocentric Motion Generation and Egocentric Motion Forecasting— expanding the scope of motion modeling for applications of wearable egocentric devices.
2. We propose UniEgoMotion, a novel unified egocentric motion model that performs reconstruction, forecasting, and generation in a single framework. It surpasses state-of-the-art baselines on egocentric motion reconstruction and, to our knowledge, is the first model to generate motion from a single egocentric image.
3. We present EE4D-Motion, a large-scale dataset of egocentric video-motion pairs, enabling the video-based scene context-aware human motion modeling.

2. Related Work

Scene-aware Motion Generation: Motion generation has been extensively studied in various settings, including character control [35, 39, 40, 53], animation [31, 34, 50, 69, 75], action-to-motion synthesis [25, 67, 77, 87], and more recently, text-to-motion generation [3, 13, 20, 26, 27, 68, 77, 96, 97]. Scene-aware motion generation focuses on generating motion grounded in a given scene context. Most approaches condition motion generation on explicit 3D scene representations, such as scene point clouds [2, 88, 102], meshes [83], voxel grids [10, 33, 43, 74], signed distance fields [32, 89, 93, 98], or specific object geometries [12, 46, 49, 100]. Some works leverage 3D scene context for navigational motion generation [83, 101], while others model human-scene [2, 10, 32, 33, 43, 74, 88, 89, 93, 98] and human-object interactions [12, 46, 49, 74, 100].

However, capturing high-quality 3D scene data requires complex setups or extensive offline reconstruction [24, 63, 79], making it impractical for real-world egocentric applications. In contrast, RGB images are easily accessible but pose challenges in extracting relevant scene context due to their limited field of view, dynamic motion, and occlusions. Yet, motion generation from image-based scene context re-

mains underexplored, with [84] being the only work proposing a two-stage GAN-based [21] model to generate human motion from a single wide-scene RGB image. We tackle a more challenging yet practical egocentric setting by introducing the egocentric motion generation task, which generates context-aware motion from a single egocentric image.

Motion Forecasting: Motion forecasting, i.e., predicting future motion based on past motion, has been widely explored in a context-independent setting. Approaches range from traditional models [47, 76, 82] to modern deep learning, including MLPs [29], RNNs [11, 18, 41, 60, 81], graph convolutional networks [14, 28], Transformers [1, 6, 58, 61], RL controllers [95], and diffusion models [4].

Similar to motion generation, scene-aware motion forecasting assumes access to a clean 3D scene as context to predict future motion [12, 59, 73, 90, 102], with a few exceptions [8, 95]. [8] uses a single wide-scene RGB image along with past poses to predict future motion. It proposes a three-stage method: stochastic goal prediction via a VAE [45], deterministic trajectory prediction, and deterministic pose generation. [95] applies an RL-based controller for motion forecasting from egocentric devices, though its evaluation is limited to simple actions like walking and running. In contrast, we address scene-aware motion forecasting in an egocentric setting using diffusion modeling [38], enabling more diverse and complex motion predictions.

Egocentric Motion Reconstruction: Unlike downward-facing camera setting [54, 64, 71, 78, 85], egocentric motion reconstruction focuses on front-facing ego cameras, which are more common in publicly available devices [15]. This introduces significant challenges due to limited body visibility and requires strong motion priors. Many works use simulation-based physical motion priors [56, 94, 95] or diffusion motion priors [30, 42, 48, 92] learned from large motion datasets. However, [48, 95] use input video to compute optical flow to estimate the ego-device's 3D trajectory, discarding valuable scene context. [9, 42, 92] depend exclusively on accurate ego-device trajectory computed via SLAM, without incorporating scene context. These approaches are suboptimal for activities where head motion is minimal, such as cooking or playing a music instrument. The closest work to ours is [30], which integrates scene point clouds and ego-image features for scene-aware egocentric motion reconstruction. In contrast, our approach does not rely on point cloud input; instead we leverage only egocentric images to capture scene context. Some concurrent works [16, 86] finetune language models on motion and text from the Nymeria [17] dataset, conditioning on egocentric inputs, for autoregressive motion reconstruction and understanding. Note that prior works often refer to this task as 'motion generation' due to its generative nature. However, we differentiate between egocentric motion generation and reconstruction based on available sensory information.

3. Method

3.1. Problem Formulation

Let $\boldsymbol{I}_{1:N} = (I_1, I_2, \cdots, I_N)$ denote a sequence of egocentric video frames captured by a head-mounted camera, where each RGB frame $I_i \in \mathbb{R}^{H\times W\times 3}$. Let $\boldsymbol{T}_{1:N} = (T_1, T_2, \cdots, T_N)$ represent the camera's 6-DOF trajectory. Modern wearable devices equipped with state-of-the-art inertial SLAM systems [15] can compute $\boldsymbol{T}_{1:N}$ in real-time. Let $\boldsymbol{X}_{1:N} = (X_1, X_2, \cdots, X_N)$ denote the 3D human motion of the user wearing the camera. Each pose X_i at timestamp i is defined by SMPL-X [65] parameters $X_i = (R_i^r, t_i^r, \theta_i, \beta_i)$ where $R_i^r \in \mathbb{R}^3$ and $t_i^r \in \mathbb{R}^3$ denotes root joint's global rotation and translation (at pelvis), $\theta_i \in \mathbb{R}^{21\times 3}$ denotes the local joint angles of the kinematic skeleton with 21 joints, and $\beta_i \in \mathbb{R}^{10}$ denotes body shape which remains constant over time ($\beta_i = \beta_j$ for all i, j). Using this notation, we define three egocentric motion tasks:

Egocentric Motion Generation aims to synthesize plausible future motion from a single egocentric image. Formally, this task involves sampling from $p(\boldsymbol{X}_{1:N}|I_1)$, where the front-facing egocentric image I_1 provides a personalized scene context. This problem is more challenging than 3D scene-aware motion generation because, without explicit 3D scene input, the model must infer geometric information from the visible scene and make plausible assumptions about occluded regions. For instance, if the right corner of a soccer field is visible, the model must reason about the likely position of the goal net. Additionally, the model must infer the ongoing action without explicit action labels or textual prompts. For example, if the egocentric image captures raised hands holding a basketball, the most plausible motion is taking a shot at the basket.

Egocentric Motion Forecasting predicts future motion given past egocentric observations, formulated as sampling from $p(\boldsymbol{X}_{n+1:N}|\boldsymbol{I}_{1:n}, \boldsymbol{T}_{1:n})$ where $n < N$. This process implicitly involves reconstructing past motion, represented as $p(\boldsymbol{X}_{1:n}|\boldsymbol{I}_{1:n}, \boldsymbol{T}_{1:n})$, followed by traditional pose forecasting. Thus, it can also be formulated as $p(\boldsymbol{X}_{n+1:N}|\boldsymbol{I}_{1:n}, \boldsymbol{X}_{1:n})$. [1] In this paper, we adopt the former definition. This task is easier than generation as the model benefits from the past observations, providing additional context and constraints on plausible future motion.

Egocentric Motion Reconstruction aims to recover motion from an egocentric video and the corresponding ego camera trajectory. It is formulated as sampling from $p(\boldsymbol{X}_{1:N}|\boldsymbol{I}_{1:N}, \boldsymbol{T}_{1:N})$. Compared to generation and forecasting, reconstruction is relatively easier due to its strong frame-aligned conditioning. Prior works [9, 30, 42, 48, 92] often refer to this as a 'generation' task because of the

[1] It is reasonable to assume that the egocentric device is placed at a fixed location with respect to the user's head and device location T_i can be derived from pose X_i using a known transform.

