Probabilistic Human Mesh Recovery in 3D Scenes from Egocentric Views **Supplementary Material**

A. Architecture Details

The detailed model architecture is illustrated in Fig. S1. The local scene encoder $E_{\mathcal{S}}$ is an MLP network consisting of several residual blocks to encode the cropped input scene point cloud of M points (translated by the estimated body translation $\hat{\gamma}$ from the camera coordinate system) into a 512-d scene feature. In the diffusion denoiser D, for each joint j, we use the 6D representation [9] to represent the joint rotations. A linear layer first maps the input noisy pose parameters $\boldsymbol{\theta}_t^j$ into a 512-d pose embedding . The timestep t is embedded by an MLP with the sinusoidal function. The pose embedding is concatenated with the corresponding context embedding (including the image feature, scene feature, timestep embedding, $\mathcal{B}$, $\mathcal{K}$, and the estimated body translation $\hat{\gamma}$) as the input feature for node j in the GCN. The GCN module consists of an input GCN layer, followed by four residual modulated GCN blocks [10] and a final GCN layer, which outputs the clean pose parameters $\hat{\boldsymbol{\theta}}_0^j$ for each joint j.

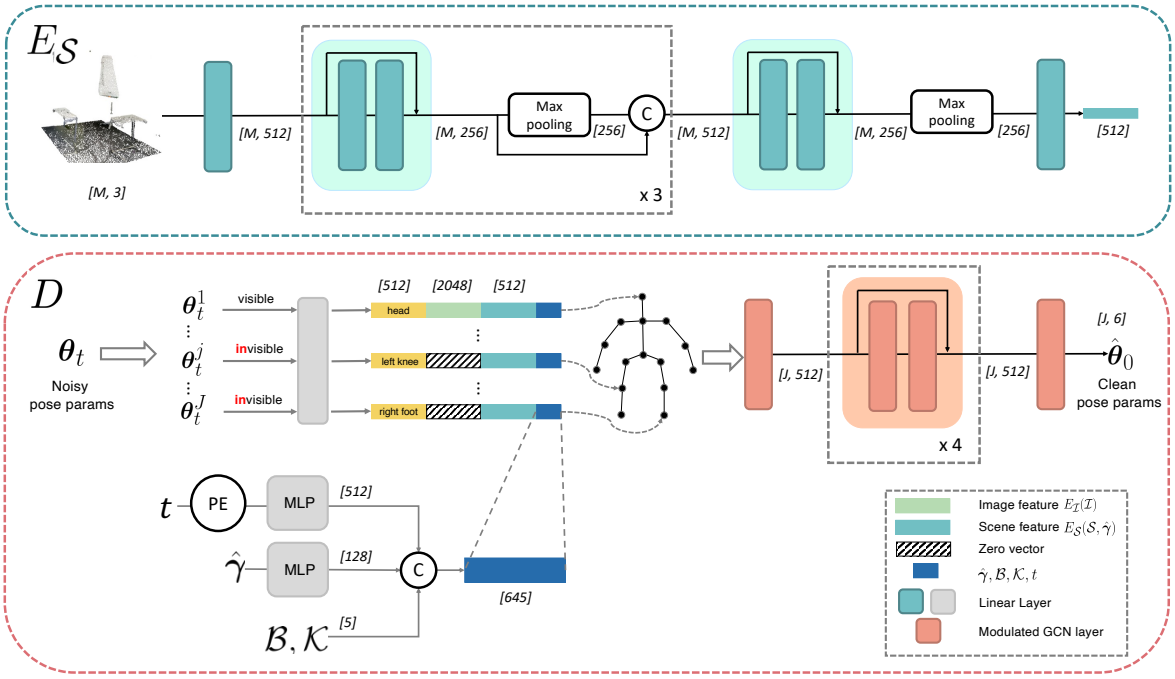

Figure S1: **Architecture details for the local scene encoder $E_{\mathcal{S}}$ and the diffusion denoise D.** Numbers in '[]' indicates the corresponding feature dimension. PE stands for positional encoding.

