Social EgoMesh Estimation

Luca Scofano*   Alessio Sampieri*   Edoardo De Matteis*
Indro Spinelli   Fabio Galasso

{scofano, sampieri}@diag.uniroma1.it   {dematteis, spinelli, galasso}@di.uniroma1.it

Sapienza University of Rome, Italy

Abstract

*Accurately estimating the 3D pose of the camera wearer in egocentric video sequences is crucial to modeling human behavior in virtual and augmented reality applications. The task presents unique challenges due to the limited visibility of the user's body caused by the front-facing camera mounted on their head. Recent research has explored the utilization of the scene and ego-motion, but it has overlooked humans' interactive nature. We propose a novel framework for **S**ocial **E**gocentric **E**stimation of body **ME**shes (SEE-ME). Our approach is the first to estimate the wearer's mesh using only a latent probabilistic diffusion model, which we condition on the scene and, for the first time, on the social wearer-interactee interactions. Our in-depth study sheds light on when social interaction matters most for ego-mesh estimation; it quantifies the impact of interpersonal distance and gaze direction. Overall, SEE-ME surpasses the current best technique, reducing the pose estimation error (MPJPE) by 53%. The code is available at SEEME.*

1. Introduction

Estimating the 3D motion of a person from an egocentric video sequence is a critical task with diverse applications in virtual reality (VR) and augmented reality (AR). This research is motivated by the necessity to realistically depict the entire body to enable total immersion in these environments [7]. The user typically wears a camera in such videos, capturing their surroundings from a first-person perspective. Large field of view (FOV) cameras can capture some body parts (hands, feet), which standard FOV cameras cannot, as illustrated from the first-person perspective in Figure 1.

The task, dubbed egopose estimation, is influenced by the type of camera used to record the sequences. Top-down head-mounted cameras have the best view of the wearer [47] but are intrusive due to the required displacement from the face. Front-facing cameras solve the ergonomic issue, but the downside is that the wearer is almost always invisible except when using very large FOV.

In [16], the authors introduce a two-stage framework for fish-eye head-mounted cameras, where a traditional SLAM algorithm [43] provides the camera rotation and translation. Simultaneously, a neural network exploits visible body parts to predict the egopose. Recently, EgoEgo [25] proposed an updated pipeline, relaxing the assumption on the large FOV. However, they still rely on (deep learning-based) SLAM [44] to estimate the head's rotation and translation, fed to a generative model [12] to synthesize the pose for the rest of the body. These approaches exploit the video to extract the head pose, ignoring the surrounding environment and the other actors in the scene that can provide essential cues for the egopose estimation.

For this reason, we propose a new social egocentric mesh estimation task and contribute with a novel probabilistic framework for **S**ocial **E**gocentric **E**stimation of body **ME**shes (**SEE-ME**). Humans have a social nature and are often involved in social activities in real and virtual worlds, accomplishing tasks that usually require cooperation and coordination. You2Me [33] explicitly captures the interplay between two persons from a chest-mounted camera in a controlled environment where two people interact by conversing, playing hand games, or doing sports. However, their approach does not extract high-level scene information and predicts deterministic skeleton-like representation that may not be suitable for AR/VR applications. In contrast with our work, [33] employs a person-centric coordinate system and does not estimate global orientation or translation, while our approach enables end-to-end learning of both of them. [33] also relies on OpenPose [4] to extract the interactee, which faces challenges in accurately estimating occluded keypoints [26]. We exploit EgoHMR [58] to estimate the interactee's mesh, demonstrating robustness

---

*Authors contributed equally.

(a) Wearer's first person view, looking at the interactee.
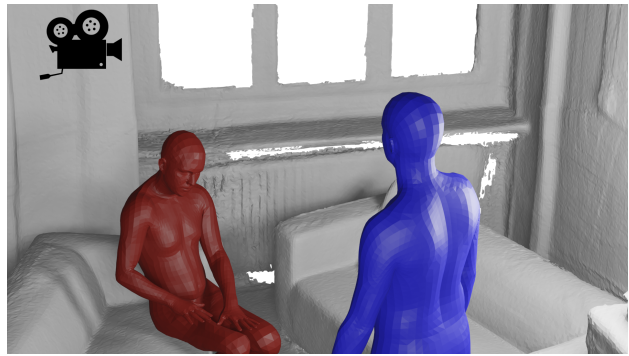
(b) Third-person view of the wearer and the interactee in a scene.

Figure 1. *(Left)* Frame from the input egocentric video stream. We experience the immersive subjective perspective of the front-facing camera wearer, but the wearer is behind the wearable and, therefore, invisible. Still, we recognize the points of interest of the wearer, parts of the scene where the action happens, and, most importantly, the interactee engaged in communication with the wearer. *(Right)* Third-view reconstruction of the *ego* mesh of the camera wearer by our proposed **SEE-ME**. Our vantage point reveals the surrounding environment, featuring a sofa and a person from an overhead perspective, leading us to infer the wearer's likely standing position. Specifically, vicinity and gaze interactions are important cues for our reconstruction as we experimentally quantify.

against occlusions, using scenes as input. In contrast, our approach leverages the scene and interactions to understand the invisible camera wearer's pose, with no input body cues required.

The motivation behind our research stems from the wearer's perspective, which highlights both their surroundings and an overlooked aspect: the presence of another actor within the scene (see Figure 1). Our unified approach links the generation of the egomesh to both the 3D scene depiction and the wearer's social interaction with the interactee. We do not rely on SLAM for the head's orientation or translation but predict them with our model. Our approach pioneers conditional Latent Diffusion in egocentric human mesh estimation, which provides a notable speedup in generation times. We build the latent space encoding human poses using a Variational Autoencoder (VAE) [20]. Subsequently, we perform conditional diffusion on this latent space. The conditioning strategies aim to guide the process by modeling the 3D point cloud of the scene in which the wearer moves and the estimated mesh of the interactee recovered from the egocentric video feed. Our model depends on the 3D scene and other actors' poses. Therefore, provided the wearer's pose, we can predict the interactee's one even if it is not paired with an egocentric video stream.

