EnvPoser: Environment-aware Realistic Human Motion Estimation from Sparse Observations with Uncertainty Modeling
Supplementary Material

Songpengcheng Xia[1] Yu Zhang [1] Zhuo Su[2] * Xiaozheng Zheng[2] Zheng Lv[2] Guidong Wang[2]
Yongjie Zhang[2] Qi Wu[1,2] Lei Chu [3] Ling Pei[1] †
[1]Shanghai Jiao Tong University [2]ByteDance [3]University of Southern California

This supplementary material provides additional information to complement our main paper. Sec. A details the network architecture and training procedures of our proposed method. Sec. B summarizes the datasets used in our study. Sec. C presents extended ablation studies, encompassing various environment points sampling strategy and further analysis of component design. Sec. D showcases additional qualitative comparisons between our approach and state-of-the-art methods. Finally, Sec. E summarizes our work, discusses its limitations, and outlines potential directions for future research.

A. Implementation Details

Network Details. Our model architecture consists of distinct sub-networks tailored to each stage. The uncertainty-aware initial motion estimation module processes sparse tracking signals and historical motion states, both pre-segmented into windows of length 40. These inputs are fed through two linear layers and positional encoding before entering a standard Transformer network with eight attention heads [6]. Subsequently, two MLP layers estimate the human motion and corresponding joint uncertainty. During the first stage of training, this module is trained in two steps: first, the motion reconstruction head is optimized, followed by training the uncertainty estimation head using Eq.(3). The hyperparameters λ_M and λ_δ are set to 1 and 0.001, respectively.

For environment-aware motion refinement module, we utilize PointNet++ [4] to extract features from the cropped environment point clouds. These features are integrated into the motion-environment attention network. In this stage, the initial motion estimates, combined with additional inputs such as head translations and head height, are concatenated into a 175-dimensional motion embedding. This embedding is projected into a 256-dimensional latent space through a linear layer. Spatial priors are computed by normalizing scene points to a human-centered coordinate system and combining distance-based salience with directional components. These priors are refined using a learnable network with two fully connected layers and ReLU activations. To integrate motion and environmental information, we employ a cross-attention mechanism that aligns scene features and spatial priors with the motion embedding. The resulting environment-refined motion representation $\boldsymbol{Z_{RM}}$ is obtained by passing the integrated features through an MLP with two fully connected layers.

Joint contact probabilities are predicted by projecting the concatenated environment-refined motion representation and sparse observations into a 256-dimensional embedding via a linear layer. This embedding is passed through a fully connected contact prediction head to estimate joint contacts.

Finally, the sparse observations and contact predictions are concatenated and processed through a decoder comprising two fully connected layers with ReLU activations, generating the final pose estimation. During the second training phase, we use Eq.(12) as the objective function. At the beginning of this phase, the parameters of the first module are fixed, and only the environment refinement module is trained. The hyper-parameters $\{\lambda_i\}_{i=1,\dots,7}$ are set to $\{2.0, 1.0, 0.75, 0.75, 0.75, 1.0, 0.1\}$.

	EgoBody		GIMO	
	MPJRE	MPJPE	MPJRE	MPJPE
(a) EnvPoser-cube	6.04	75.4	4.85	65.4
(b) EnvPoser-500	6.05	76.4	4.47	60.2
(c) EnvPoser-2000	**5.95**	76.0	4.45	59.8
EnvPoser-1000	6.00	**74.7**	**4.38**	**57.6**

Table 1. The effectiveness of environment point cloud sampling strategy.

*Project Lead
†Corresponding Author
This work was supported by the National Nature Science Foundation of China (NSFC) under Grant 62273229.