B. Body Translation Estimator

As the body translation is coupled with the body shape and pose parameters in the SMPL model, an accurate estimation of γ also relies heavily on the learning of the body pose $\boldsymbol{\theta}$ and shape $\boldsymbol{\beta}$. We adopt ProHMR [6] as the backbone for the body translation estimator, which estimates (γ, $\boldsymbol{\theta}$, $\boldsymbol{\beta}$) jointly, but with three major modifications.

First, the scale ambiguity between γ and $\boldsymbol{\beta}$ poses great challenges to the accurate global translation prediction in [6] with a single image as the input. We leverage the 3D scene point cloud $\mathcal{S}$ with a global scene encoder to extract the global scene feature. The global scene encoder has the same architecture as the local scene encoder in Fig. S1, but with the full scene point cloud $\mathcal{S}$ in the camera coordinate system as the input. The encoded global scene feature is concatenated with the image feature (encoded by a ResNet50 backbone [1]) as the conditioning input to the normalizing flow. Second, existing works [3, 4, 6]

take the cropped image (containing the target person, resized to a fixed resolution) as the input, which discards the location information in the full image camera coordinate system. The ignorance of the original camera introduces additional ambiguity and results in inaccurate estimations of global information. Besides the cropped image features, we additionally feed the bounding box information $\mathcal{B}$ (the same as in Eq. (4) in the main paper) to the network to provide global-aware features. On top of that, the predicted body is transformed back to the original camera coordinate system and the 2D keypoint reprojection loss is calculated in the full image instead of the cropped image. The projected 2D keypoints in the original image have similar perspective distortions with the person in the full image, offering better supervision for 3D predictions [7]. Last but not least, the model is further conditioned on the camera intrinsics $\mathcal{K}$ such that it can be adapted to different cameras and headsets.

Here the supervision on body pose $\boldsymbol{\theta}$ and body shape $\boldsymbol{\beta}$ provides auxiliary information for the accurate estimation of the body translation. And we employ this straightforward scene-conditioned model as the baseline method ProHMR-scene in Sec. 5.3 in the main paper. However, the 3D scene features here are not sufficiently localized to learn fine-grained human-scene interactions for local body pose (as demonstrated in Sec. 5.4 in the main paper), thus we only take the predicted translation $\hat{\boldsymbol{\gamma}}$ from this model, and propose the scene-conditioned pose diffusion model for better local pose reasoning.

To train the body translation estimator we employ the same training objectives as in [6], but with the 2D keypoint re-projection loss calculated in the full image frame. The trained model also serves as the ProHMR-scene-*orig* baseline in Sec. 5.3 in the main paper.

C. Implementation Details

Training details. The image encoder $E_{\mathcal{I}}(\mathcal{I})$ is loaded from the pretrained weights from [6] (for our method and all baseline methods). For the ProHMR baseline, we load the entire pretrained checkpoint from [6] and fine-tune it on EgoBody dataset. In our model, the body translation estimator and the local pose diffusion model are trained separately. The diffusion model is trained with the ground truth body translation $\boldsymbol{\gamma}$. During inference, we use the predicted $\hat{\boldsymbol{\gamma}}$ from the body translation estimator to crop and translate the local scene point cloud to encode the local scene feature, and feed $\hat{\boldsymbol{\gamma}}$ as the input to the diffusion model. The body pose $\boldsymbol{\theta}$ is transformed from the 6D representation to the rotation matrix to calculate $\mathcal{L}_{\text{simple}}$. The weights for $\mathcal{L}_{\text{simple}}, \mathcal{L}_{\text{3D}}, \mathcal{L}_{\text{2D}}, \mathcal{L}_{\boldsymbol{\beta}}, \mathcal{L}_{\text{coll}}, \mathcal{L}_{\text{orth}}$ are 0.001, 0.05, 0.01, 0.0005, 0.0002, and 0.1, respectively. The collision loss term $\mathcal{L}_{\text{coll}}$ is disabled for the first three epochs. The model is trained with a single TITAN RTX GPU of 24GB memory for approximately 18 epochs, with a batch size of 12, which takes around 24 hours.