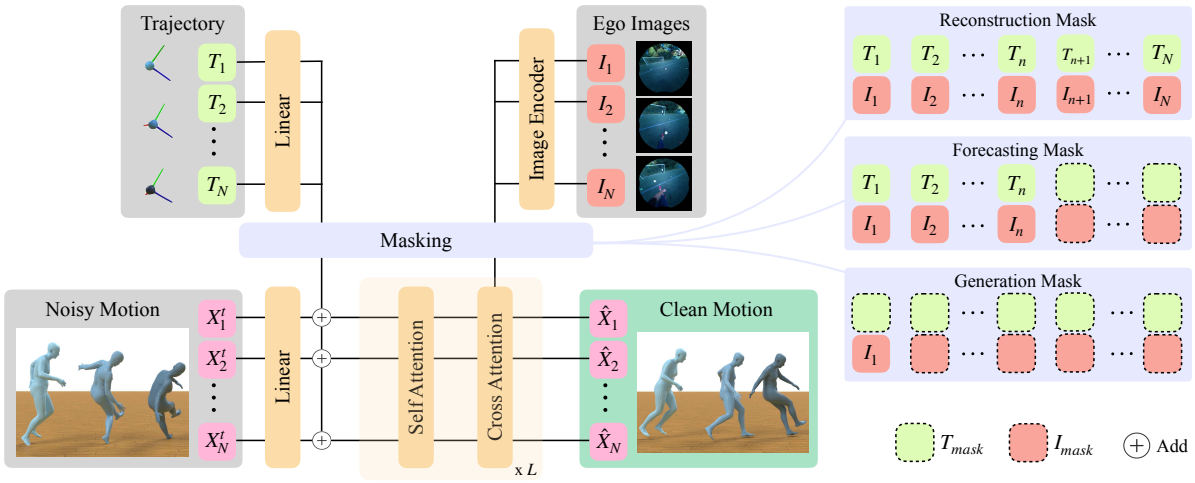

Figure 2. Overview of a denoising step in UniEgoMotion. The input noisy motion $\boldsymbol{X}^t_{1:N}$ is denoised using a transformer decoder network conditioned on the ego-device trajectory $\boldsymbol{T}_{1:N}$ and egocentric images $\boldsymbol{I}_{1:N}$. A robust image encoder is used to extract fine-grained scene context from the images. During training, conditioning inputs are randomly replaced with learnable mask tokens to simulate three tasks: egocentric reconstruction, forecasting, and generation. During inference, the learned mask tokens are used in place of any missing conditioning input, allowing a single model to perform all three tasks consistently.

body's partial visibility in egocentric views. However, we adopt the term 'reconstruction' based on its frame-wise aligned conditioning to distinguish it from egocentric generation and forecasting.

3.2. Diffusion Motion Modeling

Most diffusion-based motion models follow [38, 77] and train a conditional diffusion model [37]

$$\hat{\boldsymbol{X}} = \mathcal{M}\left(\boldsymbol{X}^t,\ t,\ \boldsymbol{C};\ \Theta\right)$$

where $\boldsymbol{X}^t$ is a noised version of the clean motion $\boldsymbol{X}$, $\hat{\boldsymbol{X}}$ is the predicted clean motion, t is the diffusion timestep, $\boldsymbol{C}$ represents optional conditioning inputs, and Θ are the learnable model parameters. For clarity, motion frame indices are omitted i.e. $\boldsymbol{X} = \boldsymbol{X}_{1:N}$. Following [38], $\boldsymbol{X}$ is sampled using a forward Gaussian diffusion process

$$q_t(\boldsymbol{X}^t|\boldsymbol{X}) = \mathcal{N}(\boldsymbol{X}^t; \sqrt{\bar{\alpha}_t}\boldsymbol{X}, (1-\bar{\alpha}_t)\boldsymbol{I})$$

where $\bar{\alpha}_t$ defines a monotonically increasing noise schedule. The model $\mathcal{M}(\cdot|\cdot\ ;\Theta)$ learns the reverse diffusion process by minimizing the following denoising loss.

$$\mathcal{L} = \mathbb{E}_{t\in[1,t_{max}],\boldsymbol{X}^t\sim q_t(\cdot|\boldsymbol{X})}\left[\left\|\boldsymbol{X} - \mathcal{M}\left(\boldsymbol{X}^t,\ t,\ \boldsymbol{C}\right)\right\|_2^2\right]$$

The conditioning input $\boldsymbol{C}$ can include text prompts, action labels, 3D scene or object geometry, or other relevant features, depending on the task. During inference, sampling starts from random Gaussian noise $\boldsymbol{X}^{t_{max}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, and iteratively denoises through

$$\boldsymbol{X}^{t-1} = \mathcal{M}(\boldsymbol{X}^t, t, \boldsymbol{C}) + \epsilon_t$$

until $t = 1$, ultimately generating a clean sequence $\boldsymbol{X}^0$.

3.3. UniEgoMotion

Instead of training separate models for egocentric reconstruction, forecasting, and generation, we use a unified approach where the conditioning $\boldsymbol{C}$ is adapted to the specific task, as described in §3.1. Recent egocentric motion reconstruction methods [9, 48, 92] fit into this framework by setting $\boldsymbol{C} = \boldsymbol{T}_{1:N}$. However, we leverage the fact that conditional diffusion models trained with classifier-free guidance [37] support sampling from both conditional and unconditional distribution. While [37] used unconditional generation to balance sample quality and diversity, we employ it specifically for egocentric motion generation. During training, we randomly set $\boldsymbol{C} = \{\boldsymbol{T}_{1:N}, \boldsymbol{I}_{1:N}\}$ to simulate the reconstruction task and $\boldsymbol{C} = \{I_1\}$ to simulate the generation task, covering the both extremes.

Forecasting can be trained by setting $\boldsymbol{C} = \{\boldsymbol{T}_{1:n}, \boldsymbol{I}_{1:n}\}$ where $n < N$. For the reconstruction-then-forecasting approach, diffusion repainting [55] can also be applied during inference. In particular, given inputs $\{\boldsymbol{T}_{1:n}, \boldsymbol{I}_{1:n}\}$, we first reconstruct the observed motion (using egocentric reconstruction) as $\boldsymbol{X}_{1:n} \sim \mathcal{M}(\cdot|\boldsymbol{T}_{1:n}, \boldsymbol{I}_{1:n})$. We then condition on $\boldsymbol{C} = \{\boldsymbol{T}_{1:n}, \boldsymbol{I}_{1:n}\}$ and sample the full motion sequence $\hat{\boldsymbol{X}}_{1:N}$, enforcing consistency by overwriting the known frames at each step as the following.

$$\hat{\boldsymbol{X}}_{1:N} \leftarrow \text{concat}(\boldsymbol{X}_{1:n}, \hat{\boldsymbol{X}}_{n+1:N})$$

3.4. Architecture

We implement UniEgoMotion using a transformer-based [80] architecture to denoise noisy motion input $\boldsymbol{X}^t_{1:N}$. Each motion input X_i is projected into a latent vec-

tor via a linear layer, $f_X(X_i)$, and then processed by multiple transformer decoder layers. For conditioning, we use full conditioning $\boldsymbol{C} = \{\boldsymbol{T}_{1:N}, \boldsymbol{I}_{1:N}\}$ during reconstruction. In generation mode, where $\boldsymbol{C} = \{I_1\}$, we use a learnable mask inputs to create full conditioning. In particular, we set $T_i = T^{\text{mask}}$ for all $i \in \{1, \cdots, N\}$ and $I_i = I^{\text{mask}}$ for all $i \in \{2, \cdots, N\}$. Forecasting conditioning is processed in a similar manner. Each T_i and I_i is projected into a latent vector using $f_T(T_i)$ and $f_I(I_i)$, where f_T is a linear layer and f_I is a ViT-based image encoder. $f_T(T_i)$ is added to $f_X(X_i)$ before passing through the transformer, while $f_I(I_i)$ is incorporated via a cross-attention mechanism.

Unlike prior works [48, 92] that discard semantic information from ego images, we integrate fine-grained scene-aware features through f_I. We show that the choice of f_I significantly impacts the accuracy and fidelity of motion prediction. Training f_I from scratch is suboptimal, as extracting scene context from images is a challenging problem in itself. To address this, we leverage a pretrained DINOv2 [62] to initialize f_I, training only the projector network. Our results show that its fine-grained features from [62] yield significant improvements over other strong image encoders [66, 70].

3.5. Motion Representation

Although SMPL-X parameters $X_i = (R_i^r, t_i^r, \theta_i, \beta_i)$ are sufficient to represent 3D body motion, they are not ideal for learning [26, 92]. The global root trajectory (R_i^r, t_i^r), defined at the pelvis, fails to exploit motion redundancies, requiring the model to learn all directions explicitly. Moreover, a mismatch exists between the egocentric conditioning inputs (T_i, I_i) and the pelvis-centric SMPL-X parameters, complicating motion reasoning. Using local joint angles further forces to learn complex forward kinematics, often leading to artifacts like foot-floor penetration and sliding.

