EgoGVAE: Ego-body Mesh Reconstruction via Guided Variational Autoencoder

Jaehun Jung and Wonjun Kim*

Konkuk University
{brian111725, wonjkim}@konkuk.ac.kr

Abstract. We address the problem of recovering the full-body mesh from only the head pose. This task has become essential for various applications based on head-mounted devices or smart glasses. The challenge of this task lies in estimating the pose information of unobserved body parts based solely on a single joint (i.e., head) trajectory. Several studies have begun to adopt head-conditioned generative models, however, such previous methods are costly and time-consuming due to the diffusion-based iterative process. As an alternative, we propose a simple yet novel method that leverages the latent space of the guidance network, which is designed as a variational autoencoder taking full-body poses as inputs. By enforcing latent distributions of this guidance network and our head-to-motion network to be similar, latent features sampled from the 'guided' distribution, i.e., distribution learned in our head-to-motion network, can be reliably decoded for natural representations of full-body poses even only with the head pose. One important advantage of the proposed method is that one-step sampling scheme achieves remarkably fast inference (more than 50 times faster) compared to diffusion-based approaches. Experimental results on benchmark datasets show that the proposed method efficiently improves the performance of ego-body mesh reconstruction. The code and model are publicly available at: https://github.com/DCVL-3D/EgoGVAE_release.

Keywords: 3D human mesh reconstruction · Egocentric input · Variational framework

1 Introduction

As the era of physical AI approaches, interactive wearable devices are expected to become mainstream in our daily life. At the same time, understanding 3D motions of a human wearer is essential for enabling diverse applications in such interactive environments. For example, the algorithm implemented in the smart glass like Project Aria [6] needs to figure out the 3D status of the wearer based solely on SLAM head poses for assistive operations. However, this task is quite challenging since most body parts of the wearer are unobservable in the egocentric input.

* Corresponding author

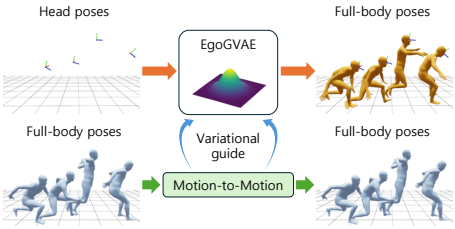

Fig. 1: Overview of the proposed method that estimates full-body poses only from head poses via the guided variational autoencoder, i.e., EgoGVAE. Note that the motion-to-motion network guides our EgoGAVE in the latent space.

In the beginning, to address this challenge, there have been meaningful attempts to utilize the hand position, which can be obtained from motion tracking sensors such as hand controllers and wristbands. This explicit observation of the hand position helps the model predict the global orientation [13] or joint-level features [34], but has the disadvantage of requiring additional hardware, leading to the limited accessibility in real-world environments. Most recently, several studies have begun to adopt the diffusion model without using such extra hardware. By taking the 3D head pose as the condition on learning the distribution of full-body poses, the prediction result of head-to-motion generation can be appropriately constrained. In addition, normalization techniques of 3D head positions, e.g., fixing starting positions of motion sequences [18] or aligning per-timestep positions to the ground plane [31], have been also applied to help train the diffusion model to focus on the intrinsic trajectory of full-body poses, regardless of the absolute 3D head position within the input sequence. Even though diffusion-based approaches for head-to-motion generation have shown promising results, those require the expensive computational cost due to the iterative process and often result in unstable full-body poses when attempting to reduce the number of denoising steps.

In this paper, we propose a simple yet novel variational method, so-called EgoGVAE, that reconstructs the full-body mesh from only the head pose of the wearer. The key idea of the proposed method is to leverage the latent space of the motion-to-motion network to guide the process of head-to-motion generation as shown in Fig. 1. Since the motion-to-motion network, which is designed to a transformer-based variational autoencoder, takes full-body poses as inputs, its latent distribution efficiently captures various characteristics of full-body poses and serves as a reliable prior for guiding the generation process. By encouraging the latent distribution of our head-to-motion network to be similar to that of the guidance network, i.e., motion-to-motion network, latent features sampled from the learned distribution can be reliably decoded for natural representations of full-body poses even only with the head pose. To align the generation process of the head-to-motion network with that of the guidance network, we concatenate the head pose with learnable tokens, which correspond to unobserved body parts. This appended embedding effectively mitigates the gap between latent spaces driven by two different inputs, i.e., head poses and full-body poses. It

is noteworthy that our guidance scheme is applied only to the training phase, leading to the fast inference with a one-step sampling compared to iterative diffusion-based approaches. The main contribution of the proposed method can be summarized as follows:

- We propose to leverage the latent space of the guidance network, which takes full-body poses as inputs, for ego-body mesh reconstruction. By enforcing latent distributions of this guidance network and our head-to-motion network to be similar, the model can easily understand the process of the head-to-motion generation and successfully generate full-body meshes with natural poses.
- We represent the latent distribution of full-body poses in a variational framework. That is, the guidance process can be efficiently achieved by simply aligning Gaussian parameters estimated from head-to-motion and motion-to-motion networks, respectively. Moreover, this design scheme, which operates with one-step sampling in inference, makes the proposed method perform very fast (more than 50 times faster) compared to diffusion-based approaches.

2 Related Work

In this Section, we give a brief review of previous studies for 3D human mesh reconstruction (HMR) from third-person inputs and discuss representative approaches specifically designed to address challenges related to the egocentric input.

2.1 3D Human Mesh Reconstruction

The majority in the field of 3D human mesh reconstruction has utilized third-person inputs. The early methods adopted a simple encoder-decoder architecture to directly regress pose and shape parameters of the SMPL model [19]. For example, Kanazawa *et al.* [14] used the end-to-end convolutional neural network with the adversarial loss to achieve the plausible mesh reconstruction. Kolotouros *et al.* [17] designed the optimization loop that enables the regression model to be trained with the strong supervision, which significantly improved the performance of human mesh reconstruction. However, in-the-wild images often contain unobservable body parts occluded by objects, other people, or body parts of the target subject. To improve the prediction accuracy under such occlusions, Kocabas *et al.* [16] proposed a part attention module that considers dependencies of visible body parts. Sun *et al.* [26] predicted body center positions of every person in a full frame, which are then used to accurately estimate SMPL parameters for multiple people. They also extended their work by explicitly considering the relative depth of each person through the bird's-eye-view body center heatmap [27]. Dwivedi *et al.* [5] utilized the discrete pose codebook as strong pose priors to enforce plausible mesh reconstructions for occluded body parts. Gwon *et*

al. [9] proposed a new concept of eigenpose to understand various occlusions in a global context. In addition, diverse methods have been introduced to further improve the performance of HMR in videos. Specifically, Ye *et al.* [30] applied both relative camera motions and tracking results to the joint optimization framework, which yields trajectories of the moving people in the world coordinate. Shin *et al.* [25] utilized 2D keypoints and camera angular velocities to estimate global human trajectories while maintaining the temporal consistency in videos. Most recently, Wang *et al.* [29] designed a multi-modal prompt encoder to refine the mesh reconstruction conditioned on the spatio-temporal information, which demonstrates the significant improvement in accuracy of HMR.