Collusion score guided sampling. In Eq. (9) in the main paper, we set a as 2. For the last 10 diffusion denoising timesteps, we ignore the Σ_t and only scale $\nabla \mathcal{J}(\boldsymbol{\theta_t}), \Sigma_t)$ by a such that the collision score guidance does not diminish too much at the end of the sampling process.

Evaluation protocol. For the evaluation, the standard PA-alignment is obtained from the full body joints, however in the highly truncated case the diverse nature of invisible body parts could deviate from the ground truth and result in inaccurate PA-alignment, thus we perform PA-alignment with only visible body joints and report the PA-MPJPE metric.

D. More Experiments

D.1. Ablation Study on Model Architecture and Per-joint Conditioning

We also conduct experiments with the following two architectures as the diffusion denoiser D to verify the effectiveness of our proposed per-joint visibility conditioning strategy and the GCN architecture: 1) a single MLP network to predict the full body pose conditioned on $\boldsymbol{c} = (E_{\mathcal{I}}(\mathcal{I}), E_{\mathcal{S}}(\mathcal{S}, \hat{\boldsymbol{\gamma}}), \hat{\boldsymbol{\gamma}}, \mathcal{B}, \mathcal{K}, t)$, *i.e.* the image feature, the scene feature, the bounding box information, the camera intrinsics and the diffusion timestep, without the per-joint visibility mask, denoted as 'full-body MLP'; 2) using the same per-joint conditioning strategy as our proposed method, but with J MLP networks to predict the pose parameters for each body joint separately, where the MLPs share the same architecture but with different weights, denoted as 'per-joint MLP'. The results are shown in Tab. S1.

For the per-joint MLP model, there is a significant drop on sample diversity for invisible body parts, as such model architecture disables the classifier-free guidance. With the classifier-free guidance, the pose for invisible body parts sampled from the model excluding the image condition can be fused into the standard sampling results, thus improving the sample diversity for invisible body parts. With a separate MLP for each body joint independently from other joints, each MLP for the invisible joint already excludes the image condition, therefore no additional classifier-free guidance can be applied on top of that to further improve diversity. Different from the GCN architecture which considers the human kinematic tree, the per-joint MLP model neglects the inter-joint dependencies, which are crucial to model human poses and human-scene interactions. Due to this reason, the *min-of-n* MPJPE for invisible joints of the per-joint MLP model is also higher compared

Table S1: **Ablation study for model design choices.** All experiments are conducted without the scene collision score guidance $\mathcal{J}(\boldsymbol{\theta_t})$. The results are reported for $n = 5$.

Method	MPJPE ↓ *-vis*	*min-of-n* MPJPE ↓ *-invis*	coll ↓	contact ↑	std ↑ *-invis*	APD ↑ *-invis*
Ours	65.10	**107.59**	**0.00225**	**0.99**	**20.30**	**25.34**
Per-joint MLP	65.35	113.62	**0.00225**	**0.99**	16.50	20.51
Full-body MLP	65.65	111.67	0.00228	0.98	11.08	12.98

Table S2: **Evaluation on PROX with trained models on EgoBody.** All results are reported for $n = 5$.

Method	MPJPE ↓ *-vis*	*min-of-n* MPJPE ↓ *-invis*	coll ↓	std ↑ *-invis*	APD ↑ *-invis*
ProHMR-scene-*orig*	117.97	217.69	0.00907	**48.88**	**59.49**
ProHMR-scene-*weak-3D*	115.53	201.37	<u>0.00839</u>	25.69	31.68
ProHMR-scene-*strong-3D*	<u>112.37</u>	<u>199.33</u>	0.00887	20.63	25.26
Ours	**107.17**	**198.72**	**0.00739**	<u>30.01</u>	<u>37.35</u>

Table S3: **Comparison with Deterministic baseline.** Here the MPJPE and PA-MPJPE are calculated for the full body.