To address these issues, we adopt a head-centric representation. We transform X_i into $(M_i^h, \boldsymbol{M}_i^j)$ using forward kinematics where $M_i^h \in \mathbb{R}^{4\times4}$ is the global SE(3) transform of the head joint, and $\boldsymbol{M}_i^j \in \mathbb{R}^{21\times4\times4}$ are those of other joints. This removes joint dependencies in the kinematic chain. Next, we derive a canonical reference frame ${}_cM_i$ per frame by eliminating pitch, roll, and height relative to the floor, ensuring that ${}_cM_i$ captures the head's global trajectory projected onto the floor. Motion $(M_i^h, \boldsymbol{M}_i^j)$ is then expressed as $({}_cM_i, {}_cM_i \odot M_i^h, {}_cM_i \odot \boldsymbol{M}_i^j)$ where the latter terms encode local canonicalized pose information. For trajectory invariance, we represent ${}_cM_i$ as its residual relative to the previous frame. For more details, please refer to the supplementary material. While [92] adopts a similar canonicalization scheme, it preserves the kinematic chain, resulting in severe foot-floor penetration and floating artifacts. Our experiments validate the effectiveness of our motion representation against [92].

3.6. EE4D-Motion Dataset

To train UniEgoMotion, we process EgoExo4D dataset [23] and develop EE4D-Motion dataset that provides synchronized egocentric videos and pseudo-ground-truth 3D motion data. Since existing datasets either lack paired egocentric videos or motion annotations, we develop a processing pipeline to fit SMPL-X [5] to EgoExo4D sequences. Our approach refines initial pose estimates through multi-view optimization, sequence-level smoothing, and quality filtering, producing 110+ hours of 3D-accurate motion data for real-world activities. Please refer to suppl. material for more details on EE4D-Motion.

4. Experiments

We follow the official split of the EgoExo4D dataset [23] to partition the EE4D-Motion dataset into training and validation sets based on capture takes. UniEgoMotion and other baselines are trained on 8-second video clips sampled at 10fps ($N = 80$), with clips extracted every 2 seconds, resulting in a total of 143K training samples. For evaluation, we sample similar video clips every 20 seconds, yielding 4400 validation samples. For forecasting, we predict 6 seconds into the future after observing first 2 seconds ($n = 20$) of egocentric inputs. We train a single UniEgoMotion model and evaluate it across all three tasks.

4.1. Metrics

We employ several metrics to evaluate motions in both 3D and semantic space. **MPJPE** calculates the mean per-joint positional error (in meters) of 22 body joints. **MPJPE-PA** applies Procrustes analysis to align ground truth and predicted motions per frame before computing MPJPE, measuring the accuracy of local pose predictions. **MPJPE-H** calculates the mean per-joint positional error (in meters) of hand joints. **Head Rotation Error** and **Head Translation Error** measure the rotation error and translation error (in meters) of the head joint, respectively, capturing the model's ability to adhere to head-aligned conditioning inputs for the reconstruction task. Rotation error is calculated as the frobenius norm of the difference rotation matrix [48]. **Foot Sliding** [34] quantifies the extent of foot sliding when the foot is close to the ground. **Foot Contact** computes the average separation (in meters) between foot and ground. It quantifies both floating and floor penetration. **Semantic Similarity** evaluates the similarity between generated and ground-truth motions similar to CLIP-score [36]. Specifically, we leverage the motion encoder from TMR [27] to embed motion into a latent space and compute the cosine distance between embeddings. **FID** measures the distributional discrepancy between generated and ground-truth motions in the latent space, akin to vanilla FID for images.

For generation and forecasting, we compute the MPJPE

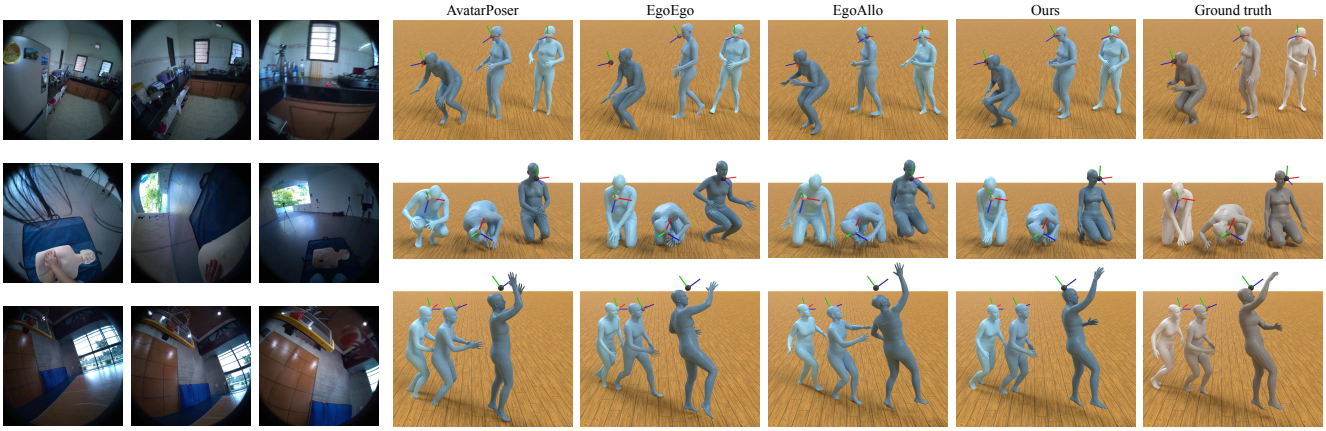

Figure 3. Qualitative comparison of Egocentric Reconstruction. The input egocentric images are shown on the left, with the corresponding ego-device trajectory visualized alongside the predictions. Baseline methods exhibit floating motion, floor penetration, and inaccurate joint localization, whereas UniEgoMotion generates reconstructions that closely align with the ground truth.

Table 1. **Egocentric Motion Reconstruction**: Comparison of the reconstruction capabilities of UniEgoMotion with prior works (top). Ablation on UniEgoMotion's model design for the reconstruction task (bottom). Note that the vanilla UniEgoMotion model uses transformer decoder architecture, head-centric motion representation, and DINOv2 visual encoder.

Method	Head Rot. Err.	Head Trans. Err.	MPJPE	MPJPE-PA	MPJPE-H	Foot Slide	Foot Contact	Semantic Sim.(↑)	FID
AvatarPoser [42]	-	-	0.116	0.068	0.240	7.85	0.042	0.872	0.082
EgoEgo [48]	-	-	0.130	0.075	0.272	3.90	0.033	0.858	0.068
EgoAllo [92]	-	-	0.163	0.071	0.273	4.10	0.056	0.885	0.043
UniEgoMotion	**0.260**	0.058	**0.100**	**0.053**	**0.180**	**3.62**	0.027	**0.918**	0.027
Transformer Encoder	0.280	0.076	0.115	0.056	0.189	4.48	0.029	0.912	**0.017**
1D U-Net	0.338	0.109	0.145	0.061	0.224	5.86	0.032	0.900	0.019
Global Motion Repre.	0.275	**0.051**	0.101	0.057	0.192	3.73	**0.025**	0.912	0.024
Pelvis-centric Repre.	0.398	0.138	0.166	0.054	0.241	3.66	0.028	0.909	0.030
CLIP encoder	0.269	0.062	0.107	0.056	0.191	4.03	0.032	0.911	0.021
EgoVideo encoder	0.332	0.101	0.132	0.060	0.211	4.73	0.032	0.897	0.041

and MPJPE-PA metrics only for the first 2 seconds of prediction, as beyond that, generated motions may remain valid despite exhibiting large joint errors. FID and Semantic Similarity compares motions in a semantic latent space and offers a more meaningful evaluation of the motion quality and scene-relevance respectively. Foot Slide and Foot Contact capture physical realism of the predicted motion.

4.2. Baselines

We compare our egocentric motion reconstruction with task-specific prior works: EgoEgo [48], EgoAllo [92], and AvatarPoser [42]. To ensure a fair evaluation, we retrain each method on the EE4D-Motion dataset using publicly available code. None of these baselines use semantic information from egocentric images for prediction. We compare these baselines on the reconstruction task and further ablate their design choices within our UniEgoMotion framework in a consistent manner. Since there are no direct baselines for egocentric motion forecasting and generation, we construct strong baselines based on state-of-the-art motion modeling practices. LSTM-forecasting is a task-specific model that sequentially processes forecasting inputs $\{f_I(I_i), f_T(T_i)\}_{i=1:n}$ using an LSTM and outputs the motion $\boldsymbol{X}_{n+1:N}$. Similarly, LSTM-generation processes $f_I(I_1)$ to generate $\boldsymbol{X}_{1:N}$. We also train a two-stage model to replicate the two-stage approach used in prior works on wide-scene image-based forecasting [8] and generation [84], but with diffusion modeling. In particular, a UniEgoMotion-trajectory model first predicts the head trajectory, followed by the standard UniEgoMotion model, which takes the predicted head trajectory as additional input. We further ablate our model by replacing the transformer decoder with a transformer encoder and a specialized 1D-UNet-based motion model [44]. We also evaluate our motion representation against simple global representation [48] and traditional pelvis-centric representation [26, 42, 92]. Finally, we examine the impact of using fine-grained features of DINOv2 versus text-optimized semantic features trained on general natural images (CLIP [70]) and in-domain egocentric images (EgoVideo [66]).

