HMD-Poser: On-Device Real-time Human Motion Tracking from Scalable Sparse Observations
(Supplementary Materials)

A. Implementation Details

Feature embedding. This module comprises eight sub-modules to process eight components of the input data independently. Each sub-module consists of a few shallow multilayer perceptions (MLPs) to process rotation, angular velocity, position, linear velocity, and acceleration, respectively. Each MLP is a single Linear layer followed by $\mathrm{LeakyReLU}$. The outputs of the MLPs within each sub-module are concatenated to form a vector of size 256.

Lightweight TSFL. This module consists of two identical blocks, and each block has two sub-blocks: an LSTM-based block to learn temporal information and a Transformer-based block to encode spatial information. For each component of the input data, we use a separate unidirectional single-layer LSTM with a hidden size of 256 to encode the historical information. The LSTM networks are orthogonally initialized. To learn how different components are spatially correlated to each other, we adopt a 3-layer $\mathrm{Transformer}$ encoder with 8 attention heads and a feed-forward hidden size of 256 to process the outputs of LSTMs at each time step. The $\mathrm{Transformer}$ encoder itself can handle the scalable inputs via masks.

Regression heads. There are two regression heads that regress the local pose parameters and the shape parameters of SMPL [5] respectively. Both are MLP networks, and each MLP network consists of a Linear layer, $\mathrm{LeakyReLU}$ activation, and another Linear layer. We represent the joint orientations by the 6D reparametrization due to its simplicity and continuity [10]. Therefore, the output feature dimension of the pose regression head is $22 \times 6 = 132$. The output feature dimension of the shape regression head is set to 16. The intermediate dimension between the two Linear layers is set to 256.

B. Additional Quantitative Comparisons

As mentioned in the main paper, we consider three input scenarios in this paper, including (a) HMD, (b) HMD+2IMUs, and (c) HMD+3IMUs. In this section, we conduct extra experiments on each separate scenario. Meanwhile, to better demonstrate the superiority of our method, we also train a variant of the existing approaches, denoted as † in the following tables, by adding a shape regression head to their original model and introducing the joint position loss in the model training. All the experiments are also conducted on the AMASS dataset [6] with two different protocols.

B.1. HMD Scenario

In this scenario, we can fairly compare our method with state-of-the-art human motion tracking methods in the HMD setting, such as AvatarPoser [3], AGRoL [2], and AvatarJLM [9]. We re-train these methods with their public source code and the ground truth body shape parameters. As shown in Tab. A, HMD-Poser is not only more accurate (lower *MPJRE* and *MPJPE*) but also generates smoother human motions (lower *Jitter*) than all previous methods (without †) on both protocol1 and protocol2. It validates that HMD-Poser achieves a new state-of-the-art on the AMASS dataset. By comparing HMD-Poser with the variants of previous methods, HMD-Poser has similar tracking accuracy to AvatarJLM† [9] on protocol2 but is significantly better than other methods on all protocols. Meanwhile, HMD-Poser is significantly better than AvatarJLM† [9] in inference speed, i.e., 205.7Hz vs 1.9Hz, as shown in the main paper.

B.2. HMD+2IMUs Scenario

To the best of our knowledge, there is no available method for comparison in this scenario. We make a minor adjustment to the existing methods [2, 3, 9] in the HMD setting by adding the IMU tracking signals to their input data and extending the dimension of their feature embedding layer. Following [7, 8], we adopt synthesized IMU data on the AMASS dataset. Detailed results are presented in Tab. B. Comparing the results in Tab. A and Tab. B, it can be concluded that the tracking accuracy, especially for the lower body, is significantly improved by adding the IMU signals from the lower legs. It demonstrates the effectiveness of our method by combining HMD with IMUs. On protocol1, HMD-Poser surpasses all existing methods including their variants in all metrics. On protocol2, HMD-Poser obtains