To assess the performance of our approach, we evaluate it on EgoBody [59]: the solely available dataset that features multiple individuals, an egocentric perspective, and an environment. We reach state-of-the-art results, establishing the efficacy of environmental and social components. In our research, we thoroughly examine the influence of social interactions on estimating the wearer's egomesh. We establish when this conditioning strategy has the highest impact. We use proxies for social interaction that can be extracted from

our setup. We demonstrate how proximity and eye contact between the wearer and the interactee bring the most from our conditioning. Moreover, we study the effect of future knowledge about the interactee's motion, as we humans use the experience to predict and react to the future movements of the people around us. Our framework, exposed to this information, provides a significant performance boost. To corroborate the strength of our approach, we evaluate our model on the dataset GIMO [61], which is egocentric and includes the environment but is not multi-person.

To summarize our contributions:

- We propose the task of social egomesh estimation. Emphasizing the role of other actors in 3D scenes to account for the lack of information about the camera wearer.

- We introduce SEE-ME, a unified framework based on latent diffusion that achieves state-of-the-art performance in predicting the wearer's pose and is also capable of predicting the interactee's pose.

- We perform an in-depth ablation study to highlight the scenarios in which modeling social interactions brings the most benefits.

2. Related Work

Pose Estimation from third-person cameras. 3D pose estimation from images and videos in a third-person perspective has seen significant research efforts in recent years. It can be broadly categorized into two main approaches. The first approach aims to predict the positions of joints from images and videos directly [21, 29, 32, 35, 48, 62]. The second approach employs a parametric human body

model [27] to estimate the parameters of the body model based on images or videos [8, 18, 21, 22, 29, 52]. Recent approaches have advanced the realism of human motion modeling by incorporating human dynamics. This is achieved through the utilization of learned priors [38] or physics-based priors [11, 36, 39, 41, 51, 55]. These priors are inherently formulated within the human coordinate frame.

However, a notable challenge arises when dealing with egocentric videos, which we consider in this study, where the whole body is often not visible because body joints are mostly hidden from view. By considering the physical and social aspects, we aim to provide we aim to overcome this lack of information.

Motion Estimation from Egocentric Video. There is a growing emphasis on pose estimation from egocentric videos, where various hardware setups are utilized, including fisheye cameras [2, 13, 26, 46, 49], outward-facing cameras [15, 33, 42, 53, 58, 59, 61], and additional inputs such as controllers and synchronized headsets [17, 60], all aimed at estimating a person's pose. None of them consider social interactions. However, there has been a growing interest in incorporating interactions between people in a given scene, as evidenced by numerous studies [19, 33, 58, 59]. Notably, [58, 59] both focus on estimating the interactee without relying on any additional cues from the camera wearer. In contrast, You2Me [33] predicts full-body wearer motions by observing the interaction poses of a second person in the camera view. While this approach is effective for keypoint estimation, it lacks critical information necessary for our task, such as body meshes, 3D scene details, and global rotation and orientation. Conversely, our approach involves probabilistic human mesh estimation, incorporating scene information.

In [25], a distinctive method is proposed for multi-hypothesis human mesh estimation. This approach decouples the problem by estimating head movement, employing it as a conditioning variable for a DDPM [12] model to generate plausible poses. In contrast, our approach eliminates the reliance on a three-block process involving SLAM [44], Gravitynet, and Headnet [25], tackling the problem as a unified mesh estimation task. We employ a Latent Diffusion Model to generate the entire body without preprocessing steps such as optical flow estimation and camera localization. Furthermore, our approach incorporates the social component as a conditioning factor.

Probabilistic models for human pose estimation. Due to limited image or video observations and inherent depth ambiguity, estimating a 3D human pose from a single image can give rise to numerous potential solutions, mainly when body truncation is a factor. Recent research endeavors have approached this challenge by framing it as a generative process or predicting multiple hypothetical poses. In various studies, a discrete set of hypotheses has been generated to address this issue, as seen in [3, 14, 23–26, 34, 50, 58]. We extend the advancements made in recent motion diffusion models [5, 6, 9, 45, 57]. Our approach further integrates conditioning on bystanders or individuals within the scene. It is crucial to highlight that only [58] considers the 3D scene constraint. Our methodology uses the latent diffusion process to capture the inherent ambiguity in pose estimation, incorporating flexible scene and social conditioning techniques.

3. Methodology

Modeling social interaction requires solving a series of challenging tasks. Note, as done in [25, 59], we pose the origin of our coordinate system as the camera of the wearer. First, we extract a 3D representation of the interactee's human body from a video stream. Subsequently, we need to position this representation into a depiction of the 3D environment where the interaction occurs. Hence, this is the information that SEE-ME exploits to recover the camera wearer's mesh using only the egocentric video stream, where the wearer itself is not visible. We showcase our pipeline in Figure 2. Before using social interaction and scene description to drive the generation process, we train part of our model to encode human motion in a latent space (Sec. 3.1). This dramatically reduces the dimensionality required to represent a person. In this space, we learn to synthesize human motion from pure noise using latent diffusion processes (Sec. 3.2).