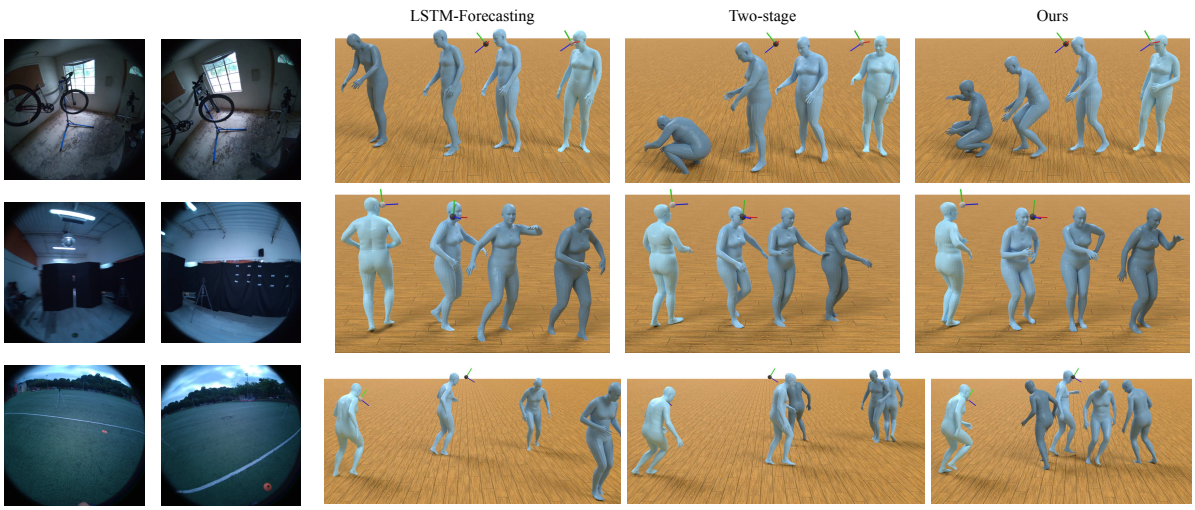

Figure 4. Qualitative comparison of Egocentric Forecasting for predicting future motion using the first 2 seconds of egocentric video and trajectory input. The LSTM baseline predicts an average future motion and suffers from foot sliding, while the Two-stage baseline produces damped motion. In contrast, our model successfully predicts complex motions, such as squatting down to repair a bike tire (top), performing a salsa dance (middle), and executing a dribbling drill around a dome cone (bottom).

Table 2. The baselines and ablations are evaluated on egocentric motion forecasting (left) and generation (right). The metrics reported include **J**: MPJPE, **J-PA**: MPJPE-PA, **J-H**: MPJPE-H, **FS**: Foot Slide, **FC**: Foot Contact, and **SS**: Semantic Similarity. MPJPE metrics are computed over the first two seconds of future predictions (0-2s for generation and 2-4s for forecasting). *Two-stage baseline replicates the trajectory-to-motion prediction framework used in prior works on image-based motion forecasting [8] and motion generation [84].

Method	Egocentric Motion Forecasting							Egocentric Motion Generation						
	J (2-4s)	J-PA (2-4s)	J-H (2-4s)	FS	FC	SS (↑)	FID	J (0-2s)	J-PA (0-2s)	J-H (0-2s)	FS	FC	SS (↑)	FID
LSTM	0.238	**0.066**	0.330	7.23	0.031	0.849	0.058	**0.216**	**0.067**	**0.308**	6.83	0.028	0.809	0.090
Two-stage* [8, 84]	0.253	0.072	0.361	3.55	**0.026**	**0.850**	**0.038**	0.222	0.072	0.323	4.35	0.026	**0.822**	0.037
UniEgoMotion	**0.206**	0.071	**0.308**	**2.60**	**0.026**	0.849	0.047	0.226	0.070	0.321	**2.89**	**0.025**	0.817	0.043
Transformer Encoder	0.213	0.073	0.315	3.13	0.027	0.846	0.041	0.231	0.072	0.330	3.40	0.026	0.814	**0.034**
1D U-Net	0.239	0.075	0.346	4.11	0.027	0.840	0.056	0.259	0.079	0.360	4.06	0.029	0.802	0.035
Global Motion Repre.	0.293	0.076	0.405	3.16	0.027	0.841	0.046	0.228	0.072	0.328	3.65	0.025	0.821	0.035
Pelvis-centric Repre.	0.245	0.074	0.354	3.56	0.030	0.838	0.042	0.232	0.071	0.327	3.94	0.029	0.814	0.039
CLIP encoder	0.214	0.073	0.315	3.75	0.030	0.844	0.043	0.238	0.073	0.333	3.57	0.028	0.816	0.037
EgoVideo encoder	0.228	0.077	0.331	3.65	0.030	0.835	0.060	0.236	0.074	0.322	3.50	0.030	0.812	0.059

4.3. Results & Discussion

Egocentric Motion Reconstruction: Tab. 1 shows the results and ablation study for the egocentric reconstruction task, and Fig. 3 shows the qualitative comparison of the reconstruction baselines. UniEgoMotion's egocentric reconstruction capabilities outperform specialized baselines – AvatarPoser[42], EgoEgo [48], and EgoAllo [92], in both reconstruction and semantic metrics. EgoAllo's motion representation with a kinematic chain and local joint angles, often struggles to reconstruct motion accurately grounded on the floor. As a result, it shows frequent foot sliding and floating motion, leading to its high Foot Contact error in Tab. 1. We further verify the benefits of egocentric motion representation by training UniEgoMotion with a pelvis-centric representation. We also show that transformer cross-attention is better suited for our flexible conditioning setting [72] compared to encoder-based architectures [30, 48] and specialized 1D-Unet [44]. Thanks to the explicit use of fine-grained image features, UniEgoMotion captures visible semantic cues more effectively and achieves the lowest MPJPE-* errors along with the highest motion quality and semantic similarity to the ground-truth motion.

Although UniEgoMotion with CLIP [70] image encoder shows strong performance [30], the fine-grained features of DINOv2 [62] are more suitable for extracting task-relevant scene context, which is often not the central focus of the image. Surprisingly, the in-domain contrastive video features of EgoVideo [66] perform slightly worse than CLIP, suggesting that generalized scene context is more important than ego-action centric image features.

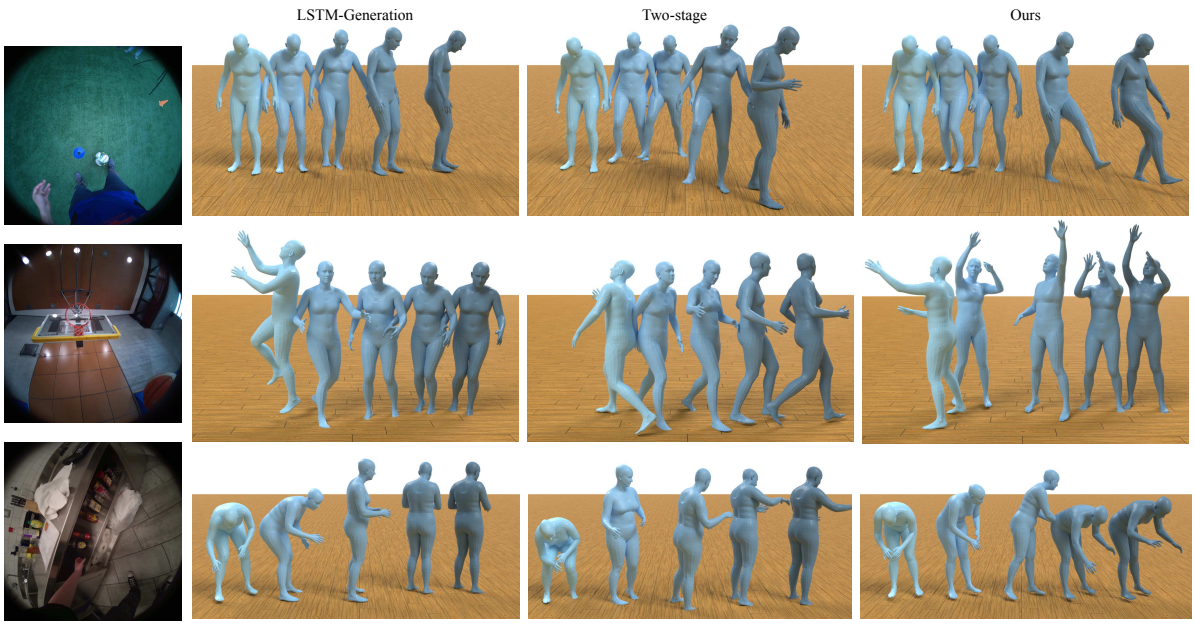

Figure 5. Qualitative comparison of Egocentric Motion Generation from a single egocentric image input. Compared to the LSTM and Two-stage baseline, our model leverages the fine-grained image features for more accurate motion generation, demonstrating soccer juggling (top), a basketball shooting drill (middle), and interaction with the lower cabinet on the *left* side of the person.