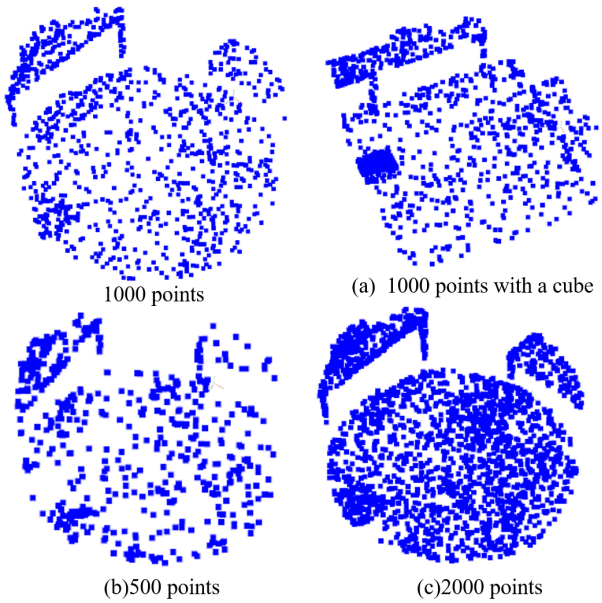

Figure 1. Environmental point cloud with different sampling strategies.

	EgoBody		GIMO	
	MPJRE	MPJPE	MPJRE	MPJPE
EnvPoser-w/o Contact	6.18	77.6	4.49	59.4
EnvPoser	**6.00**	**74.7**	**4.38**	**57.6**

Table 2. The effectiveness of contact estimation module.

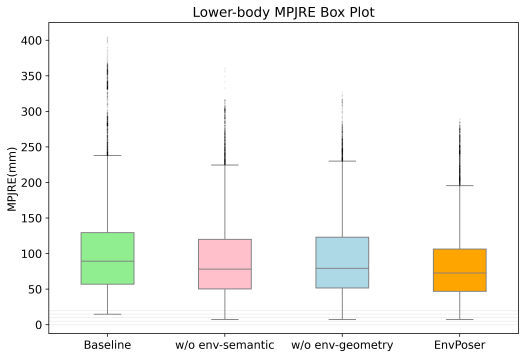

Figure 2. Qualitative results of lower-body MPJPE box plot for ablation study on GIMO dataset.

B. Datasets Details

To comprehensively evaluate the efficacy of our proposed method, we conduct experiments on two challenging public motion-scene interaction datasets: Egobody [7] and GIMO [9]. These datasets are carefully selected for their diverse and complex representations of human motion within interactive and immersive environments.

- **Egobody:** The Egobody dataset is a robust egocentric dataset designed to capture 3D human motion during social interactions in immersive virtual environments. It includes 125 sequences collected from 36 participants, with an equal distribution of 18 males and 18 females. These participants engage in a wide range of social activities across 15 distinct indoor scenes, making the dataset highly diverse. Egobody offers detailed annotations for various interaction scenarios, enabling precise evaluation of motion estimation techniques in dynamic and interactive settings. Following the official split in [7], the dataset is divided into 65 sequences for training and 43 sequences for testing. Its focus on immersive, egocentric perspectives provides valuable data for analyzing motion within constrained and socially active environments.
- **GIMO:** The GIMO dataset is notable for its multi-modal nature, offering a rich combination of body pose sequences, detailed environmental scans, and eye gaze data. Motion data were collected using a combination of HoloLens devices and IMU-based motion capture suits, providing precise motion trajectories and body pose sequences. Additionally, an iPhone 12 was used to scan the surrounding ambient scenes, resulting in high-quality environmental reconstructions. This multi-modal setup enables the dataset to capture the intricate relationships between human motion and environmental contexts. GIMO is particularly valuable for its emphasis on interactions within detailed 3D environments, offering a comprehensive perspective that bridges the gap between motion and scene understanding.

To ensure a diverse range of motion data, we also leverage the AMASS [3] dataset during training. Specifically, our model's uncertainty-aware initial human motion estimation module is first trained on AMASS to capture diverse motion patterns. Subsequently, the entire model is fine-tuned on motion-environment interaction datasets [7, 9], incorporating the environment-aware motion refinement module in the second stage.

C. Additional Analysis

In this section, we present more experimental results of our proposed method (EnvPoser).