Problem formalization. Our objective is to generate a plausible representation of the camera wearer $\mathbf{P^w}$, conditioned on the other person in the scene $\mathbf{P^i}$ and the 3D scene $\mathbf{S}$. We define the sequence of poses of the camera wearer as $\mathbf{P^w} = \{\mathbf{p^w}_k\}_{k=1}^{F} \in \mathbb{R}^{F \times V}$, where F and V are the number of frames and the parameters of the pose vector. Similarly, $\mathbf{P^i} = \{\mathbf{p^i}_k\}_{k=1}^{F} \in \mathbb{R}^{F \times V}$. Following literature [6, 28, 56, 57], our poses vectors $\mathbf{P^w}$ and $\mathbf{P^i}$ contain pose, translation and rotation and represents the SMPL [27] body representation. Finally, $\mathbf{S}$ is defined as a 3D scene point cloud, denoted as $\mathbf{S} \in \mathbb{R}^{N \times 3}$, where N is the number of points.

3.1. Latent Human Representation

VAEs consist of an encoder-decoder generative architecture trained to minimize the reconstruction error. We employ it to reduce the pose vector V dimensionality, projecting onto a manifold of feasible poses. In this framework, the encoder network $\mathcal{E}$, parameterized by ϕ, generates lower-dimensional embeddings $\mathbf{z} \in \mathbb{R}^{n \times D}$ as outputs based on the input poses $\mathbf{P} = \{\mathbf{p}_k\}_{k=1}^{F}$. Following VAE literature [20],
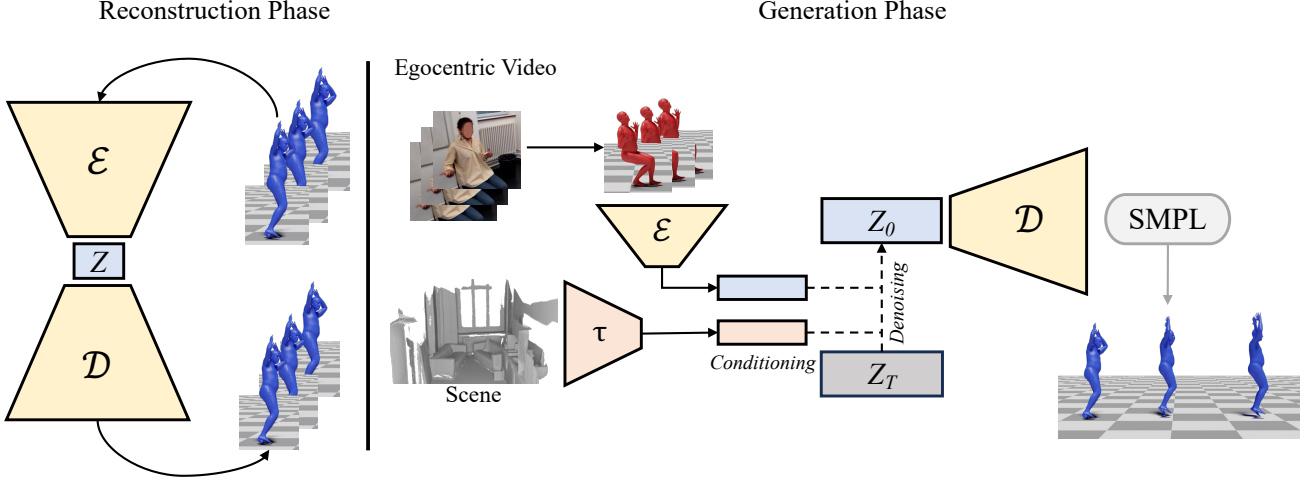
Figure 2. SEE-ME framework. On the left, we present VAE's training to learn a meaningful latent space by solving reconstruction tasks. On the right, we extract and process our conditioning strategies. Corresponding to the 3D point cloud representation of the scene and the interactee's pose extracted from the video sequence. After the conditional denoising process, we can output a SMPL representation of the wearer's pose.

$q_\phi(\mathbf{z}|\mathbf{p})$ approximates the true posterior distribution of the latent space with a multivariate Gaussian with a diagonal covariance structure:

$$q_\phi(\mathbf{z}|\mathbf{p}) = \mathcal{N}(\mathbf{z}|\mu_\phi(\mathbf{p}), \sigma_\phi^2(\mathbf{p})\mathbf{I}), \qquad (1)$$

where $\mu_\phi(\cdot)$ and $\sigma_\phi^2(\cdot)$ are the encoder's outputs. We sample from the approximated posterior $\mathbf{z}_i \sim q_\phi(\mathbf{z}|\mathbf{p}_i)$ using:

$$\mathbf{z}_i = \boldsymbol{\mu}_i + \boldsymbol{\sigma}_i^2 \odot \boldsymbol{\rho}, \qquad (2)$$

where $\mathbf{z}_i$ is a one-dimensional vector of size D, and $\boldsymbol{\rho}$ is sampled from a standard multivariate Gaussian distribution. The decoder network $\mathcal{D}$ parametrized by θ maps the sampled values back to body poses $p_\theta(\mathbf{p}|\mathbf{z})$ mapped into body meshes with the differentiable SMPL model. The network parameters are obtained by optimizing the Evidence Lower Bound (ELBO) objective, as described in [20]. The motion decoder $\mathcal{D}$ relies on a transformer decoder architecture [6, 37] with a cross-attention mechanism, taking f^* zero motion tokens as queries, where f^* is the target sequence length, and a latent $\mathbf{z} \in \mathbb{R}^{1 \times D}$ as memory, ultimately generating a human motion sequence $\hat{\mathbf{p}}_{1:f^*}$.