Dataset	Method	MPJRE↓	MPJPE↓	MPJVE↓	Jitter↓	H-PE↓	U-PE↓	L-PE↓	R-PE↓
Protocol 1	AvatarPoser [3]	2.94	5.84	26.60	13.97	4.58	3.24	9.59	5.05
	AvatarPoser† [3]	2.94	5.17	27.17	14.45	3.64	2.71	8.73	4.39
	AGRoL [2]	2.70	5.73	19.08	7.65	4.29	3.16	9.44	5.15
	AGRoL† [2]	3.32	6.58	23.81	11.45	4.32	3.38	11.20	5.77
	AvatarJLM [9]	2.81	5.03	20.91	6.94	2.01	3.00	7.96	4.58
	AvatarJLM† [9]	2.56	3.89	20.91	7.46	**1.60**	2.01	6.62	3.32
	HMD-Poser(Ours)	**2.28**	**3.19**	**17.47**	**6.07**	1.65	**1.67**	**5.40**	**3.02**
Protocol 2	AvatarPoser [3]	4.68	6.62	33.16	10.79	3.93	2.97	11.89	5.30
	AvatarPoser† [3]	4.64	6.63	33.54	10.77	3.30	2.81	12.14	5.42
	AGRoL [2]	4.38	6.74	**24.14**	6.33	3.53	3.02	12.11	5.86
	AGRoL† [2]	4.82	8.17	33.82	15.75	5.69	3.75	14.56	6.63
	AvatarJLM [9]	4.45	5.96	27.50	6.91	2.30	2.97	10.28	5.22
	AvatarJLM† [9]	4.28	**5.43**	27.14	6.89	**1.88**	**2.32**	9.93	**4.67**
	HMD-Poser(Ours)	**4.27**	5.44	30.15	**5.62**	2.56	2.44	**9.77**	4.83

Table A. Evaluation results in the HMD scenario. We retrain existing approaches with their public source code and the ground truth body shape parameters. † denotes a variation of the existing models by adding a shape regression head to their original model and introducing the joint position loss in the model training. The best results are in **bold**.

Dataset	Method	MPJRE↓	MPJPE↓	MPJVE↓	Jitter↓	H-PE↓	U-PE↓	L-PE↓	R-PE↓
Protocol 1	AvatarPoser [3]	2.51	4.99	22.02	11.16	4.58	3.22	7.53	4.98
	AvatarPoser† [3]	2.52	4.24	23.17	12.09	3.69	2.68	6.49	4.23
	AGRoL [2]	2.25	4.81	15.13	8.44	4.25	3.09	7.28	4.95
	AGRoL† [2]	2.76	5.25	16.17	7.98	5.20	3.48	7.82	5.39
	AvatarJLM [9]	2.38	4.24	18.72	7.39	2.00	2.90	6.16	4.34
	AvatarJLM† [9]	2.12	2.95	18.78	7.53	1.48	1.89	4.48	3.06
	HMD-Poser(Ours)	**1.83**	**2.27**	**13.28**	**5.96**	**1.39**	**1.51**	**3.35**	**2.74**
Protocol 2	AvatarPoser [3]	3.87	4.58	25.98	9.75	3.74	2.88	7.03	4.99
	AvatarPoser† [3]	3.90	4.70	26.76	10.08	3.10	2.77	7.49	5.29
	AGRoL [2]	**3.64**	4.69	**17.22**	7.46	3.54	2.95	7.20	5.40
	AGRoL† [2]	4.00	5.63	23.37	14.53	3.95	3.32	8.98	6.78
	AvatarJLM [9]	3.89	4.49	22.64	6.34	2.21	2.89	6.41	4.84
	AvatarJLM† [9]	3.77	3.69	22.25	6.04	1.78	2.20	5.83	4.38
	HMD-Poser(Ours)	3.66	**3.68**	20.29	**6.22**	**1.65**	**2.14**	**5.92**	**4.51**

Table B. Evaluation results in the HMD+2IMUs scenario. Note that all previous methods are modified in this scenario by adding the IMU tracking signals to their input data and extending the dimension of their feature embedding layer.

the lowest position error and *Jitter* among all methods, but slightly higher *MPJRE* and *MPJVE* than AGRoL [2].