Metrics	METRO [8]	EFT [2]	SPIN [5]	SPIN-scene	Ours
MPJPE	98.5	102.1	106.5	91.6	**80.4**
PA-MPJPE	66.9	64.8	67.1	**64.5**	**64.5**

to the proposed GCN architecture, indicating that the generated body pose for visible body parts cannot cover the ground truth distribution well enough.

For the full-body MLP model, the full body pose is conditioned on the image and scene feature. Without the explicit per-joint visibility information, the network can hardly achieve the precise per-body-part control, therefore struggling to balance between the accuracy and diversity for different body parts (with the lowest diversity compared with other two models in Tab. S1). On the contrary, our proposed per-joint conditioning strategy can leverage the joint visibility to achieve both accuracy for visible joints and diversity for invisible joints, together with plausible human-scene interactions.

D.2. Evaluation on PROX Dataset

We also evaluate the trained model on PROX dataset, a third-person view dataset with monocular RGB frames for human-scene interaction scenarios. Due to the large domain gap of camera-body distances between EgoBody ($1\sim3m$) and PROX ($2\sim5m$), predicted body translations are not accurate on PROX for all methods since they are trained on EgoBody. To better analyze our model's capability on scene-conditioned 3D body estimation on PROX, we report the numbers with ground truth body translations for all methods. Our model outperforms the baselines (Tab. S2), with more accurate local pose, more plausible interactions with the scene and relatively high diversity for unseen body parts.

D.3. Comparison with Deterministic Methods

Here we show the full-body accuracy of deterministic baselines (Tab. S3): they lag behind our method by a considerable margin. Results show that even conditioning the network with scene features (SPIN-scene, by encoding scene point clouds with an additional scene feature on top of SPIN) cannot perform comparably with our method. This validates (1) the advantage of our probabilistic formulation with highly ambiguous poses; (2) feeding the network with scene features alone is insufficient, and our scene-guided diffusion sampling effectively addresses this.

E. More Qualitative Results

More qualitative examples and diverse sampling results of our proposed method are shown in Fig. S2 and Fig. S3, respectively. While obtaining accurate pose estimations aligning with the input image, our method also achieves impressive sample diversity for the unobserved body parts, with plausible human-scene interaction relationships.

Figure S2: **More qualitative examples.** For both left and right sides: (a) the input egocentric image; (b) the rendered body mesh overlay on the input image; (c) the rendered body mesh in the 3D scene.

F. Limitations and Future Work

Apart from the static human pose, human motions also play an important role in human behavior understanding from the egocentric view. One of the limitations of the proposed method is that it only allows per-frame human mesh recovery given a single frame input. Human motion estimation from an egocentric temporal sequence in 3D scenes would be an exciting future work and enable more real-life AR/VR applications. Besides, the current model relies on a two-stage pipeline, which estimates the global body translation and local body pose in separate stages. However, the global translation, local body pose and body shape are coupled together, and equally important for learning the interactions between the human body and the 3D environment. A unified end-to-end model to learn the body parameters altogether would be desired and potentially provide better reasoning about human-scene interaction relationships.