Egocentric Motion Forecasting & Generation: Tab. 2 shows the quantitative results for the forecasting and generation tasks. Qualitative results are shown in Fig. 4 & 5 respectively. Since the LSTM-based forecasting/generation baselines deterministically output the average of plausible future motions, they perform well on comparison metrics such as MPJPE and Semantic Similarity. However, the generated average motion exhibits significant foot-sliding and floor-penetration, resulting in lower motion quality, particularly affecting the Motion FID score. The extensive two-stage baseline performs slightly worse on some metrics due to a mismatch between the distributions of the generated and groundtruth trajectories. However, it achieves strong motion quality metrics, benefiting from the vanilla UniEgoMotion model used in the second stage. In comparison, UniEgoMotion enables one-shot high-quality forecasting and generation within a unified model, delivering strong overall performance across all metrics.

We also evaluate various design aspects of UniEgoMotion in the forecasting and generation tasks, and the results are consistent with those observed in the reconstruction task. Motion representation plays an important role in generating high-quality motion, as reflected in metrics such as Foot Slide. The choice of image encoder has a slightly smaller impact than in reconstruction, as high-level scene context may be sufficient for generating/forecasting scene-relevant motion. However, a fine-grained image encoder still provides meaningful benefits in all metrics and noticeably improves motion realism, as observed in FC and FS metrics. Interestingly, 1D-UNet [44], which focuses on local motion reasoning, performs noticeably worse in both generation and forecasting. For additional ablation and qualitative comparisons, please refer to the suppl. material.

5. Conclusion & Future Work

We present UniEgoMotion, a unified framework for egocentric motion reconstruction, forecasting, and generation. Unlike previous methods, it extracts scene context from egocentric images, enabling scene-aware motion synthesis without explicit 3D scene. By integrating fine-grained visual features, our approach improves motion accuracy and realism. We also introduce EE4D-Motion, a large-scale dataset from EgoExo4D, offering time-synchronized egocentric video and pseudo-ground-truth 3D motion data. UniEgoMotion outperforms state-of-the-art methods in egocentric motion reconstruction while enabling novel egocentric forecasting and generation capabilities. Our experiments emphasize the need for scene-aware motion reasoning. We show that instead of specialized motion architectures, a well-structured simple model with a strong context encoder and egocentric motion representation achieves superior results. Looking ahead, we plan to explore egocentric scene-motion interactions and leverage multimodal annotations for applications like in-context motion generation from text prompts. We believe UniEgoMotion provides a strong benchmark to drive future research in egocentric motion analysis and generation.

Acknowledgment This research was conducted at Stanford University with support from Panasonic Holdings Corporation and partial funding from NIH grant R01AG089169.

References

[1] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. A spatio-temporal transformer for 3d human motion prediction. In *2021 International Conference on 3D Vision (3DV)*, pages 565–574. IEEE, 2021. 3

[2] Joao Pedro Araújo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Jiajun Wu, Deepak Gopinath, Alexander William Clegg, and Karen Liu. Circle: Capture in rich contextual environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21211–21221, 2023. 2

[3] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Sinc: Spatial composition of 3d human motions for simultaneous action generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9984–9995, 2023. 2

[4] German Barquero, Sergio Escalera, and Cristina Palmero. Belfusion: Latent diffusion for behavior-driven human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2317–2327, 2023. 3

[5] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 561–578. Springer, 2016. 5, 17

[6] Yujun Cai, Lin Huang, Yiwei Wang, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Xu Yang, Yiheng Zhu, Xiaohui Shen, et al. Learning progressive joint propagation for human motion prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 226–242. Springer, 2020. 3

[7] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36:11454–11468, 2023. 17

[8] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 387–404. Springer, 2020. 2, 3, 6, 7, 14

[9] Angela Castillo, Maria Escobar, Guillaume Jeanneret, 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*, pages 4221–4231, 2023. 3, 4

[10] Zhi Cen, Huaijin Pi, Sida Peng, Zehong Shen, Minghui Yang, Zhu Shuai, Hujun Bao, and Xiaowei Zhou. Generating human motion in 3d scenes from text descriptions. In *CVPR*, 2024. 2

[11] Hsu-kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. Action-agnostic human pose forecasting. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 1423–1432. IEEE, 2019. 3

[12] Enric Corona, Albert Pumarola, Guillem Alenya, and Francesc Moreno-Noguer. Context-aware human motion prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6992–7001, 2020. 2, 3

[13] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9760–9770, 2023. 2

[14] Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11467–11476, 2021. 3

[15] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. 3, 16

[16] Hong et al. Egolm: Multi-modal language model of egocentric motions. *CVPR*, 2025. 3, 16

[17] Ma et al. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *ECCV*, 2024. 3, 16

[18] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE international conference on computer vision*, pages 4346–4354, 2015. 3

[19] Stuart Geman and Donald E. McClure. Statistical methods for tomographic image reconstruction. 1987. 17

[20] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1396–1406, 2021. 2

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 3

[22] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022. 16

[23] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 2, 5, 16, 17

[24] Qiao Gu, Zhaoyang Lv, Duncan Frost, Simon Green, Julian Straub, and Chris Sweeney. Egolifter: Open-world 3d segmentation for egocentric perception. In *European Conference on Computer Vision*, pages 382–400. Springer, 2024. 2

[25] 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. 2

[26] 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*, pages 5152–5161, 2022. 2, 5, 6, 15

[27] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*, pages 580–597. Springer, 2022. 2, 5

[28] Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Multi-person extreme motion prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13053–13064, 2022. 3

[29] Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Back to mlp: A simple baseline for human motion prediction. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 4809–4819, 2023. 3

[30] Vladimir Guzov, Yifeng Jiang, Fangzhou Hong, Gerard Pons-Moll, Richard Newcombe, C Karen Liu, Yuting Ye, and Lingni Ma. Hmd2: Environment-aware motion generation from single egocentric head-mounted device. *arXiv preprint arXiv:2409.13426*, 2024. 3, 7, 16

[31] Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)*, 39(4):60–1, 2020. 2

[32] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2282–2292, 2019. 2

[33] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael Black. Stochastic scene-aware motion prediction. In *Proceedings of the International Conference on Computer Vision 2021*, 2021. 2

[34] Chengan He, Jun Saito, James Zachary, Holly Rushmeier, and Yi Zhou. Nemf: Neural motion fields for kinematic animation. *Advances in Neural Information Processing Systems*, 35:4244–4256, 2022. 2, 5

[35] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020. 2

[36] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 5

[37] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4

[38] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3, 4

[39] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (ToG)*, 35(4):1–11, 2016. 2

[40] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017. 2

[41] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 5308–5317, 2016. 3

[42] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *European conference on computer vision*, pages 443–460. Springer, 2022. 2, 3, 6, 7, 14

[43] Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1737–1747, 2024. 2

[44] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2151–2162, 2023. 6, 7, 8

[45] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013. 3

[46] Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas Guibas. Nifty: Neural object interaction fields for guided human motion synthesis, 2023. 2

[47] Andreas M Lehrmann, Peter V Gehler, and Sebastian Nowozin. Efficient nonlinear markov models for human motion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1314–1321, 2014. 3

[48] Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17142–17151, 2023. 2, 3, 4, 5, 6, 7, 14, 16

[49] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. In *European Conference on Computer Vision*, pages 54–72. Springer, 2024. 2

[50] Peizhuo Li, Kfir Aberman, Zihan Zhang, Rana Hanocka, and Olga Sorkine-Hornung. Ganimator: Neural motion synthesis from a single sequence. *ACM Transactions on Graphics (TOG)*, 41(4):1–12, 2022. 2