3.2. Ego-Mesh Estimation via Latent DDPMs

We utilize the latent-DDPM framework introduced in [40] and adapted for human motion synthesis in [6] where the diffusion process occurs on a condensed, low-dimensional motion latent space. Latent DDPM have a different approximated posterior $g(\mathbf{z}_t|\mathbf{z}_{t-1})$, denoted *diffusion process*, which gradually gradually converts latent representations $\mathbf{z}_0 = \mathbf{z}$ into random noise $\mathbf{z}_T$ in T timesteps:

$$g(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \sqrt{\bar{\alpha}_t}\mathbf{z}_{t-1}, (1 - \bar{\alpha}_t)\mathbf{I}), \qquad (3)$$

where $\bar{\alpha}_t$ is a scaling factor specific to timestep t. Then, the reverse process dubbed *denoising* gradually refines the noised vector to a suitable latent representation $\mathbf{z}_0$. Following [6, 10, 12, 40], we use the notation $\{\mathbf{z}_t\}_{t=0}^T$ to denote the sequence of noised latent vectors, with $\mathbf{z}_{t-1} = \epsilon_\psi(\mathbf{z}_t, t)$ representing the denoising operation at time step t. Here, ϵ_ψ refers to a denoising autoencoder trained to predict the denoised version of its input.

$$Loss := \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\mathbf{I}),t}[\|\epsilon - \epsilon(\mathbf{z}_t, t)\|_2^2]. \qquad (4)$$

During the denoiser's training, the encoder and decoder remain frozen. In the subsequent diffusion reverse stage, $\epsilon_\psi(\mathbf{z}_t, t)$ predicts $\hat{\mathbf{z}}_0$ through a series of T iterative denoising steps. Following this, the decoder $\mathcal{D}$ translates $\hat{\mathbf{z}}_0$ into poses $\mathbf{P}^i$ and meshes in a single forward pass.

3.3. Social Conditioning

We introduce a novel conditioning strategy focused on embedding social interaction knowledge into the generation of the wearer's mesh. A primary advantage of DDPM is its capability to incorporate signals, represented as $\mathbf{c}$, that guide the trajectories of the conditional denoising process $\epsilon_\psi(\mathbf{z}_t, t, \mathbf{c})$. In our approach, we use a pre-trained VAE encoder specifically to encode the interactee's pose $\mathbf{P}^i$ as a key conditioning element. During training, ground truth poses are utilized, while at test time, we employ the state-of-the-art EgoHMR algorithm [58]. Additionally, we incorporate scene conditioning by compressing the point-cloud representation $\mathbf{S}$ of the environment with an MLP network τ_η, as interactions are shaped by both social cues and the surroundings.

To inject these embedded conditions into the transformer-based denoiser, we concatenate different conditions and apply cross-attention [6] with the denoiser latent. Then, the conditional objective is defined as follows:

$$\mathcal{L} = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\mathbf{I}), t, \mathbf{P_i}, \mathbf{S}} [\|\epsilon - \epsilon_\psi(\mathbf{z}_t, t, \mathcal{E}_\phi(\mathbf{P_i}), \tau_\eta(\mathbf{S})\|_2^2], \quad (5)$$

where $\mathcal{E}_\phi$ and τ_η stay frozen.

4. Experiments

In this section, we validate our model against the state of the art and showcase the qualitative results of our approach. Additionally, we perform several ablation studies highlighting the impact of social interactions. Below, we define the dataset and the reference metrics.

Dataset. We employ the EgoBody [59] dataset to assess our technique, the sole available egocentric and social dataset featuring an environment and body meshes defined by parametric body models [27] for both the wearer and the interactee. The dataset is recorded using Microsoft Hololens2 [1] and contains 125 sequences at 30 fps. The total number of frames is 220k, but, as in [59], we selectively use frames where the interactee is within the wearer's field of view. The training set comprises 90000 images, the validation set consists of 23000 images, and the test set includes 62000 images. These recordings occur in 15 indoor scenes, each accompanied by 3D representations, which are recorded using an iPhone 12 Pro Max running a 3D Scanner App (for additional information, refer to [25]).

To enhance comparability, we also include the GIMO dataset [61] and benchmark SEE-Me without Int.ee—meaning the variant of SEE-ME based solely on the scene, as GIMO contains only single-person videos. Although this specific setup excludes social cues, we conduct this evaluation to further assess our overall framework. This dataset features egocentric views in single-body environments, and we adapt our model accordingly. It has been acquired from 11 different subjects performing actions in 19 3D scenes, scanned using an Apple iPhone 13 Pro Max. The dataset includes up to 125,400 egocentric images captured at 30 fps with a Hololens2 and 217 trajectories captured by an Inertial Measurement Units (IMU) system. In total, GIMO contains approximately 129,000 frames

Baselines. To estimate the wearer's pose from a first-person perspective, we evaluate our model against EgoEgo [25], the existing state-of-the-art model designed for estimating the wearer's pose from an egocentric video. EgoEgo is a probabilistic model based on DDPM [12] that leverages head poses as conditioning to generate a plausible pose. The model consists of two modules. Firstly, it

estimates the head pose by using two sub-modules. The first module estimates the global orientation and translation of the camera based on optical flow. The second module involves head pose estimation using DROID-SLAM [44]. By combining these estimates, the output is the global position of the head, comprised of pose, orientation, and trajectory. During the second phase, the estimated head is employed as a conditioning factor for a DDPM model, facilitating the generation of realistic poses. We employ EgoEgo's [25] configuration, where body poses are represented relative to the initial pose. Concerning the scene alignment, we adopt an approach in line with EgoHMR [58].

Metrics. We consider the evaluation metrics commonly employed in the current literature on egocentric human pose estimation [25, 58, 59].
Specifically, we evaluate the accuracy of our model by computing the error on the keypoints extracted from the SMPL body representation, ignoring the mesh form factor. We use SMPL due to its widespread adoption for representing virtual humans. The added realism brought by meshes allows for augmenting the fidelity of interactions, for example, by modeling collisions.