2.2 Ego-body Mesh Reconstruction

Although many methods have been explored for 3D human mesh reconstruction, the problem of estimating the 3D model from the egocentric input is still challenging due to the lack of visual clues for body parts of the target subject. To tackle this challenge, there have been impactful attempts to utilize explicit observations of the hand position. Specifically, Jiang *et al.* [13] used the trajectory of the head and two hands to estimate the global orientation, which serves as the root transform for the forward kinematic chain of the human body. Zheng *et al.* [34] also utilized such explicit observations to generate the initial joint-level features, which are calibrated through spatio-temporal transformers and subsequently fed into the SMPL regressor. Du *et al.* [4] designed a simple diffusion model to generate full-body poses with sparse IMU tracking signals in real-time. Feng *et al.* [7] first predicted upper-body poses by using positions of the head and two hands and then leveraged these prediction results as the condition for the diffusion-based generation process of lower-body poses. On the one hand, there have been several attempts to utilize visible hand observations when they appear in the field of view based on head-mounted cameras [1–3, 12]. For example, Jiang *et al.* [12] proposed to exploit intermittent hand positions and orientations only when inside a headset's field of view. Barquero *et al.* [2] proposed a rolling prediction model that leverages previously generated past motions and forecasted future motions to enable smooth transitions when hand positions are lost and later recovered.

On the other hand, to avoid unnecessary constraints in terms of using hand signals, a few studies have predicted full-body poses based solely on head poses. Li *et al.* [18] extracted the 3D head pose by refining SLAM results with the gravity direction and optical flow features, and injected it into the diffusion model to generate full-body poses. Similarly, Yi *et al.* [31] also utilized the diffusion model, but assumed that head poses are directly computed from SLAM systems introduced in Project Aria [6]. They mapped these poses onto the floor plane at each timestep to achieve both spatial and temporal invariance, helping the model train more robustly. While previous approaches have adopted diffusion models to generate full-body poses conditioned on head poses, we shift our focus to the efficient variational method that leverages the generation process learned from full-body poses as guidance for reliable ego-body mesh reconstruction.

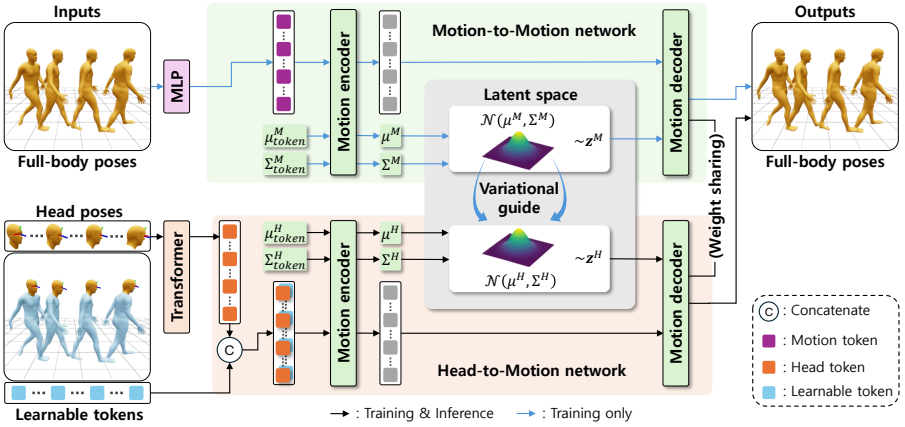

Fig. 2: The overall architecture of the proposed method for recovering full-body poses from head poses. A motion-to-motion network is used to guide the generation process of the head-to-motion network in the latent space. Note that both tokens for mean and variance are appended to input tokens of each network to formulate latent distributions as Gaussian distributions in a variational framework.

3 Proposed Method

The proposed method consists of two main networks, i.e., head-to-motion network and motion-to-motion network, both formulated as variational models. The core of the proposed method is to leverage the latent space of the motion-to-motion network for guiding the generation process of the head-to-motion network. The overall architecture of the proposed method is illustrated in Fig. 2

3.1 Task Definition

The goal of this task is to reconstruct a sequence of full-body meshes from the head trajectory. The input is defined as a sequence of T head poses, $W_{1:T} = \{w_1, w_2, ..., w_T\}$. Here, $w_t \in \mathbb{R}^{16}$ represents the head pose at timestep t, which is concatenated by the relative motion, the canonicalized orientation, and the absolute height of the head position. Note that we adopt the continuous 6D representation [36] for rotational components. These poses are defined in the world coordinate system and canonicalized to remove the effect of global rotation and translation by following the representation scheme used in [31]. This normalization yields the head trajectory that reflects only the motion of the wearer independent of global movements.

We represent the full-body pose as a sequence of T timesteps, $X_{1:T} = \{x_1, x_2, ..., x_T\}$. At each timestep t, we predict the 3D human model, i.e., $x_t = \{\theta_t, \beta, \psi_t\}$. Here, $\theta_t \in \mathbb{R}^{21\times6}$ and $\beta \in \mathbb{R}^{16}$ denote local joint rotations and time-invariant shape parameters, respectively, of the SMPL-H model [19]. $\psi_t \in \mathbb{R}^{21}$ denotes the probability of the per-joint contact. Note that the global root pose is

not directly estimated. Instead, it is recovered later through a global alignment step that uses the input head pose, as described in [31]. Therefore, the core objective of this task is to learn the head-to-motion network that can predict the corresponding sequence of full-body poses $X_{1:T}$ from the given head trajectory $W_{1:T}$, thereby enabling the reconstruction of reliable full-body meshes.