[51] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European conference on computer vision*, pages 280–296. Springer, 2022. 17

[52] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36: 25268–25280, 2023. 16, 17

[53] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel Van De Panne. Character controllers using motion vaes. *ACM Transactions on Graphics (TOG)*, 39(4):40–1, 2020. 2

[54] Yuxuan Liu, Jianxin Yang, Xiao Gu, Yao Guo, and Guang-Zhong Yang. Egohmr: Egocentric human mesh recovery via hierarchical latent diffusion model. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9807–9813. IEEE, 2023. 3

[55] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022. 2, 4

[56] Zhengyi Luo, Ryo Hachiuma, Ye Yuan, and Kris Kitani. Dynamics-regulated kinematic policy for egocentric pose estimation. *Advances in Neural Information Processing Systems*, 34:25019–25032, 2021. 3

[57] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 16

[58] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 474–489. Springer, 2020. 3

[59] Wei Mao, Miaomiao Liu, Richard Hartley, and Mathieu Salzmann. Contact-aware human motion forecasting. 2022. 2, 3

[60] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2891–2900, 2017. 3

[61] Angel Martínez-González, Michael Villamizar, and Jean-Marc Odobez. Pose transformers (potr): Human motion prediction with non-autoregressive transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2276–2284, 2021. 3

[62] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 2, 5, 7

[63] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20133–20143, 2023. 2

[64] Jinman Park, Kimathi Kaai, Saad Hossain, Norikatsu Sumi, Sirisha Rambhatla, and Paul Fieguth. 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*, pages 1837–1849, 2023. 3

[65] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 3

[66] Baoqi Pei, Guo Chen, Jilan Xu, Yuping He, Yicheng Liu, Kanghua Pan, Yifei Huang, Yali Wang, Tong Lu, Limin Wang, and Yu Qiao. Egovideo: Exploring egocentric foundation model and downstream adaptation. *arXiv preprint arXiv:2406.18070*, 2024. 5, 6, 7

[67] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10985–10995, 2021. 2

[68] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497. Springer, 2022. 2

[69] Sigal Raab, Inbal Leibovitch, Peizhuo Li, Kfir Aberman, Olga Sorkine-Hornung, and Daniel Cohen-Or. Modi: Unconditional motion synthesis from diverse data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13873–13883, 2023. 2

[70] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 5, 6, 7

[71] Helge Rhodin, Christian Richardt, Dan Casas, Eldar Insafutdinov, Mohammad Shafiei, Hans-Peter Seidel, Bernt Schiele, and Christian Theobalt. Egocap: egocentric marker-less motion capture with two fisheye cameras. *ACM Transactions on Graphics (TOG)*, 35(6):1–11, 2016. 3

[72] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 7

[73] Luca Scofano, Alessio Sampieri, Elisabeth Schiele, Edoardo De Matteis, Laura Leal-Taix'e, and Fabio Galasso. Staged contact-aware global human motion forecasting. *ArXiv*, abs/2309.08947, 2023. 2, 3

[74] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Transactions on Graphics*, 38(6):178, 2019. 2

[75] Sebastian Starke, Yiwei Zhao, Taku Komura, and Kazi A Zaman. Local motion phases for learning multi-contact character movements. *ACM Trans. Graph.*, 39(4):54, 2020. 2

[76] Graham W Taylor, Geoffrey E Hinton, and Sam Roweis. Modeling human motion using binary latent variables. *Advances in neural information processing systems*, 19, 2006. 3

[77] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 2, 4, 16

[78] Denis Tome, Thiemo Alldieck, Patrick Peluse, Gerard Pons-Moll, Lourdes Agapito, Hernan Badino, and Fernando De la Torre. Selfpose: 3d egocentric pose estimation from a headset mounted camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):6794–6806, 2020. 3

[79] Vadim Tschernezki, Ahmad Darkhalil, Zhifan Zhu, David Fouhey, Iro Laina, Diane Larlus, Dima Damen, and Andrea Vedaldi. Epic fields: Marrying 3d geometry and video understanding. *Advances in Neural Information Processing Systems*, 36:26485–26500, 2023. 2

[80] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 4, 16

[81] Borui Wang, Ehsan Adeli, Hsu-kuang Chiu, De-An Huang, and Juan Carlos Niebles. Imitation learning for human pose prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7124–7133, 2019. 3

[82] Jack Wang, Aaron Hertzmann, and David J Fleet. Gaussian process dynamical models. *Advances in neural information processing systems*, 18, 2005. 3

[83] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes, 2020. 2

[84] Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. Scene-aware generative network for human motion synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12206–12215, 2021. 2, 3, 6, 7, 14

[85] Jian Wang, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, Diogo Luvizon, and Christian Theobalt. 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*, pages 13157–13166, 2022. 3

[86] Jian Wang, Rishabh Dabral, Diogo Luvizon, Zhe Cao, Lingjie Liu, Thabo Beeler, and Christian Theobalt. Ego4o: Egocentric human motion capture and understanding from multi-modal input. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22668–22679, 2025. 3, 16

[87] Zhenyi Wang, Ping Yu, Yang Zhao, Ruiyi Zhang, Yufan Zhou, Junsong Yuan, and Changyou Chen. Learning diverse stochastic human-action generators by learning smooth latent transitions. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12281–12288, 2020. 2

[88] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. *Advances in Neural Information Processing Systems*, 35:14959–14971, 2022. 2

[89] Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as you say interact as you can: Language-guided human motion generation with scene affordance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 433–444, 2024. 2

[90] Chaoyue Xing, Wei Mao, and Miaomiao Liu. Scene-aware human motion forecasting via mutual distance prediction. In *European Conference on Computer Vision*, pages 128–144. Springer, 2024. 2, 3

[91] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems*, 2022. 17

[92] Brent Yi, Vickie Ye, Maya Zheng, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. Estimating body and hand motion in an ego-sensed world. *arXiv preprint arXiv:2410.03665*, 2024. 2, 3, 4, 5, 6, 7, 14, 15, 16

[93] Hongwei Yi, Justus Thies, Michael J Black, Xue Bin Peng, and Davis Rempe. Generating human interaction motions in scenes with text control. In *European Conference on Computer Vision*, pages 246–263. Springer, 2024. 2

[94] Ye Yuan and Kris Kitani. 3d ego-pose estimation via imitation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 735–750, 2018. 3

[95] Ye Yuan and Kris Kitani. Ego-pose estimation and forecasting as real-time pd control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10082–10092, 2019. 2, 3, 16

[96] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14730–14740, 2023. 2

[97] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE transactions on pattern analysis and machine intelligence*, 46(6):4115–4128, 2024. 2

[98] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J. Black, and Siyu Tang. PLACE: Proximity learning of articulation and contact in 3D environments. In *International Conference on 3D Vision (3DV)*, 2020. 2

[99] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *European conference on computer vision*, pages 180–200. Springer, 2022. 16

[100] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. In *European Conference on Computer Vision*, pages 518–535. Springer, 2022. 2

[101] Yan Zhang and Siyu Tang. The wanderings of odysseus in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20481–20491, 2022. 2

[102] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, C Karen Liu, and Leonidas J Guibas. Gimo: Gaze-informed human motion prediction in context. In *European Conference on Computer Vision*, pages 676–694. Springer, 2022. 2, 3

UniEgoMotion: A Unified Model for Egocentric Motion Reconstruction, Forecasting, and Generation

Supplementary Material

A. Qualitative Comparison

See Fig. 6 for a qualitative visualization of egocentric motion reconstruction, with vertex errors color-coded. Please refer to the supplementary video to view UniEgoMotion's results on egocentric motion reconstruction, forecasting, and generation, as well as comparisons with baselines.

B. Baselines

Egocentric Motion Reconstruction

We compare the egocentric motion reconstruction capabilities of UniEgoMotion with task-specific prior works: EgoEgo [48], EgoAllo [92], and AvatarPoser [42]. To ensure a fair evaluation, we retrain each method on the EE4D-Motion dataset using their publicly available code. Following EgoEgo's experimental setup, we exclude hand tracking from AvatarPoser and instead provide a constant input for hand trajectories. Both EgoEgo and AvatarPoser use as inputs the head trajectories derived from motion annotations rather than from the Aria device's SLAM system, resulting in perfect head tracking by design. Similarly, EgoAllo also uses a fixed transformation between head and device trajectory. Therefore, we omit their head tracking metrics from the evaluation. For EgoAllo, we evaluate the output of the motion diffusion model directly, without applying the post-processing optimization step.