Effectiveness of the Environment Point Clouds Sampling Strategy. EnvPoser processes the environmental point cloud through clipping and normalization in the Environmental Point Cloud Embedding module. Specifically, we first clip the environmental point clouds based on the human's global position, as illustrated in Fig. 1. To evaluate the impact of the clipping and sampling strategies on model performance, we experimented with various sampling methods. In this study, we replaced the circular sampling approach with a square sampling method, shown in Fig. 1(a). In Tab. 1, we could find that using circular sam-

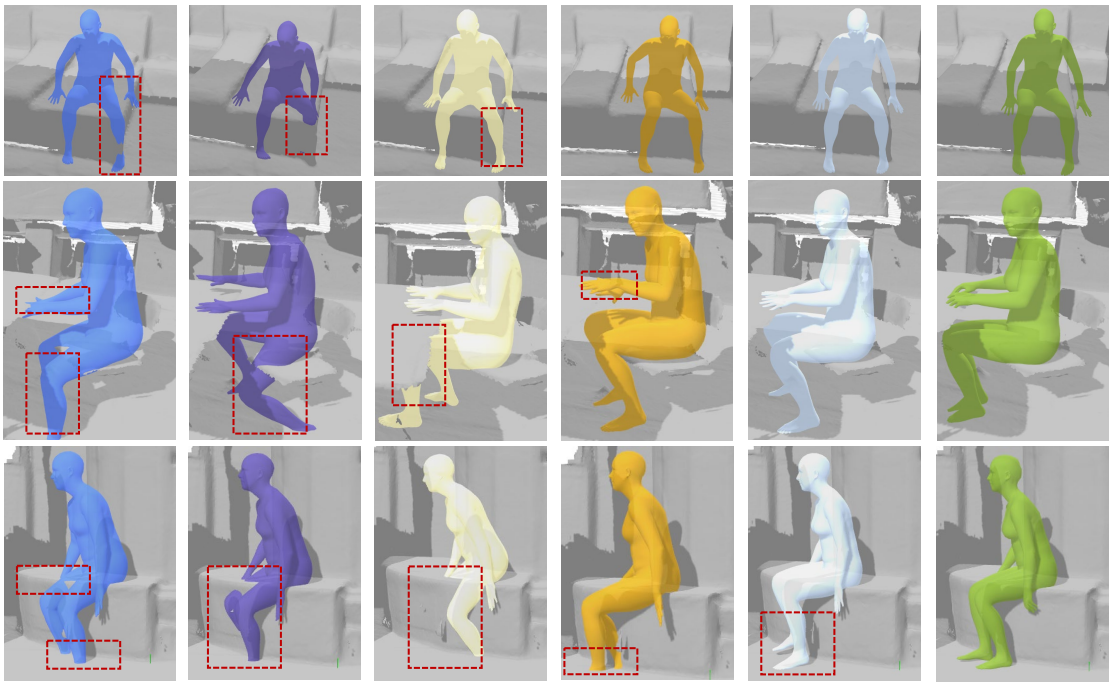

Figure 3. Visualization of various sitting motions from EgoBody and GIMO Datasets.

pling improves motion estimation performance to some extent.

Subsequently, based on circular clipping, we adjusted the number of sampled points to 500 and 2000 to further analyze the impact of point cloud sampling strategies on model performance. Sampling 500 environmental points as input leads to performance degradation due to the sparsity of the point cloud. Conversely, increasing the number of sampled points to 2000 does not consistently improve performance. On the EgoBody dataset, EnvPoser-2000 achieves a marginal improvement of 0.05 in the MPJRE metric, but on the GIMO dataset, using more points results in a decline in performance. These results indicate that sampling 1000 environmental points within the selected area is sufficient to refine the initial motion estimates. Considering the trade-off between motion estimation accuracy and computational efficiency, this paper adopts circular sampling with 1000 environmental points as the optimal configuration.

Effectiveness of the Contact Estimation Module. With the environment-refined motion representation $\boldsymbol{Z}_{RM}$ obtained through environment-semantic attention, EnvPoser first estimates contact probabilities and subsequently regresses full-body motion. To validate the effectiveness of the contact probability estimation, we conducted an ablation experiment by removing this step and directly regressing full-body motion from the environment-refined motion representation $\boldsymbol{Z}_{RM}$. This variant, referred to as EnvPoser-w/o Contact, was evaluated on the EgoBody and GIMO datasets.