3.2 Head-to-Motion via Variational Guidance

The proposed motion-to-motion network models a latent distribution of realistic full-body poses, providing a variational guide for the process of head-to-motion generation. To do this, we adopt the transformer-based variational autoencoder that encodes pose sequences into a Gaussian latent distribution in a similar way to [24]. Specifically, local joint rotations θ_t at each timestep are tokenized to form the sequence of pose features. Then, mean token $\boldsymbol{\mu}^M_{token}$ and variance token $\boldsymbol{\Sigma}^M_{token}$ are appended to this input sequence as shown in Fig. 2. This appended input is processed by the transformer encoder, and the corresponding output embeddings are used to produce parameters $\boldsymbol{\mu}^{\boldsymbol{M}}$ and $\boldsymbol{\Sigma}^{\boldsymbol{M}}$ of the Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}^M, \boldsymbol{\Sigma}^M)$. It is noteworthy that the motion-to-motion network learns the sequence-level latent distribution, which captures various characteristics of full-body poses and serves as the motion prior for the proposed framework.

The head-to-motion network mirrors the variational framework of the guidance network, i.e., motion-to-motion network, to construct its own latent distribution $\mathcal{N}(\boldsymbol{\mu}^H, \boldsymbol{\Sigma}^H)$, which is later aligned with $\mathcal{N}(\boldsymbol{\mu}^M, \boldsymbol{\Sigma}^M)$. Unlike the guidance network, the head-to-motion network is required to estimate the latent distribution of full-body poses from only the head trajectory $W_{1:T}$. To supplement the lack of visual clues, we introduce learnable tokens $L_{token} \in \mathbb{R}^{T\times dim}$ corresponding to the rest of body parts, where T denotes the total number of timesteps. More concretely, the input sequence for the encoder of the head-to-motion network is constructed in several steps as follows. The head trajectory $W_{1:T} \in \mathbb{R}^{T\times 16}$ is first transformed by a lightweight transformer to produce head embeddings $E_W \in \mathbb{R}^{T\times dim}$. Then, learnable tokens L_{token} are concatenated with these embeddings E_W and fed into the motion encoder (see Fig. 2). Consequently, the head-to-motion network aligns its generation process with that of the guidance network through learnable tokens.

In order to enforce two latent distributions to be close to each other, we minimize the symmetric Kullback–Leibler divergence during training as follows:

$$\begin{aligned}\mathcal{L}^{H,M}_{\mathrm{KL}} = & D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{\mu}^H, \boldsymbol{\Sigma}^H)||\mathcal{N}(\boldsymbol{\mu}^M, \boldsymbol{\Sigma}^M)) \\ & + D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{\mu}^M, \boldsymbol{\Sigma}^M)||\mathcal{N}(\boldsymbol{\mu}^H, \boldsymbol{\Sigma}^H)).\end{aligned} \tag{1}$$

In addition, to regularize latent distributions of both networks, we encourage each distribution toward a standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, which is defined as follows [24]:

$$\mathcal{L}^{normal}_{\mathrm{KL}} = D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{\mu}^H, \boldsymbol{\Sigma}^H)||\mathcal{N}(\mathbf{0}, \mathbf{I})) + D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{\mu}^M, \boldsymbol{\Sigma}^M)||\mathcal{N}(\mathbf{0}, \mathbf{I})). \tag{2}$$

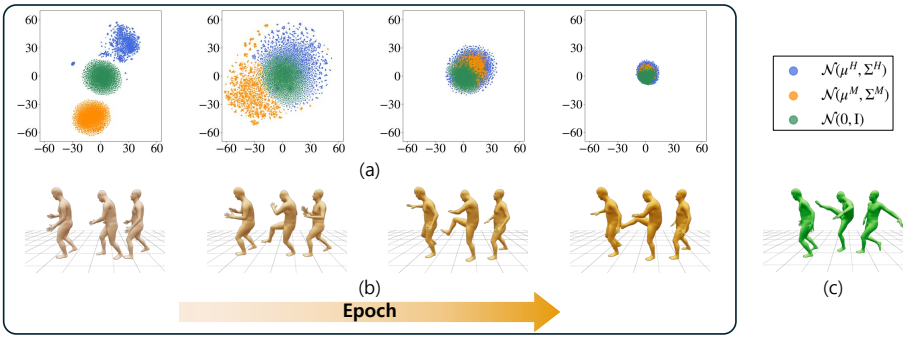

Fig. 3: Visualization for the alignment process of latent distributions. (a) The t-SNE [21] visualization of latent distributions over training epochs. (b) Results of ego-body mesh reconstruction at corresponding epochs. (c) Ground Truth.

Based on these two terms, the guidance is efficiently conducted in a variational scheme. The effect of the proposed guidance scheme is shown in Fig. 3. As can be seen, the latent distribution of the head-to-motion network $\mathcal{N}(\boldsymbol{\mu}^H, \boldsymbol{\Sigma}^H)$ is disjoint from that of the motion-to-motion network $\mathcal{N}(\boldsymbol{\mu}^M, \boldsymbol{\Sigma}^M)$ in the early stage of training (see the left side of Fig. 3(a)). As the training epoch increases, these two distributions are progressively aligned to each other, and simultaneously converge to the normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ (see the right side of Fig. 3(a)). Qualitative results of ego-body mesh reconstruction at the corresponding epoch are also shown in Fig. 3(b). It is observed that full-body poses generated from head poses become more natural as latent distributions become aligned.

Along with this guidance, $\mathbf{z}^M$ and $\mathbf{z}^H$ are sampled from each distribution via the reparameterization trick [15], and both are fed into the shared motion decoder, which is composed of transformer blocks [28], to generate full-body poses $X_{1:T}$ during training. Specifically, in the head-to-motion network, the decoder receives the latent feature $\mathbf{z}^H$ as the key and value, while the output of the motion encoder, i.e., concatenated embeddings of head and learnable tokens, is adopted as the query. In contrast, in the motion-to-motion network, the decoder takes $\mathbf{z}^M$ as the key and value, and motion tokens encoded from full-body poses are used as the query to reconstruct $X_{1:T}$. Here, $\mathbf{z}$ represents the single latent feature of full-body poses over the entire sequence, while query tokens interact with $\mathbf{z}$ to decode this sequence-level information into poses at each timestep, maintaining temporal consistency with the input sequence. At the inference time, a latent feature $\mathbf{z}^H$ is sampled in one step from the guided latent distribution $\mathcal{N}(\boldsymbol{\mu}^H, \boldsymbol{\Sigma}^H)$ and subsequently decoded to reconstruct full-body meshes. It is noteworthy that this process achieves the competitive performance at a remarkably fast inference speed compared to previous methods.