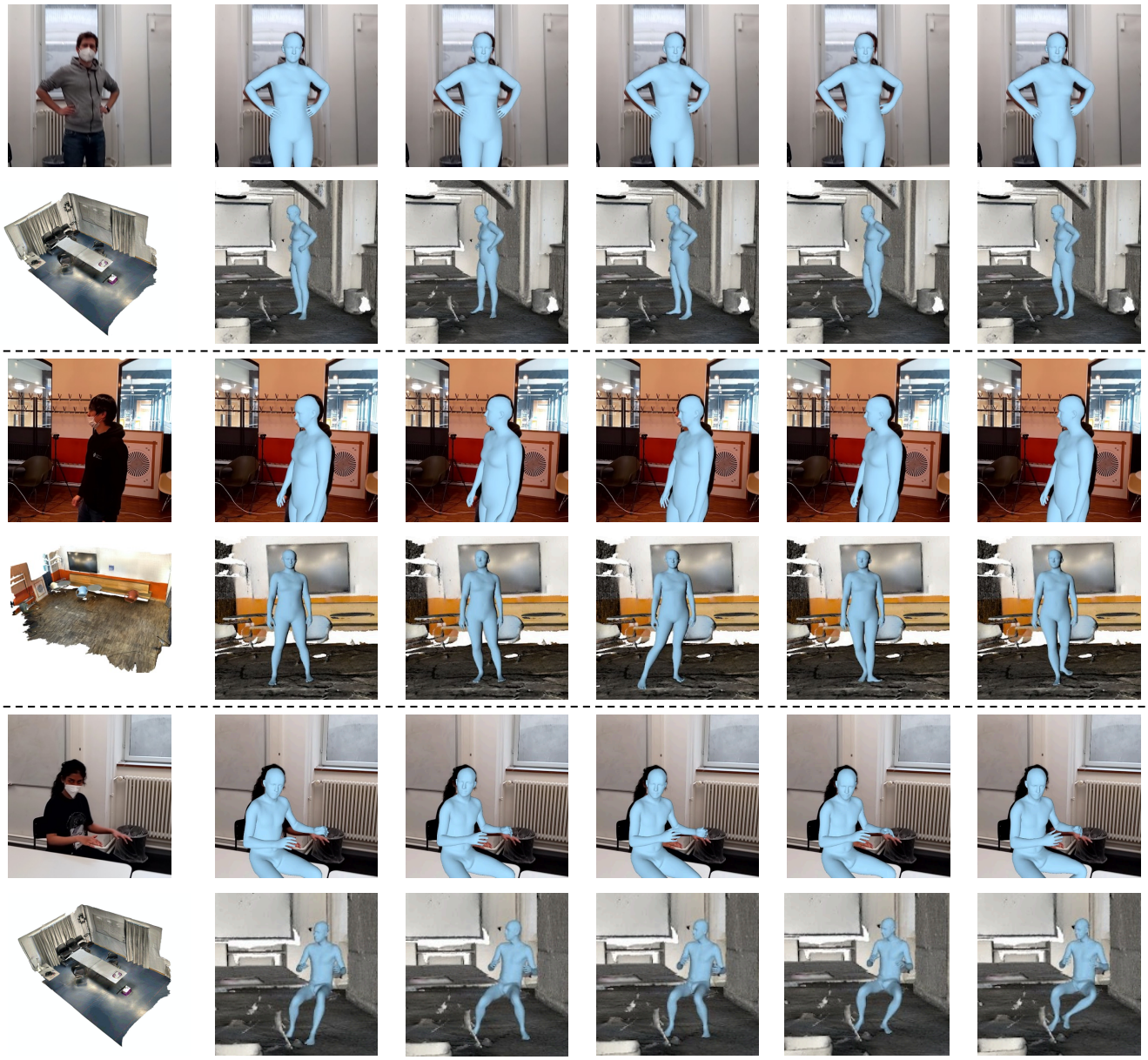

Figure S3: **More examples for diverse sampling.** Each row shows five different sample results given the input image and 3D environment (the first column).

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[2] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human pose fitting towards in-the-wild 3D human pose estimation. 2021. 3

[3] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 1

[4] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Proceedings International Conference on Computer Vision (ICCV)*, pages 11127–11137. IEEE, Oct. 2021. 1

[5] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2252–2261, 2019. 3

[6] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11605–11614, 2021. 1, 2

[7] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. CLIFF: Carrying location information in full frames into human pose and shape estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 590–606. Springer, 2022. 2

[8] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021. 3

[9] Yi Zhou, Connelly Barnes, Lu Jingwan, Yang Jimei, and Li Hao. On the continuity of rotation representations in neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[10] Zhiming Zou and Wei Tang. Modulated graph convolutional network for 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11477–11487, 2021. 1