As shown in Tab. 2, omitting the contact probability estimation results in degraded motion reconstruction performance compared to EnvPoser. Specifically, the MPJPE metric declines by 3.7% on the EgoBody dataset and 3.0% on the GIMO dataset, highlighting the importance of contact probability estimation in enhancing reconstruction accuracy.

Additional Comparison on Environment Refinement Module. To further illustrate the effectiveness of our environment refinement module, we present box plots of lower-body position estimation errors (Lower-body MPJPE) on the Gimo dataset for EnvPoser and its three variants: ① Baseline, ② w/o env-semantic, and ③ w/o env-geometry, as defined in Sec.4.

As shown in Fig. 2, EnvPoser achieves the lowest maximum Lower-body MPJPE and exhibits significantly fewer outliers, demonstrating that the integration of semantic and geometric environmental constraints enables the most robust and accurate estimation across diverse motions. Comparing variants ② and ③ with ①, we observe that incorporating environmental information significantly improves lower-body motion estimation accuracy. The constraints provided by environmental information not only enhance overall accuracy but also effectively reduce outliers in lower-body estimation, underscoring the importance of leveraging both environmental semantic and geometric aspects for refining motion estimates.

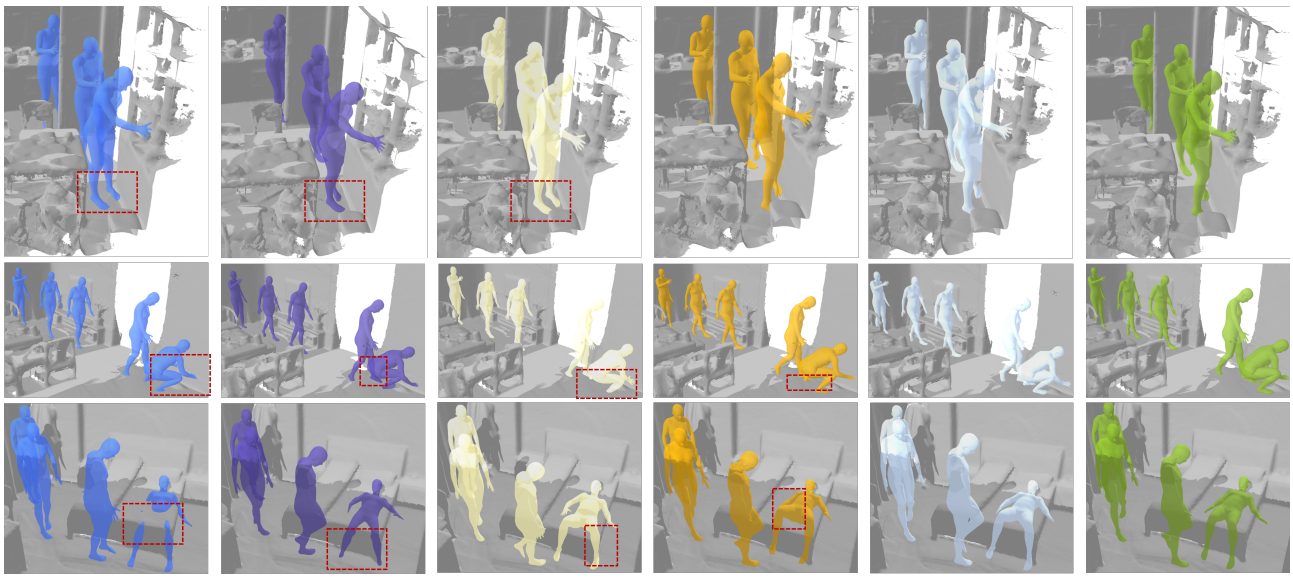

Figure 4. Visualization of full-body estimation on three test sequences from GIMO Datasets.

D. Additional Qualitative Results

In this section, We show more visualization results of our method compared to the state-of-the-art methods.