3.3 Loss Function

Both the head-to-motion network and the motion-to-motion network are jointly trained based on three loss terms, i.e., reconstruction loss $\mathcal{L}_{\text{rec}}$, KL regularization

$\mathcal{L}_{\text{KL}}$, and velocity loss $\mathcal{L}_{\text{vel}}$. First, $\mathcal{L}_{\text{rec}}$ supervises both networks by comparing their predicted full-body poses $\hat{X}^{H}_{1:T}$ and $\hat{X}^{M}_{1:T}$ with the ground truth $X_{1:T}$. Each predicted motion sequence $\hat{X}_{1:T} = \{\hat{x}_t\}_{t=1}^{T}$ comprises local joint rotations $\hat{\theta}_t$, body shape parameters $\hat{\beta}_t$, and per-joint contact probabilities $\hat{\psi}_t$. For $\hat{\theta}_t$ and $\hat{\beta}_t$, we utilize the rotation loss $\mathcal{L}_{\text{rot}} = \frac{1}{T}\sum_{t=1}^{T} \|\hat{\theta}_t - \theta_t\|_1$ and the shape parameter loss $\mathcal{L}_{\text{shape}} = \frac{1}{T}\sum_{t=1}^{T} \|\hat{\beta}_t - \beta\|_1$, where θ_t and β denote the corresponding ground truth of joint rotations and shape parameters, respectively. T indicates the total number of time steps as mentioned. Note that such joint rotations and body shape parameters are also used to compute the position loss $\mathcal{L}_{\text{pos}}$ as follows:

$$\mathcal{L}_{\text{pos}} = \frac{1}{T}\sum_{t=1}^{T} \Big\| \text{FK}\big(\hat{\theta}_t, \hat{\beta}_t\big) - \text{FK}\big(\theta_t, \beta\big) \Big\|_1, \tag{3}$$

where FK denotes forward kinematics of the SMPL-H model [19]. In addition, the contact loss $\mathcal{L}_{\text{contact}}$ is applied to supervise per-joint contact probabilities $\hat{\psi}_t$ using the binary cross-entropy loss [33]. Therefore, the reconstruction loss $\mathcal{L}_{\text{rec}}$ is formulated as follows:

$$\mathcal{L}_{\text{rec}} = \mathcal{L}_{\text{rot}} + \mathcal{L}_{\text{pos}} + \lambda_{\text{shape}}\mathcal{L}_{\text{shape}} + \lambda_{\text{contact}}\mathcal{L}_{\text{contact}}, \tag{4}$$

where weighting factors λ_{shape} and λ_{contact} are both set to 0.003 to balance their scales. $\mathcal{L}_{\text{KL}}$ is defined as the sum of $\mathcal{L}^{H,M}_{\text{KL}}$ and $\mathcal{L}^{\text{normal}}_{\text{KL}}$, which are computed by using Eqs. (1) and (2). The velocity loss $\mathcal{L}_{\text{vel}}$ is employed to enforce the temporal smoothness, as follows [34]:

$$\mathcal{L}_{\text{vel}} = \frac{1}{T-1}\sum_{t=2}^{T} \Big\| \big(\hat{p}_t - \hat{p}_{t-1}\big) - \big(p_t - p_{t-1}\big) \Big\|_1, \tag{5}$$

where $\hat{p}$ and p denote the predicted joint position and the ground truth, respectively. The total loss function is defined as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}}, \tag{6}$$

where λ_{KL} and λ_{vel} are balancing factors, which are empirically set to 0.0004 and 0.003 respectively.

3.4 Inference for Arbitrary-length Inputs

To process sequences of arbitrary length T during the test time, we employ a sliding window strategy following previous approaches [12, 13, 34]. Specifically, for initial 128 frames, we utilize the full set of predictions generated from a single forward pass. For subsequent frames, the window slides forward one frame at a time, and the last predicted full-body pose is adopted as the final output. This strategy is naturally adaptable to an online setting. The proposed method enables immediate inference by repeatedly filling the window with the first observed head pose, rather than waiting to buffer initial 128 frames.

Table 1: Performance comparisons of ego-body mesh reconstruction based on the AMASS [22] dataset (best results and second results are shown in bold and underlined, respectively).

Methods	AMASS					
	MPJPE ↓	PA-MPJPE ↓	Ground ↓	$\mathbf{T}_{head}$ ↓	Jitter ↓	Foot sliding ↓
AvatarPoser† [13]	142.2	127.5	48.1	42.9	4.74	13.5
EgoPoser† [12]	143.9	121.0	49.3	40.8	5.27	10.6
EgoEgo [18]	167.4	145.8	47.1	54.9	4.22	11.7
EgoAllo [31]	<u>119.7</u>	<u>101.1</u>	<u>26.3</u>	<u>6.2</u>	<u>4.06</u>	<u>10.2</u>
EgoGVAE (Ours)	**106.7**	**89.9**	**21.6**	**5.5**	**3.08**	**8.2**

† indicates modified versions of baselines that are retrained using only head pose inputs.

Table 2: Performance comparisons of cross-validation based on the RICH [10] dataset (best results and second results are shown in bold and underlined, respectively). Note that all methods here are trained by using the AMASS [22] dataset.

Methods	RICH					
	MPJPE ↓	PA-MPJPE ↓	Ground ↓	$\mathbf{T}_{head}$ ↓	Jitter ↓	Foot sliding ↓
AvatarPoser† [13]	297.5	287.4	281.8	75.6	4.43	12.4
EgoPoser† [12]	314.4	305.9	161.3	108.5	5.12	**9.5**
EgoEgo [18]	210.7	180.1	93.6	134.5	**3.61**	12.7
EgoAllo [31]	<u>193.1</u>	<u>172.9</u>	<u>71.9</u>	<u>7.7</u>	5.07	15.9
EgoGVAE (Ours)	**179.9**	**165.9**	**59.0**	**5.4**	<u>3.87</u>	<u>11.9</u>

† indicates modified versions of baselines that are retrained using only head pose inputs.

4 Experimental Results

4.1 Implementation Details

All experiments were conducted by using the PyTorch framework [23], running on an Intel E5-1650 v4@3.60GHz CPU and an NVIDIA RTX 3090 Ti GPU. We employed the AdamW optimizer [20] to train all model parameters, with momentum factors β_1 and β_2 set to 0.9 and 0.999, respectively. The proposed method is trained for 300 epochs with the batch size of 32. The learning rate is fixed to 1×10^{-4} throughout the training process. Inputs are randomly windowed to a length of 128 from pairs of head and full-body motion sequences for training.