The Mean Per Joint Position Error (MPJPE) measures the average Euclidean distance between the predicted and ground truth 3D positions of individual joints across a sequence of frames:

$$\text{MPJPE} = \frac{1}{J \times T} \sum_{j=1}^{J} \sum_{t=1}^{T} \|\mathbf{p}_{j,t} - \mathbf{g}_{j,t}\|^2. \quad (6)$$

We measure the Orientation Error using the Frobenius norm of the 3x3 rotation matrix of the reference joint $\mathbf{A}_{pred}$ predicted and ground truth $\mathbf{A}_{GT}$, expressed as follows:

$$\text{Orientation Error} = ||\mathbf{A}_{pred} - \mathbf{A}_{GT}^{-1} - \mathbf{I}||_2. \quad (7)$$

To evaluate the error of the generated motion translation, we use the Euclidean distance between the predicted trajectory $\mathbf{r}^{pred}$ and the ground truth $\mathbf{r}^{GT}$:

$$\text{Translation Error} = \frac{1}{T} \sum_{t=1}^{T} ||\mathbf{r}_t^{pred} - \mathbf{r}_t^{GT}||^2. \quad (8)$$

We then compute the acceleration for predicted poses $\mathbf{a}^{pred}$ and ground truth $\mathbf{a}^{GT}$, and we define the Acceleration Error as their Euclidean distance. We employ millimeters (mm) for MPJPE and Translation, millimeters per second squared (mm/s^2) for Acceleration, and the Frobenius norm between the rotation matrices for Orientation.

4.1. Comparison with SOTA

We quantitatively assess the camera wearer's pose using the results from the EgoBody [59] dataset. As the

Table 1. Egocentric Pose Estimation Results. We activate and deactivate Scene and interactee conditioning to assess their contribution. In each version, our framework improves SoA performances by a large margin. We obtain the best results when the scene and interactee conditioning are present.

Models	Conditioning		EgoBody [59] Dataset			
	Scene	Interactee	MPJPE (mm)	Orientation Error	Translation Error (mm)	Acceleration Error (mm/s^2)
EgoEgo [25]			268	**0.40**	207	10.8
SEE-ME w/o Scene		✓	138	0.51	174	3.07
SEE-ME w/o Int.ee	✓		130	0.51	171	2.69
SEE-ME	✓	✓	**126**	0.48	**164**	**2.67**

GIMO [61] dataset lacks any interactee, we benchmark only the SEE-ME variant, focusing solely on the scene. Furthermore, we conduct several studies to explore the impact of social relation proxies and their effects.

Social egocentric human pose estimation. Table 1 compares our model quantitatively against the current leading technique [25]. We retrained the HeadNet module of their model on the EgoBody [59] dataset, specifically targeting head rotation and translation distance estimation. The rest of the model's modules, initially trained on a large-scale motion capture dataset like AMASS [31], were frozen during this process.

We improve the performance of MPJPE from 268mm to 126mm over the state-of-the-art [25] yielding a 53% enhancement in MPJPE, 21% in Translation, and a significant 75% in Acceleration. The impact of the interactee is evident, even when the wearer is not engaged in active interaction with it. Furthermore, compared to [25], conditioning solely on the interactee enhances the performance from 268mm to 138mm in MPJPE, giving a 49% increase and a substantial 72% improvement in the Acceleration error between predicted and ground truth joints.

Predicting the wearer's pose conditioning on the scene alone allows a direct comparison with EgoEgo, and our model proves again to boost performances. Going from 268mm to 130mm, we get a 51% MPJPE improvement, predicting the wearer's pose from the scene alone We get 17% and 75% improvements in translation and acceleration errors, respectively.

Egocentric human pose estimation. To reinforce the effectiveness of our approach, we further assess SEE-ME using the GIMO dataset, as there is a scarcity of 3D social egocentric environmental datasets aside from EgoBody. Thus we set to assess the quality of SEE-ME by the sole consideration of the scene. The results shown in Table 2 are comparable to EgoEgo and they should be paralleled with SEE-ME w/o Int.ee in the prior experiment, as the interactee is absent in GIMO.

SEE-ME w/o Int.ee outperforms EgoEgo on the orien-

tation estimation but it yields a larger translation error as EgoEgo leverages information from SLAM.

Overall, on the general MPJPE performance, SEE-ME w/o Int.ee outperforms EgoEgo by 7.2%, which reasserts the quality of the proposed model, across different settings.

4.2. Qualitative results

We show the results obtained from SEE-ME qualitatively, comparing them to the state-of-the-art [25]. In Figure 3, the input footage (a) is processed to extract the interactee's pose by EgoHMR [58] (b), which is then ingested by SEE-ME, alongside the encoding of the scene point cloud. The qualitative reconstructions match the quantitative evaluation of Table 1 as SEE-ME (d) adheres better to the ground truth, than EgoEgo (c) does. E.g. the actor stands up in (d) to attend to the interactee.

Figure 4 depicts a different qualitative study case: the interactee's pose (red) conditions the generation of the ego actor (blue), which SEE-ME successfully positions in front of it, taking plausible poses while conversing.

4.3. Ablation studies

Social Interaction ablation. We explore the influence of social relations through ablation studies to gain a deeper understanding of the interactee's impact. To quantify interaction, we initially employ a straightforward proxy: the distance between individuals. This is achieved by categorizing distances into three ranges based on the root joint. As illustrated in Table 3, the proximity of the interactee and wearer correlates with more pronounced effects on the final results. Notably, the most substantial improvements are observed in MPJPE (by 6%), Translation Error (by 9%), and Acceleration (by 23%) when considering proximate interactions.