Additional Qualitative Comparison on Sitting Motion. The most common interaction between human motion and the environment involves sitting, such as sitting on chairs, sofas, or beds. Fig. 3 compares the performance of our method, EnvPoser, with other approaches in visualizing sitting motions across various scenarios. Notably, EnvPoser effectively adapts to different object shapes, generating realistic and plausible sitting poses. In contrast, methods that lack environmental information [8], such as AvatarPoser [2] and AGRoL [1], can estimate sitting motions but often produce diverse and inconsistent lower-body poses due to the absence of joint observations. This results in unrealistic and impractical sitting motion estimates that fail to align with actual human motions.

S2Fusion [5], which incorporates environmental information, shows notable improvements in sitting motion estimation compared to other competing methods. However, it performs slightly less accurately than EnvPoser in estimating upper-body interactions with objects during sitting motions. These findings highlight the effectiveness of EnvPoser, which leverages environmental information from both semantic and geometric perspectives to refine full-body motion estimation, ensuring more accurate and contextually appropriate results.

Additional Qualitative Comparison on GIMO Dataset. Fig. 4 presents motion estimation results on the GIMO dataset, showcasing three representative sequences with human motion visualized at key moments. Errors in comparison methods are highlighted with red bounding boxes for clarity. As illustrated in Fig. 4, EnvPoser exhibits robust performance, effectively handling challenging scenarios such as lying down and navigating narrow passages. These results further demonstrate the effectiveness and reliability of our proposed method.

Figure 5. Qualitative results on real data from VR device.

Additional Qualitative Comparison on Realistic Self-collected Data. We validated our model using data collected from real-world VR devices. The environment mesh within the motion range was pre-scanned, and human motion was estimated based on three 6DoF tracking signals captured by the VR devices. Fig. 5 showcases the full-body motion estimation results of EnvPoser, with a more comprehensive performance demonstration available in the supplementary video material.

E. Discussions and Future Works

Despite the effectiveness of our proposed framework, there are limitations that provide opportunities for future research and development.

Static Environment Assumptions: Our current model assumes a static environment, which does not account for dynamic interactions involving multiple users or moving objects. While this simplification facilitates efficient motion estimation, it limits the model's applicability in real-world scenarios where dynamic changes are common. For example, interactions in crowded spaces or with moving objects, such as passing a ball, are not effectively captured. To address this, future work could incorporate a third-person perspective or additional external sensors to estimate the movements of other users and objects. These additions could complement our existing environmental refinement module by providing more robust motion-context interaction capabilities.

Mesh Quality in Complex Environments: In real-time applications, the quality of pre-scanned environment meshes can vary significantly, especially in complex environments. Low-quality meshes may introduce inaccuracies in environmental constraints, impacting overall motion estimation performance. Exploring methods to enhance real-time mesh quality or mitigate the effects of noisy environment inputs will be essential.

Leveraging Raw Ego-centric Visual Data: Lastly, the model currently relies on pre-scanned point clouds for environmental context. Future work could extend this by incorporating raw visual data, such as images or video streams, to infer contact points and environmental semantics directly through 2D or 3D understanding. This approach could enable real-time processing of dynamic environments while reducing the dependency on pre-scanned data. For example, 2D semantic segmentation combined with depth estimation could enhance the system's ability to handle occlusions and complex interactions in cluttered scenes.

In summary, while our method achieves state-of-the-art performance in human motion estimation with sparse tracking signals, addressing these limitations through dynamic interactions, improved environmental modeling, and adaptive strategies will further enhance its robustness and applicability. These directions hold promise for broadening the utility of our approach in increasingly complex and interactive environments.

References

[1] Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, and Artsiom Sanakoyeu. Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 481–490, 2023. 4

[2] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *European conference on computer vision*, pages 443–460. Springer, 2022. 4

[3] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 2

[4] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 1

[5] Jiangnan Tang, Jingya Wang, Kaiyang Ji, Lan Xu, Jingyi Yu, and Ye Shi. A unified diffusion framework for scene-aware human motion estimation from sparse signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21251–21262, 2024. 4

[6] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 1

[7] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *European conference on computer vision*, pages 180–200. Springer, 2022. 2

[8] Xiaozheng Zheng, Zhuo Su, Chao Wen, Zhou Xue, and Xiaojie Jin. Realistic full-body tracking from sparse observations via joint-level modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14678–14688, 2023. 4

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