4.2 Datasets and Evaluation Metrics

Dataset. For the performance evaluation of the proposed method, two representative benchmarks, i.e., AMASS [22] and RICH [10], are employed. We follow the settings used in previous work [31] with the same split of the AMASS dataset. For the RICH dataset, we use the test split only for the evaluation as introduced in [10]. Given synthetic device poses for both datasets, we adopt the annotation process [31], where a central pupil frame is anchored between vertices corresponding to left and right pupils in the mesh.

Evaluation metrics. For the performance evaluation on 3D human mesh datasets, we use four metrics, i.e., mean per joint position error (MPJPE) [11],

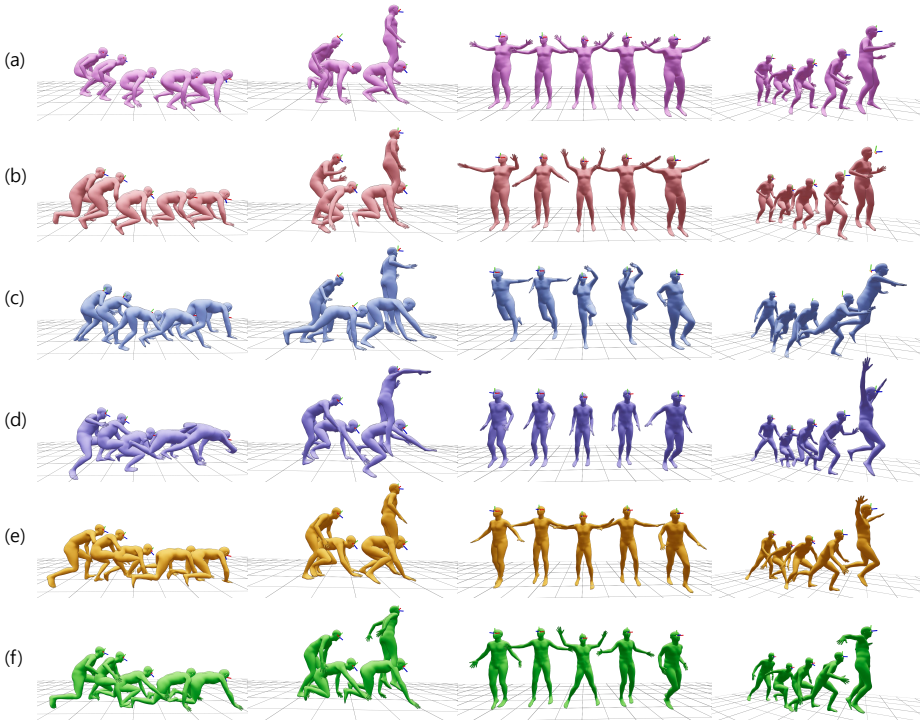

Fig. 4: Results of ego-body mesh reconstruction on the AMASS [22] dataset. (a) Results by AvatarPoser [13]. (b) Results by EgoPoser [12]. (c) Results by EgoEgo [18]. (d) Results by EgoAllo [31]. (e) Results by the proposed method (i.e., EgoGVAE). (f) Ground Truth.

Procrustes-aligned mean per joint position error (PA-MPJPE) [35], grounding metric (Ground) [34], and mean head joint position error ($\mathbf{T}_{\text{head}}$) [31]. Specifically, MPJPE measures the average value of the Euclidean distance between the predicted 3D joint and the corresponding ground truth. PA-MPJPE indicates MPJPE calculated after applying the Procrustes analysis to the estimated body mesh. Note that Ground denotes the average value of the vertical distance between the lowest position of predicted joints and that of ground truth joints to evaluate floating and ground penetration artifacts. $\mathbf{T}_{\text{head}}$ computes the average value of the Euclidean distance between the predicted position of the head joint and the corresponding ground truth.

Besides the above pose accuracy, we also use two metrics to assess the quality of the generated motion, i.e., Jitter [8] and Foot sliding metrics, as defined in [34]. Specifically, Jitter measures the average value of the jerk, i.e., the time derivative of acceleration, over the predicted motion sequence to evaluate the temporal smoothness. Foot sliding computes the average horizontal displacement of foot joints between consecutive frames to detect sliding artifacts.

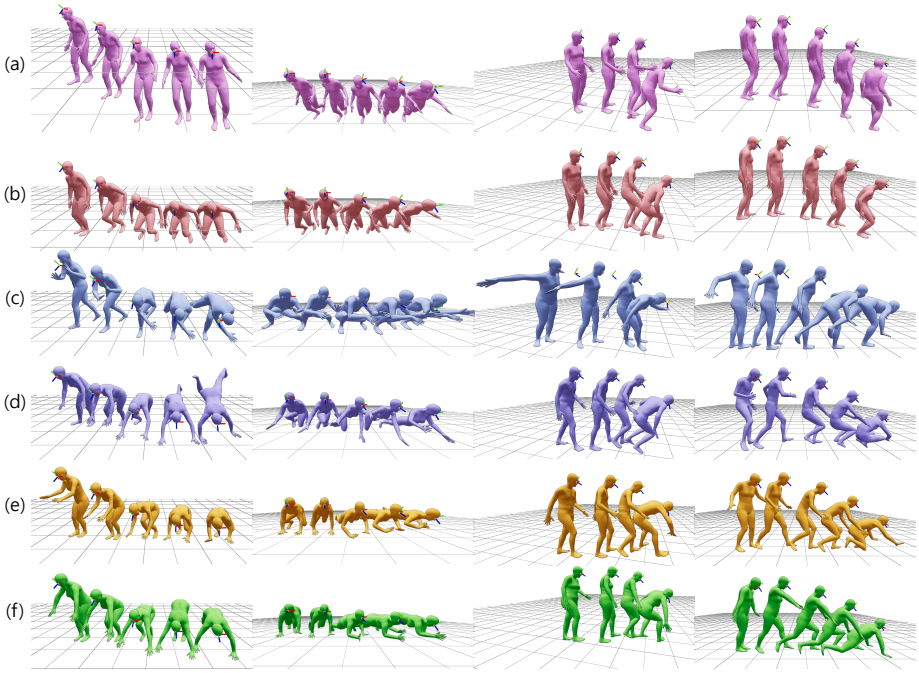

Fig. 5: Results of ego-body mesh reconstruction on the RICH [10] dataset. (a) Results by AvatarPoser [13]. (b) Results by EgoPoser [12]. (c) Results by EgoEgo [18]. (d) Results by EgoAllo [31]. (e) Results by the proposed method (i.e., EgoGVAE). (f) Ground Truth.