The diminished performance in Translation and Acceleration beyond the 2-meter range is attributed to the wearer moving toward the interactee. As individuals draw closer, there is a heightened relative acceleration and translation, making it more susceptible to errors.

The investigation in Table 4 aims to ascertain mutual gaze fixation between two individuals using information derived from their respective head rotation matrices. Mu-

(a) Video stream of the wearer's egocentric view, not used in the model.

(b) Wearer's view of the interactee, extracted from the video stream.

(c) Prediction of the wearer's pose by EgoEgo.

(d) Prediction of the wearer's pose by SEE-ME.
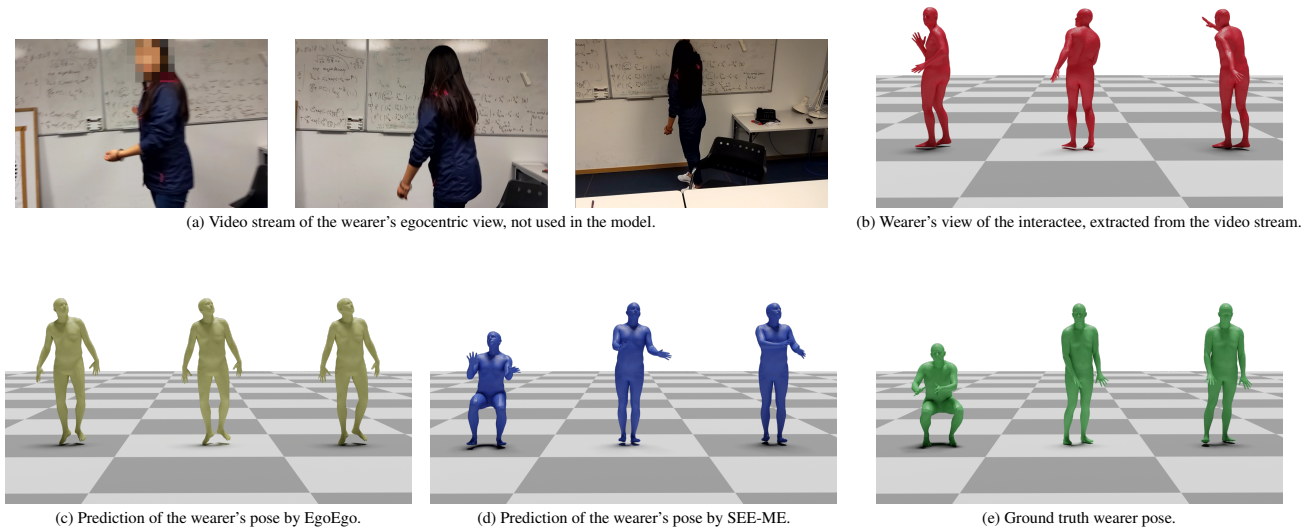

(e) Ground truth wearer pose.

Figure 3. Front view of 3 frames extracted from an egocentric sequence. We compare SEE-ME (blue) with EgoEgo (yellow), and the ground truth (green). In red we have the interactee's poses, extracted from the egocentric video sequence next to it, but not used in our model.

Models	MPJPE (mm)	Orientation Error	Translation Error (mm)
PoseReg [54]	189	1.51	1528
Kinpoly-OF [30]	404	1.52	1739
EgoEgo [25]	152	0.67	**356**
SEE-ME w/o Int.ee	**141**	**0.61**	843

Table 2. Quantitative results on GIMO [61] dataset.

Wearer-Interactee Distance Threshold (m)	MPJPE (mm)	Orientation Error	Translation Error (mm)	Acceleration Error (mm/s^2)
-	126	**0.48**	164	2.67
$d > 2$	132	**0.48**	183	2.76
$1 < d < 2$	128	0.49	161	2.73
$d < 1$	**119**	**0.48**	**156**	**2.06**

Table 3. Ablation study on the interpersonal distance between wearer and interactee. Conditioning on the interactee's pose works best when in close proximity.

tual gaze, a critical component of non-verbal communication, can indicate interpersonal engagement. The proposed method relies on converting rotation matrices to Euler angles, extracting gaze directions, and evaluating the angular deviation between these vectors. We employ $30°$ and $60°$ thresholds to determine whether two individuals are making eye contact. We look at two data subsets for each threshold: one satisfying the condition and the other not. Remarkably, both thresholds yield improved results when satisfied, affirming the initial intuition. The most substantial improvement occurs when individuals are directly looking at each other ($30°$), leading to enhancements in MPJPE from 127mm to 117mm (by 8%), from 175mm to 137mm in Translation Error (by 22%), and also enhancing the Orientation Error (14%), and Acceleration (by 9%) for when

the condition not being met.

Finally, we examine a scenario in which the wearer possesses knowledge of the interactee's future movements. This is achieved by conditioning on future frames rather than on the present. As indicated in Table 5, the MPJPE and the Translation Error exhibit improvements of 2% and 22%, respectively. The latter's enhancement can be attributed to the wearer's anticipation of the interactee's movements, reducing Translation error.

Implementation details. Our framework consists of three main components: the encoder, the decoder, and the denoiser. The reconstruction phase, which has 24 million parameters, is handled by the encoder and decoder, while the generation phase involves the denoiser (9 million parameters) and the frozen decoder. Each component has 9

Wearer-Interactee looking at each other	FOV	MPJPE	Orientation Error	Translation Error	Acceleration Error
-	-	126	0.48	164	2.67
No	60	127	0.48	173	2.71
Yes		**123**	**0.48**	**157**	**2.57**
No	30	127	0.50	175	2.79
Yes		**117**	**0.43**	**137**	**2.54**

Table 4. Ablation study on gaze directions. By considering an angle of 60 and 30 degrees, we asses if the wearer and the interactee are looking at each other. If this is the case, the conditioning boots improve the performance even more.