Although EgoEgo and EgoAllo also adopt diffusion-based formulation for motion reconstruction, their approach differ from ours in their choice of motion representation and model architecture. For instance, EgoEgo assumes a constant body shape and uses a global motion representation, whereas EgoAllo uses a head-centric representation that explicitly includes the head-to-pelvis transformation and preserves the kinematic chain. More importantly, none of these baselines utilize semantic information from egocentric video for motion prediction. We compare these methods on the reconstruction task and also ablate their design choices separately within our UniEgoMotion framework in a consistent manner.

Egocentric Motion Forecasting & Generation

For egocentric motion forecasting and generation, the most relevant baselines [8, 84] are two-stage models that generate or forecast human motion from third-person RGB images. They first predict the root trajectory (typically pelvis) and then generate the full-body human motion using a global motion representation. To replicate these baselines faithfully, we train a separate UniEgoMotion variant that uses global motion representation and predicts only the root trajectory. This output is then provided as an additional conditioning input to the standard UniEgoMotion model (also with global motion representation) for full-body motion prediction. We also train separate autoregressive LSTM-based baselines with a comparable model capacity for both forecasting and generation tasks. Since these models lack a generative component, their outputs tend to regress toward the mean of all plausible futures. As a result, they show lower error in direct comparison metrics such as MPJPE. However, their 'averaged' prediction suffer from reduced motion diversity and realism, as shown in semantic metrics and qualitative visualization (see supplementary video).

C. Ablation on Conditioning Inputs

We evaluate UniEgoMotion under two ablation settings: without trajectory input and without video input. Additionally, we train two single-modality variants of UniEgoMotion. Egocentric reconstruction results in Tab. 3 shows that both signals are useful for optimal reconstruction performance, thereby validating our use of video input, unlike prior baselines. Interestingly, the separately trained single-modality variants offer no significant advantage over the original UniEgoMotion model when evaluated under the same conditions. Without video input, UniEgoMotion still outperforms baselines on most metrics. However, when the trajectory input is removed, the model is forced to implicitly solve visual odometry problem (a significantly harder task), leading to large errors on absolute metrics (head tracking, MPJPE, MPJPE-H). Despite this, it maintains accuracy in local pose metrics (MPJPE-PA, semantic similarity) and realism (FID), showing its ability to infer plausible motion from video alone.

EgoEgo [48] employed an off-the-shelf monocular visual SLAM on egocentric video and trained an additional module to predict scale and the gravity vector to derive gravity-aligned metric SLAM trajectory. Their results showed that using predicted metric SLAM trajectory leads to only a minor degradation in pose metrics compared to using ground-truth trajectories. In our work, we assume access to inertial SLAM trajectories for both our method and the baselines to decouple motion analysis from trajectory estimation and focus our evaluation on motion tasks.

Forecasting and generation results follow similar trends, with both input modalities contributing to optimal performance. Notably, the model without video input performs

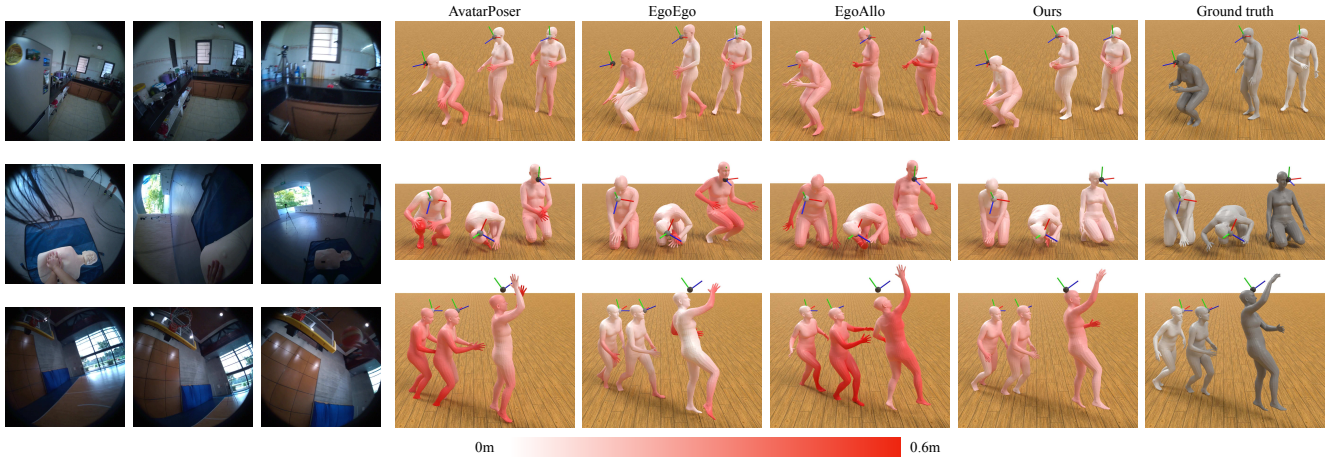

Figure 6. Qualitative comparison of Egocentric Reconstruction, with absolute vertex errors color-coded. The input egocentric images are shown on the left, with the corresponding ego-device trajectory visualized alongside the predictions.

Table 3. **Ablation on Conditioning Inputs**: We evaluate UniEgoMotion in two ablation settings–without video and without trajectory input. Additionally, we train two single-modality variants of UniEgoMotion by conditioning only on trajectory or only on video.

	Egocentric Motion Reconstruction								
Method	Head Rot. Err.	Head Trans. Err.	MPJPE	MPJPE-PA	MPJPE-H	Foot Slide	Foot Contact	Semantic Sim.(↑)	FID
UniEgoMotion	**0.260**	0.058	**0.100**	**0.053**	**0.180**	3.62	0.027	**0.918**	0.027
w/o video	0.278	**0.057**	0.115	0.066	0.234	3.64	0.026	0.878	0.030
w/o trajectory	0.539	0.280	0.290	0.059	0.352	2.95	0.024	0.885	0.033
UniEgoMotion (w/o video)	0.293	0.063	0.119	0.067	0.239	3.49	0.025	0.877	**0.026**
UniEgoMotion (w/o trajectory)	0.535	0.292	0.299	0.060	0.362	**2.70**	**0.023**	0.886	0.035

	Egocentric Motion Forecasting							Egocentric Motion Generation						
Method	J (2-4s)	J-PA (2-4s)	J-H (2-4s)	FS	FC	SS (↑)	FID	J (0-2s)	J-PA (0-2s)	J-H (0-2s)	FS	FC	SS (↑)	FID
UniEgoMotion	**0.206**	0.071	**0.308**	2.60	0.026	**0.849**	**0.047**	**0.226**	**0.070**	**0.321**	2.89	0.025	0.817	**0.043**
w/o video	0.255	0.090	0.378	**2.43**	0.028	0.782	0.058	0.356	0.100	0.449	**2.36**	0.027	0.696	0.065
w/o trajectory	0.322	**0.070**	0.414	2.66	0.025	0.838	**0.047**	**0.226**	**0.070**	**0.321**	2.89	0.025	0.816	**0.043**
UniEgoMotion (w/o video)	0.276	0.095	0.400	2.69	0.028	0.767	0.067	0.379	0.108	0.483	3.03	0.027	0.684	0.044
UniEgoMotion (w/o trajectory)	0.318	**0.070**	0.404	2.50	**0.024**	0.842	0.050	0.228	**0.070**	**0.321**	2.71	**0.024**	**0.820**	0.044

worse, as it lacks scene context necessary for generating or forecasting relevant motion.

D. Motion Representation

Although SMPL-X parameters $X_i = (R_i^r, t_i^r, \theta_i, \beta_i)$ are sufficient to represent 3D body motion, they are not always ideal for learning [26, 92]. The global parameterization of the root trajectory (R_i^r, t_i^r), defined at the pelvis, does not exploit motion invariances, forcing the model to learn all movements in every direction separately. Moreover, a mismatch exists between the conditioning information (T_i, I_i), defined in the egocentric frame, and the SMPL-X parameters X_i, defined in the pelvis-centric frame. This misalignment complicates the reasoning between pelvis-centric motion and egocentric conditioning inputs. Additionally, using local joint angles forces the model to reason complex forward kinematics of the SMPL-X skeleton, often resulting in suboptimal motion with noticeable artifacts such as foot-floor penetration and foot sliding.