4.3 Performance Evaluation

Quantitative evaluation. To demonstrate the effectiveness of the proposed method in ego-body mesh reconstruction, we compare ours with previous methods in ego-body mesh reconstruction, i.e., AvatarPoser [13], EgoPoser [12], EgoEgo [18], and EgoAllo [31], using input sequences of 128 frames. Note that AvatarPoser [13] and EgoPoser [12] utilize both head and hand positions as inputs, which makes them not directly comparable to our head-only setting. To be fair, we retrained both models using only head pose inputs on the AMASS [22] dataset. As shown in Table 1, the proposed method achieves the meaningful performance improvement compared to previous approaches. Specifically, the proposed method achieves MPJPE of 106.7, PA-MPJPE of 89.9, Ground of 21.6, and $\mathbf{T}_{head}$ of 5.5 on the AMASS dataset with the input sequence of 128 frames, which demonstrates 10.9% improvement in MPJPE, 11.1% improvement in PA-MPJPE, 17.9% improvement in Ground, and 11.3% improvement in head position error, respectively. Moreover, the proposed method also achieves lower Jitter of 3.08 and lower Foot sliding of 8.2 compared to the best-performed method, i.e., EgoAllo [31]. Based on results shown in Table 1, we could confirm that the proposed guidance scheme is effective to alleviate the ill-posed properties of head-to-motion generation while providing the reliable performance of ego-body

Table 3: Performance comparisons of ego-body mesh reconstruction based on the real-world EgoBody [32] dataset (best results and second results are shown in bold and underlined, respectively).

Methods	EgoBody		
	MPJPE ↓	PA-MPJPE ↓	Jitter ↓
EgoEgo [18]	179.9	156.0	**1.40**
EgoAllo [31]	<u>172.9</u>	<u>149.7</u>	1.77
EgoGVAE	**162.9**	**140.4**	<u>1.42</u>

Table 4: Efficiency comparison with diffusion-based methods (best results and second results are shown in bold and underlined, respectively).

Methods	128 frames		
	Params(M)↓	FLOPs(G)↓	Time(sec)↓
EgoEgo [18]	**10.97**	3025.13	7.222
EgoAllo [31]	50.45	<u>378.16</u>	<u>1.510</u>
EgoGVAE	<u>13.88</u>	**3.45**	**0.026**

mesh reconstruction. Furthermore, we conduct cross-validation by training all methods on the AMASS [22] dataset and evaluating them on the RICH [10] dataset. The result of the performance comparison is shown in Table 2. As can be seen, non-diffusion models, i.e., AvatarPoser [13] and EgoPoser [12], suffer from the significant performance drop under this setting. In contrast, the proposed method still achieves the reliable performance on the benchmark dataset compared to state-of-the-art methods. Besides, we compare the proposed method with diffusion-based methods, such as EgoEgo [18] and EgoAllo [31], on the real-world EgoBody [32] dataset (see Table 3). This result shows that the proposed method still provides the reliable performance of ego-body mesh reconstruction, even when head pose inputs are captured by real VR/AR devices and contain realistic noise.

Qualitative evaluation. Several results of ego-body mesh reconstruction for the AMASS dataset are shown in Fig. 4. We can see that the proposed method successfully recovers 3D human mesh only with head trajectories. Specifically, in the third column of Fig. 4, previous methods generate plausible full-body poses, but they occasionally tend to generate poses with either static arms or arbitrary movements. In contrast, the proposed method reconstructs plausible arm poses based on the latent distribution guided by full-body poses. In particular, the proposed method yields robust results even when the significant change in the head trajectory occurs as shown in the second and fourth columns of Fig. 4. This is further evident in its ability to effectively estimate full-body poses under extreme conditions, e.g., motions during exercise, as shown in Fig. 5. In particular, previous methods often fail to accurately predict lower-body poses, which leads to the significant performance drop due to effects of foot penetration or floating legs, whereas the proposed method yields plausible full-body poses even under such complicated scenarios. Based on this, it is thought that our EgoGVAE works robustly for various types of full-body poses without significant distortions of the pose structure.

Efficiency. We compare the model efficiency of the proposed method with that of diffusion-based methods, i.e., EgoEgo [18] and EgoAllo [31], as shown in Table 4. We analyze three metrics, i.e., the number of parameters (Params), floating point operations per second (FLOPs), and inference time (Time), which are conducted based on sequences of 128 frames with RTX 3090 Ti GPU. As

Table 5: Performance analysis of the proposed method according to changes in key components of the head-to-motion network on the AMASS [22] dataset.

Guidance	Learnable tokens	AMASS	
		MPJPE ↓	PA-MPJPE ↓
✗	✗	128.3	110.3
✗	✓	125.6	108.6
✓	✗	113.7	94.5
✓	✓	106.7	89.9

Table 6: Performance analysis of the proposed method according to different combinations of loss terms on the AMASS [22] dataset.

$\mathcal{L}_{KL}^{normal}$	$\mathcal{L}_{vel}$	AMASS	
		MPJPE ↓	PA-MPJPE ↓
✗	✗	116.5	99.4
✗	✓	114.6	97.9
✓	✗	109.8	91.8
✓	✓	106.7	89.9

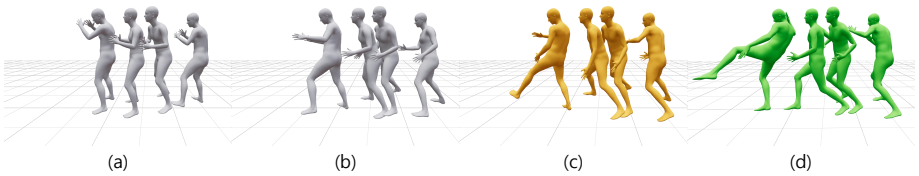

Fig. 6: (a) Result by removing both guidance and learnable tokens, i.e., baseline. (b) Result by baseline + guidance. (c) Result by baseline + guidance + learnable tokens (our method). (d) Ground Truth.

can be seen, diffusion-based approaches perform with high computational costs, however, the proposed method shows a huge advantage in the efficiency of the inference time. Specifically, EgoEgo [18] has a low number of parameters, i.e., 10.97M, but the iterative process in its diffusion model requires 7.222 seconds and 3025.13G FLOPs to process 128 frames. The state-of-the-art method, i.e., EgoAllo [31], reduces the sampling steps from 1000 to 30, but it significantly increases the number of parameters to 50.45M and still requires 1.510 seconds and 378.16G FLOPs to process 128 frames. In contrast, the proposed method achieves the lowest FLOPs of 3.45G and the fastest inference time of 0.026 seconds, which is more than 50 times faster compared to diffusion-based methods.