Wearer's input = present Interactee's input = x	MPJPE	Orientation Error	Translation Error	Acceleration Error
x = present	126	**0.48**	164	**2.67**
x = future	**123**	**0.48**	**128**	3.40

Table 5. Ablation study on wearer poses conditioned on the interactee's present and future ones. Even a little glance into the future reduces the MPJPE.
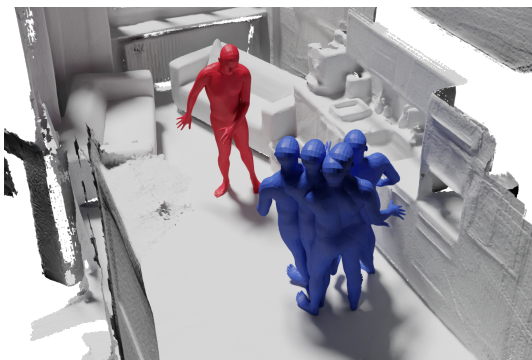

Figure 4. The interactee (red) influences the wearer's motion (blue).

layers 4 heads, and a latent space whose dimensionality is 256. The encoder and the decoder both use standard transformer layers. Based on [57], the denoiser combines classical self-attention on the latent vector with linear cross-attention between the latent vector and textual input. The batch size is 64 for the first phase and 128 for the second one, and the AdamW optimizer is used with a learning rate of 10^{-4}. The diffusion steps are set to 1000 and 20 during training and inference. We train on 8 Tesla V100 GPUs for 3k epochs for both phases.

Limitations and Future Work. We do not inject any social relations knowledge externally but leave the model free of learning from the dataset. This means that the quality of the predictions is highly dependent on the training data distribution. While we recognize the presence of social relations in the utilized dataset, EgoBody [58], we acknowledge that these relationships could be more effectively leveraged in more diverse and socially engaging datasets, such as Ego-Humans [19] for which currently both mesh recovery and

3D point cloud are not available. Additionally, we recognize the potential to enhance our model by exploiting large-motion datasets such as AMASS, and possibly others such as KIT, during the VAE training for the reconstruction phase. This would allow us to capture movements beyond the constraints of a single dataset. Currently, our model works without taking any wearer's input, but considering its egocentric observed body parts as additional data could be a research path to explore, and which could further improve our performances.

5. Conclusions

In conclusion, accurately determining the 3D pose of the camera wearer in egocentric video sequences is pivotal for advancing human behavior modeling in virtual and augmented reality applications. Despite the challenges posed by limited visibility, SEE-ME has demonstrated that the pose of the wearer can be reconstructed to an improved level of accuracy. By utilizing a latent probabilistic diffusion model, our approach integrates conditioning techniques that effectively capture both social interactions and the surrounding environment. Moreover, it is straightforward and does not require any extra preprocessing compared to other methods that utilize localization techniques. This development shows great potential for improving the realism and precision of egocentric video-based human behavior modeling, particularly for applications in augmented reality/virtual reality (AR/VR) and embodied AI.

References

[1] Microsoft hololens2. https://www.microsoft.com/en-us/hololens. 5

[2] Hiroyasu Akada, Jian Wang, Soshi Shimada, Masaki Takahashi, Christian Theobalt, and Vladislav Golyanik. Unrealego: A new dataset for robust egocentric 3d human motion capture. In *European Conference on Computer Vision*, 2022. 3

[3] Benjamin Biggs, Sébastien Ehrhardt, Hanbyul Joo, Benjamin Graham, Andrea Vedaldi, and David Novotný. 3d multi-bodies: Fitting sets of plausible 3d human models to ambiguous image data. *ArXiv*, abs/2011.00980, 2020. 3

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 1

[5] Ziyi Chang, Edmund J. C. Findlay, Hao Zhang, and Hubert P. H. Shum. Unifying human motion synthesis and style transfer with denoising diffusion probabilistic models. In *VISIGRAPP*, 2022. 3

[6] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, and Gang Yu. Executing your commands via motion diffusion in latent space. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18000–18010, 2022. 3, 4, 5

[7] Manuela Chessa, Guido Maiello, Lina K. Klein, Vivian C. Paulun, and Fabio Solari. Grasping objects in immersive virtual reality. *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 1749–1754, 2019. 1

[8] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1964–1973, 2020. 3

[9] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9760–9770, 2022. 3

[10] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *ArXiv*, abs/2105.05233, 2021. 4

[11] Erik Gärtner, Mykhaylo Andriluka, Hongyi Xu, and Cristian Sminchisescu. Trajectory optimization for physics-based reconstruction of 3d human pose from monocular video. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13096–13105, 2022. 3

[12] Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020. 1, 3, 4, 5

[13] Ryosuke Hori, Ryo Hachiuma, Mariko Isogawa, Dan Mikami, and H. Saito. Silhouette-based 3d human pose estimation using a single wrist-mounted 360° camera. *IEEE Access*, PP:1–1, 2022. 3

[14] Ehsan Jahangiri and Alan Loddon Yuille. Generating multiple diverse hypotheses for human 3d pose consistent with 2d joint detections. *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 805–814, 2017. 3

[15] Hao Jiang and Kristen Grauman. Seeing invisible poses: Estimating 3d body pose from egocentric video. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3501–3509, 2016. 3

[16] Hao Jiang and Vamsi Krishna Ithapu. Egocentric pose estimation from human vision span. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10986–10994, 2021. 1