To address these issues, we adopt a head-centric motion representation instead of a pelvis-centric one. We transform the SMPL-X parameters $X_i = (R_i^r, t_i^r, \theta_i, \beta_i)$ into $(M_i^h, \boldsymbol{M}_i^j)$ using forward kinematics where $M_i^h \in \mathbb{R}^{4\times4}$ is the global SE(3) transform of the head joint, and $\boldsymbol{M}_i^j \in$

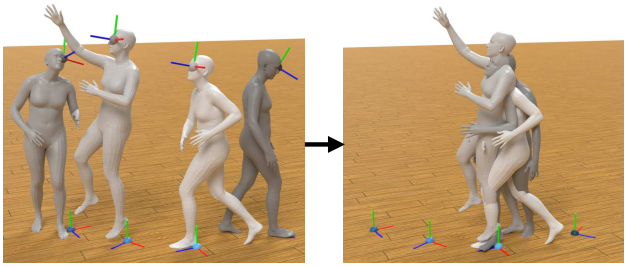

Figure 7. Our egocentric motion representation decomposes motion into two components: (1) the egocentric trajectory projected onto the floor by removing pitch, roll, and height, and (2) the body pose relative to this projected egocentric trajectory.

$\mathbb{R}^{21\times4\times4}$ are the global SE(3) transforms of other joints. This eliminates the dependency of each joint on its parent in the kinematic chain. Next, we derive a canonical reference frame ${}_cM_i$ for each frame by projecting the head transform M_i^h onto the floor. In particular, ${}_cM_i$ represents the *global* 3D transform of the head joint after removing the pitch and roll angle (keeping only yaw) and removing its height t_z relative to the floor (+Z direction). We then express the motion $(M_i^h, \boldsymbol{M}_i^j)$ as $({}_cM_i, {}_cM_i \odot M_i^h, {}_cM_i \odot \boldsymbol{M}_i^j)$, where ${}_cM_i$ captures the head's global trajectory projected onto the floor, and $({}_cM_i \odot M_i^h, {}_cM_i \odot \boldsymbol{M}_i^j)$ encode local canonicalized pose information. To achieve trajectory invariance, we represent ${}_cM_i$ as its residual relative to the previous frame ${}_cM_{i-1}^{-1} \odot {}_cM_i$. Following standard practice, we incorporate additional redundant information, such as joint locations and foot contact labels, into our motion representation.

While our motion representation is similar to the canonicalization in [92], it differs in that [92] retains the kinematic chain and defines local joint rotations relative to parent joints. Since all body joint information in our approach is defined relative to the floor, it naturally facilitates better reasoning about foot-floor contact. We validate the effectiveness of our motion representation through ablation studies and demonstrate that while [92] exhibits significant foot-floor penetration or floating artifacts, UniEgoMotion produces high-quality motion.

E. Why Not Text Conditioning

Many motion generation approaches [16, 77, 86] rely on text-based conditioning, where a clear textual prompt defines the intended motion or action. This explicit guidance simplifies the generation process. In contrast, our work focuses on passive conditioning using sensor data (e.g., video and device trajectory), where motion must be inferred without direct user input. While this introduces greater ambiguity, it also enables broader applicability in real-world scenarios such as continuous gait monitoring or fall prediction, where explicit user inputs are typically unavailable. Nonetheless, we believe that egocentric motion generation and forecasting from text prompts are promising future directions for many assistive applications. Our work, along with datasets like EE4D-Motion (with action narrations from EgoExo4D) and Nymeria [17], offers a promising starting point for such research.

F. Training Details

We train UniEgoMotion for 350 epochs using a batch size of 64 and the AdamW optimizer with a weight decay of 0.01. The learning rate is initialized at 3e-5 and decayed to 3e-6 after epoch 300. The model follows a standard transformer architecture [80], comprising 12 decoder layers with a latent dimension of 768. Training is conducted on 8-second motion sequences (80 steps at 10 fps), enabling long-horizon motion prediction. To improve training efficiency, DINOv2 features are precomputed and cached. End-to-end training takes approximately 2 days on a single NVIDIA L40S GPU. For diffusion, we use cosine noise scheduling with 1000 steps, consistent with prior works [48, 77], though effective motion synthesis has been demonstrated with very few diffusion steps [30, 77]. During training, we alternate between reconstruction and generation tasks with equal 0.5 probability by randomly masking the input sequence.

G. EE4D-Motion Dataset

Training UniEgoMotion requires paired egocentric videos and 3D human motion data within real-world environments. However, capturing 3D human motion in everyday activity settings—such as kitchens, offices, and sports fields—is challenging due to the cumbersome setup of motion capture systems. Existing large-scale 3D motion datasets [52, 57] lack paired egocentric videos, while most egocentric datasets either lack 3D motion annotations [22, 23], are small-scale [95], or have limited scene-motion correlation and diversity [48, 99]. The Nymeria dataset [17] stands out with 200+ hours of daily activity egocentric videos paired with motion capture of simple skeleton sequences, but it does not provide the standard SMPL motion representation.

To bridge this gap, we process the large-scale EgoExo4D dataset [23] to generate pseudo-ground-truth 3D motion data. We refer to this processed dataset as EE4D-Motion, which consists of 208 hours of time-synchronized 3D motion data and egocentric videos, alongside other EgoExo4D annotations. This dataset serves as an extensive benchmark for multimodal motion research.

EgoExo4D Source Data

EgoExo4D provides synchronized egocentric and exocentric video recordings of diverse activities, including cooking, dance, sports, music, healthcare, and bike repair. Egocentric videos were captured using Project Aria glasses [15]

along with the 3D trajectory of the ego camera. While EgoExo4D includes 3D body joint annotations for a subset of the dataset, these annotations are sparse, noisy, discontinuous, and lack joint angle information, making them unsuitable for motion tasks. Thus, we develop a processing pipeline to fit the SMPL-X body model to the continuous frames of EgoExo4D captures.

Fitting Pipeline

Our pipeline leverages off-the-shelf models for pose estimation and follows a two-stage fitting approach [5, 52] to obtain 3D-accurate motion groundtruth. We exclude rock climbing sequences to focus on motions occurring on a flat surface. Our pipeline consists of the following steps.

Detection & Tracking: We detect [51] and track the egocentric camera wearer in each exo view. When multiple people are present, we use the Aria 3D trajectory to identify the person of interest.

Pose Estimation: For each bounding box, we estimate 2D keypoints [91] and obtain an initial SMPL-X parameter estimate using an off-the-shelf HMR model [7]. However, single-view HMR estimates suffer from depth ambiguity and jitter in 3D translation.

Per-Frame Fitting: We initialize SMPL-X fitting by averaging HMR estimates across exo views. The fitting optimizes SMPL-X parameters $(R^r, t^r, \theta, \beta)$ using the following energy term [5]:

$$\begin{aligned}\mathcal{L}_{\text{fitting}} &= \lambda_\theta E_\theta(\theta) + \lambda_\beta E_\beta(\beta) \\ &+ \lambda_{2d} \sum_v \sigma\left(\pi_v(J(R^r, t^r, \theta, \beta)) - K_v^{2d}\right)\end{aligned}$$

where E_θ and E_β are priors for pose and shape, respectively, J is the SMPL-X 3D joint regressor, π_v is the 2D projection operator using known camera intrinsics and extrinsics of view v, K_v^{2d} represents detected 2D joints, σ is the robust Geman-McClure function [5, 19], and λ_* are energy weights.

Sequence-Level Optimization: After per-frame fitting, we refine results at the sequence level by fixing the body shape β as the average across the sequence, incorporating egocentric view detections, and adding a temporal jitter penalty to enforce smooth motion.

Filtering & Quality Control: We filter out segments with excessive jitter caused by erroneous device trajectories, suboptimal off-the-shelf model predictions, or severe occlusions across all exo views. After filtering, we retain 110 hours of smooth and accurate EE4D-Motion data for UniEgoMotion training.

Through this pipeline, EE4D-Motion provides 3D-accurate motion annotations aligned with egocentric video, enabling us to train and evaluate UniEgoMotion model.

Motion Annotations Quality

EE4D-Motion annotations can be noisy in scenes with poor exocentric visibility (e.g., kitchen, COVID testing) or large camera distances (e.g., basketball). EgoExo4D's own pose annotations are sparse and jittery, resulting in high pose error of $\sim$0.24m for EgoEgo, as reported by the authors of EgoExo4D [23], compared to $\sim$0.16m on our smoother and denser annotations. Unlike EgoEgo's synthetic dataset, where motions are scene-agnostic, EE4D-Motion provides contextually grounded motion aligned with real-world environments, which is essential for both generation and forecasting tasks.