4.4 Ablation Study

In this subsection, we conduct comparative experiments on the AMASS [22] dataset to verify the effectiveness of the proposed method. Specifically, we analyze the impact of key components on ego-body mesh reconstruction by systematically evaluating performance variations of our model.

Effect of EgoGVAE. First of all, we check the effect of the guidance by the motion-to-motion network in a variational way. As can be seen in Table 5, it is easy to see that enforcing the latent distribution of the head-to-motion network to be similar to that of the motion-to-motion network significantly improves the reconstruction accuracy (128.3 → 113.7 in MPJPE and 110.3 → 94.5 in PA-MPJPE). In what follows, we also check the effect of learnable tokens, which are appended to input tokens for consideration of other body parts except the

Table 7: Performance analysis of the proposed method according to different network designs for guiding the head-to-motion network on the AMASS [22] dataset.

Methods	MPJPE ↓	PA-MPJPE ↓	Ground ↓	Jitter ↓
Ours (learning a latent-to-latent mapping)	128.1	111.9	40.4	3.26
Ours (using a pre-trained motion prior)	111.2	94.5	23.9	3.19
Ours (default setting)	106.7	89.9	21.6	3.08

head. From the result shown in Table 5, we can see that this appended input is helpful for ego-body mesh reconstruction rather than simply using the head token. By using these two components together, i.e., the proposed method, the best performance can be achieved. We also provide the visual comparison in Fig. 6. As can be seen, the lower-body motions are not accurately reconstructed without guidance and learnable tokens (see Fig. 6(a)). By adopting the guidance, the dynamics of the lower-body motions has been increased as shown in Fig. 6(b). Finally, full-body poses become more flexible and natural with learnable tokens (see Fig. 6(c)).

Effect of loss terms. We analyze the effect of each loss term and the corresponding result is shown in Table 6. Note that core loss terms, i.e., the reconstruction loss $\mathcal{L}_{rec}$ and the symmetric Kullback-Leibler divergence $\mathcal{L}_{KL}^{H,M}$, are applied to all the experiments for the stability of the learning process. As shown in Table 6, the performance is significantly dropped by removing $\mathcal{L}_{KL}^{normal}$. This means that regularization of latent distributions plays an important role in our guidance-based learning scheme. Furthermore, we can see that removing $\mathcal{L}_{vel}$ results in the degradation of 3.1 in MPJPE and 1.9 in PA-MPJPE, respectively. This analysis confirms that both loss terms are essential for the best performance.

Effect of network designs. We investigate different network designs for guiding the head-to-motion network with the motion-to-motion network, and the corresponding results are shown in Table 7. The first row presents the result of the variant where a mapping between the latent space of the head-to-motion network and that of the motion-to-motion network is learned instead of aligning latent distributions of two networks. The second row shows the result of using a pre-trained motion-to-motion network as a fixed motion prior without joint training. As can be seen, both variants exhibit the performance degradation across all metrics compared to our default setting. This result demonstrates that aligning latent distributions of two networks through joint training is effective for guiding the process of head-to-motion generation.

4.5 Quantitative Evaluation on Extended Settings

Longer sequence evaluation. During test time, we adopt the sliding window strategy described in subsection 3.4 to process longer sequences. To verify stability in this setting, we compare ours with diffusion-based methods, i.e., EgoEgo [18] and EgoAllo [31], using input sequences of 256 frames. Specifically, we conduct this evaluation by filtering out sequences in the AMASS [22] dataset shorter than 256 frames. Note that baselines follow the protocol of EgoAllo [31]

Table 8: Performance comparisons based on the AMASS [22] dataset following the inference setting of longer sequences (best results and second results are shown in bold and underlined, respectively).

Methods	256 frames		
	MPJPE ↓	PA-MPJPE ↓	Jitter ↓
EgoEgo [18]	140.8	105.5	<u>4.94</u>
EgoAllo [31]	<u>113.4</u>	<u>91.8</u>	5.29
EgoGVAE	**95.8**	**73.9**	**4.54**

Table 9: Performance comparisons based on the AMASS [22] dataset following the online inference setting (best results and second results are shown in bold and underlined, respectively).

Methods	Online setting		
	MPJPE ↓	Jitter ↓	Time (ms) ↓
AvatarPoser [13]	<u>142.2</u>	<u>4.74</u>	**1.5**
EgoPoser [12]	143.9	5.27	<u>1.7</u>
EgoGVAE	**119.3**	**4.63**	26

to process longer sequences. As shown in Table 8, the proposed method achieves the best performance across all metrics, notably the lowest Jitter of 4.54. This result could confirm that our learned latent distribution ensures the temporal consistency beyond the training length.

Online evaluation. To further validate the real-time performance, we compare ours with non-diffusion baselines, i.e., AvatarPoser [13] and EgoPoser [12], following the inference process described in subsection 3.4. For this evaluation, we utilize the AMASS [22] dataset without cropping sequences into fixed-length subsequences. The result of the performance comparison is shown in Table 9. While previous methods achieve faster inference compared to our method, they exhibit notably lower reconstruction accuracy. In contrast, the proposed method achieves the lowest MPJPE of 119.3 and the lowest Jitter of 4.63, with an inference time of 26 milliseconds per frame, which is still enough for real-time applications.

5 Conclusion

We propose a simple yet powerful method for full-body mesh reconstruction from only the head pose of the wearer. The key idea of the proposed method is to leverage the latent space of the motion-to-motion network, which is a variational autoencoder that takes full-body poses as inputs, to guide the head-to-motion network. By enforcing two latent distributions, which are encoded from this guidance network and the head-to-motion network respectively, to be similar, the proposed method can produce natural full-body poses given the head pose. Experimental results on benchmark datasets show that our EgoGVAE reliably conducts ego-body mesh reconstruction even with complicated scenarios. Moreover, the proposed method works very fast compared to previous methods while showing the significant performance improvement.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2026-25471545).