[17] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Rene Fender, Larissa Laich, Patrick Snape, and Christian Holz. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *European Conference on Computer Vision*, 2022. 3

[18] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2017. 3

[19] Rawal Khirodkar, Aayush Bansal, Lingni Ma, Richard A. Newcombe, Minh Vo, and Kris Kitani. Egohumans: An egocentric 3d multi-human benchmark. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19750–19762, 2023. 3, 8

[20] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. 2, 3, 4

[21] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5252–5262, 2019. 2, 3

[22] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2252–2261, 2019. 3

[23] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11585–11594, 2021. 3

[24] Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3d human pose estimation with mixture density network. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9879–9887, 2019. 3

[25] Jiaman Li, C. Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17142–17151, 2022. 1, 3, 5, 6, 7

[26] Yuxuan Liu, Jianxin Yang, Xiao Gu, Yao Guo, and Guangzhong Yang. Egohmr: Egocentric human mesh recovery via hierarchical latent diffusion model. *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9807–9813, 2023. 1, 3

[27] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2015. 3, 5

[28] Yunhong Lou, Linchao Zhu, Yaxiong Wang, Xiaohan Wang, and Yezhou Yang. Diversemotion: Towards diverse human motion generation via discrete diffusion. *ArXiv*, abs/2309.01372, 2023. 3

[29] Zhengyi Luo, S. Alireza Golestaneh, and Kris M. Kitani. 3d human motion estimation via motion compression and refinement. In *Asian Conference on Computer Vision*, 2020. 2, 3

[30] 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. 7

[31] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5441–5450, 2019. 6

[32] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Trans. Graph.*, 36:44:1–44:14, 2017. 2

[33] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. You2me: Inferring body pose in egocentric video via first and second person interactions. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9887–9897, 2019. 1, 3

[34] Tuomas P. Oikarinen, Daniel C. Hannah, and Sohrob Kazerounian. Graphmdn: Leveraging graph structure and deep learning to solve inverse problems. *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9, 2020. 3

[35] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1263–1272, 2016. 2

[36] Xue Bin Peng, Angjoo Kanazawa, Jitendra Malik, P. Abbeel, and Sergey Levine. Sfv: Reinforcement learning of physical skills from videos. *ACM Trans. Graph.*, 37:178, 2018. 3

[37] Mathis Petrovich, Michael J. Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. *ArXiv*, abs/2204.14109, 2022. 4

[38] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11468–11479, 2021. 3

[39] Davis Rempe, Leonidas J. Guibas, Aaron Hertzmann, Bryan C. Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *Symposium on Computer Animation*, 2020. 3

[40] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021. 4

[41] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, Patrick P'erez, and Christian Theobalt. Neural monocular 3d human motion capture with physical awareness. *ACM Transactions on Graphics (TOG)*, 40:1 – 15, 2021. 3

[42] Takaaki Shiratori, Hyun Soo Park, Leonid Sigal, Yaser Sheikh, and Jessica K. Hodgins. Motion capture from body-mounted cameras. *ACM SIGGRAPH 2011 papers*, 2011. 3

[43] Shinya Sumikura, Mikiya Shibuya, and Ken Sakurada. Openvslam: A versatile visual slam framework. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 2292–2295, New York, NY, USA, 2019. Association for Computing Machinery. 1

[44] Zachary Teed and Jia Deng. DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. *Advances in neural information processing systems*, 2021. 1, 3, 5

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

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

[47] D. Tome, P. Peluse, L. Agapito, and H. Badino. xr-egopose: Egocentric 3d human pose from an hmd camera. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7727–7737, Los Alamitos, CA, USA, nov 2019. IEEE Computer Society. 1

[48] Hsiao-Yu Fish Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *Neural Information Processing Systems*, 2017. 2

[49] Jian Wang, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, Diogo Carbonera Luvizon, and Christian Theobalt. Scene-aware egocentric 3d human pose estimation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13031–13040, 2022. 3

[50] Tom Wehrbein, Marco Rudolph, Bodo Rosenhahn, and Bastian Wandt. Probabilistic monocular 3d human pose estimation with normalizing flows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11179–11188, 2021. 3

[51] Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. Physics-based human motion estimation and synthesis from videos. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11512–11521, 2021. 3

[52] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023. 3

[53] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Vladislav Golyanik, Shaohua Pan, Christian Theobalt, and Feng Xu. Egolocate: Real-time motion capture, localization, and mapping with sparse body-mounted sensors. *ACM Transactions on Graphics (TOG)*, 42:1 – 17, 2023. 3

[54] 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. 7

[55] Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason M. Saragih. Simpoe: Simulated character control for 3d human pose estimation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7155–7165, 2021. 3

[56] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xiaodong Shen. Generating human motion from textual descriptions with discrete representations. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14730–14740, 2023. 3

[57] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *ArXiv*, abs/2208.15001, 2022. 3, 8

[58] Siwei Zhang, Qianli Ma, Yan Zhang, Sadegh Aliakbarian, Darren P. Cosker, and Siyu Tang. Probabilistic human mesh recovery in 3d scenes from egocentric views. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7955–7966, 2023. 1, 3, 4, 5, 6, 8

[59] 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*, 2021. 2, 3, 5, 6

[60] Xiaozheng Zheng, Z. Su, Chao Wen, Zhou Xue, and Xiaojie Jin. Realistic full-body tracking from sparse observations via joint-level modeling. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14632–14642, 2023. 3

[61] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, Karen Liu, and Leonidas J. Guibas. Gimo: Gaze-informed human motion prediction in context. In *European Conference on Computer Vision*, 2022. 2, 3, 5, 6, 7

[62] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G. Derpanis, and Kostas Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4966–4975, 2015. 2