References

1. Aliakbarian, S., Saleh, F., Collier, D., Cameron, P., Cosker, D.: HMD-NeMo: Online 3D avatar motion generation from sparse observations. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. pp. 9622–9631 (2023) 4
2. Barquero, G., Bertsch, N., Marramreddy, M., Chacón, C., Arcadu, F., Rigual, F., He, N.S., Palmero, C., Escalera, S., Ye, Y., Kips, R.: From sparse signal to smooth motion: Real-time motion generation with rolling prediction models. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. pp. 1850–1860 (2025) 4
3. Chi, S., Huang, P.H., Sachdeva, E., Ma, H., Ramani, K., Lee, K.: Estimating ego-body pose from doubly sparse egocentric video data. In: Proc. Adv. Neural Inform. Process. Syst. vol. 37, pp. 55178–55203 (2024) 4
4. Du, Y., Kips, R., Pumarola, A., Starke, S., Thabet, A., Sanakoyeu, A.: Avatars Grow Legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. pp. 481–490 (2023) 4
5. Dwivedi, S.K., Sun, Y., Patel, P., Feng, Y., Black, M.J.: TokenHMR: Advancing human mesh recovery with a tokenized pose representation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. pp. 1323–1333 (2024) 3
6. Engel, J., Somasundaram, K., Goesele, M., Sun, A., Gamino, A., Turner, A., Talattof, A., Yuan, A., Souti, B., Meredith, B., et al.: Project Aria: A new tool for egocentric multi-modal ai research. arXiv:2308.13561 (2023) 1, 4
7. Feng, H., Ma, W., Gao, Q., Zheng, X., Xue, N., Xu, H.: Stratified avatar generation from sparse observations. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. pp. 153–163 (2024) 4
8. Flash, T., Hogan, N.: The coordination of arm movements: an experimentally confirmed mathematical model. J. Neurosci. **5**(7), 1688–1703 (1985) 10
9. Gwon, M.G., Um, G.M., Cheong, W.S., Kim, W.: Eigenpose: Occlusion-robust 3D human mesh reconstruction. IEEE Trans. Image Process. **34**, 2379–2391 (2025) 4
10. Huang, C.H.P., Yi, H., Höschle, M., Safroshkin, M., Alexiadis, T., Polikovsky, S., Scharstein, D., Black, M.J.: Capturing and inferring dense full-body human-scene contact. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. pp. 13274–13285 (2022) 9, 11, 12
11. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Trans. Pattern Anal. Mach. Intell. **36**(7), 1325–1339 (2013) 9
12. Jiang, J., Streli, P., Meier, M., Holz, C.: EgoPoser: Robust real-time egocentric pose estimation from sparse and intermittent observations everywhere. In: Proc. Eur. Conf. Comput. Vis. pp. 277–294 (2024) 4, 8, 9, 10, 11, 12, 15
13. Jiang, J., Streli, P., Qiu, H., Fender, A., Laich, L., Snape, P., Holz, C.: AvatarPoser: Articulated full-body pose tracking from sparse motion sensing. In: Proc. Eur. Conf. Comput. Vis. pp. 443–460 (2022) 2, 4, 8, 9, 10, 11, 12, 15
14. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. pp. 7122–7131 (2018) 3
15. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: Proc. Int. Conf. Learn. Represent. (2014) 7
16. Kocabas, M., Huang, C.H.P., Hilliges, O., Black, M.J.: PARE: Part attention regressor for 3D human body estimation. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. pp. 11127–11137 (2021) 3

17. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. pp. 2252–2261 (2019) 3
18. Li, J., Liu, K., Wu, J.: Ego-body pose estimation via ego-head pose estimation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. pp. 17142–17151 (2023) 2, 4, 9, 10, 11, 12, 13, 14, 15
19. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. In: ACM Trans. Graph. pp. 851–866 (2023) 3, 5, 8
20. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: Proc. Int. Conf. Learn. Represent. (2019) 9
21. Maaten, L.v.d., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**, 2579–2605 (2008) 7
22. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. pp. 5442–5451 (2019) 9, 10, 11, 12, 13, 14, 15
23. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: Proc. Adv. Neural Inform. Process. Syst. pp. 1–4 (2017) 9
24. Petrovich, M., Black, M.J., Varol, G.: TEMOS: Generating diverse human motions from textual descriptions. In: Proc. Eur. Conf. Comput. Vis. pp. 480–497 (2022) 6
25. Shin, S., Kim, J., Halilaj, E., Black, M.J.: WHAM: Reconstructing world-grounded humans with accurate 3D motion. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. pp. 2070–2080 (2024) 4
26. Sun, Y., Bao, Q., Liu, W., Fu, Y., Black, M.J., Mei, T.: Monocular, one-stage, regression of multiple 3D people. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. pp. 11179–11188 (2021) 3
27. Sun, Y., Liu, W., Bao, Q., Fu, Y., Mei, T., Black, M.J.: Putting people in their place: Monocular regression of 3D people in depth. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. pp. 13243–13252 (2022) 3
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Proc. Adv. Neural Inform. Process. Syst. vol. 30 (2017) 7
29. Wang, Y., Sun, Y., Patel, P., Daniilidis, K., Black, M.J., Kocabas, M.: PromptHMR: Promptable human mesh recovery. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. pp. 1148–1159 (2025) 4
30. Ye, V., Pavlakos, G., Malik, J., Kanazawa, A.: Decoupling human and camera motion from videos in the wild. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. pp. 21222–21232 (2023) 4
31. Yi, B., Ye, V., Zheng, M., Li, Y., Müller, L., Pavlakos, G., Ma, Y., Malik, J., Kanazawa, A.: Estimating body and hand motion in an ego-sensed world. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. pp. 7072–7084 (2025) 2, 4, 5, 6, 9, 10, 11, 12, 13, 14, 15
32. Zhang, S., Ma, Q., Zhang, Y., Qian, Z., Kwon, T., Pollefeys, M., Bogo, F., Tang, S.: EgoBody: Human body shape and motion of interacting people from head-mounted devices. In: Proc. Eur. Conf. Comput. Vis. pp. 180–200 (2022) 12
33. Zhang, Y., Zhang, H., Hu, L., Zhang, J., Yi, H., Zhang, S., Liu, Y.: ProxyCap: Real-time monocular full-body capture in world space via human-centric proxy-to-motion learning. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. pp. 1954–1964 (2024) 8

34. Zheng, X., Su, Z., Wen, C., Xue, Z., Jin, X.: Realistic full-body tracking from sparse observations via joint-level modeling. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. pp. 14678–14688 (2023) 2, 4, 8, 10
35. Zhou, X., Zhu, M., Pavlakos, G., Leonardos, S., Derpanis, K.G., Daniilidis, K.: MonoCap: Monocular human motion capture using a cnn coupled with a geometric prior. IEEE Trans. Pattern Anal. Mach. Intell. **41**(4), 901–914 (2018) 10
36. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. pp. 5745–5753 (2019) 5