B.3. HMD+3IMUs Scenario

In this scenario, the input setting is closest to that of 6IMUs-based tracking methods. Therefore, we compare our HMD-Poser in the HMD+3IMUs scenario with state-of-the-art methods in this category, i.e., Transpose [7] and PIP [8]. For a fair comparison, we add the global positions of the headset and hand controllers to the input data of the baselines [7, 8]. The results are summarized in Tab. C. In this scenario, our HMD-Poser can surpass all previous methods in all metrics on both protocol1 and protocol2. Comparing the results in Tab. B and Tab. C, the tracking accuracy of HMD-Poser is further improved which validates the effectiveness of adding IMU to the pelvis.

C. Additional Ablation Studies

Effect of the model size. The number of blocks N in the lightweight TSFL network is a key hyper-parameter in our HMD-Poser. As shown in Tab. D, we see a clear downward tendency for both *MPJPE* and *Jitter* when increasing N from 1 to 2. However, this tendency becomes less pro-

Dataset	Method	MPJRE↓	MPJPE↓	MPJVE↓	Jitter↓	H-PE↓	U-PE↓	L-PE↓	R-PE↓
Protocol 1	Transpose [7]	3.05	4.57	22.41	7.98	3.83	3.05	6.76	4.62
	TransPose† [7]	3.02	3.99	23.32	8.65	3.58	2.72	5.82	4.23
	PIP [8]	2.45	4.54	19.02	8.13	4.54	3.15	6.53	4.54
	PIP† [8]	2.31	2.84	17.43	6.99	3.00	2.16	3.82	2.86
	HMD-Poser(Ours)	**1.73**	**1.89**	**11.03**	**5.35**	**1.27**	**1.46**	**2.46**	**2.37**
Protocol 2	Transpose [7]	4.31	5.29	28.18	5.16	7.38	3.86	7.36	4.80
	Transpose† [7]	3.94	4.73	29.11	6.02	5.60	3.42	6.61	4.57
	PIP [8]	3.61	4.16	22.22	6.89	4.28	2.97	5.89	4.30
	PIP† [8]	3.80	4.21	26.55	7.54	4.97	3.04	5.90	4.28
	HMD-Poser(Ours)	**3.49**	**3.13**	**16.17**	**4.93**	**1.81**	**2.17**	**4.51**	**3.88**

Table C. Evaluation results in the HMD+3IMUs scenario. For a fair comparison, we also add head and hand positions to the input data of all baseline methods.

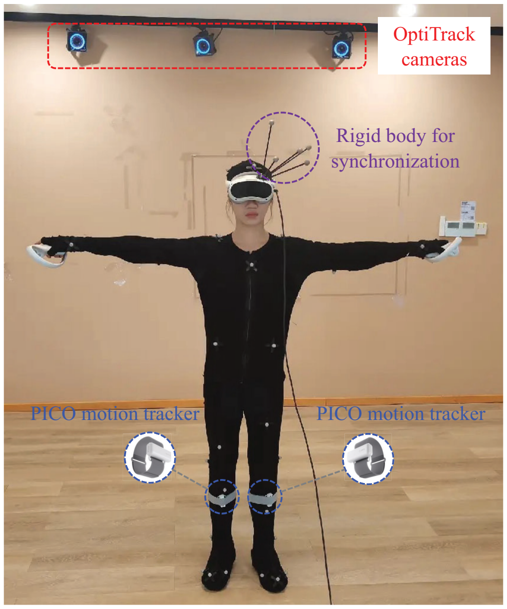

Figure A. Setup for real-data collection with HMD + 2IMUs.

nounced as N continues to increase. Hence, we use $N = 2$ in our final configuration which could obtain satisfactory results in both tracking accuracy and inference speed.

Effect of each loss term. As shown in the main paper, HMD-Poser is trained with five different loss terms. Among these loss terms, $\mathcal{L}_{ori}$, $\mathcal{L}_{lrot}$ and $\mathcal{L}_{joint}$ are essential terms for model training. We illustrate the contributions of the left two loss terms, i.e., $\mathcal{L}_{grot}$ and $\mathcal{L}_{smooth}$, in a leave-one-term-out manner. As shown in Tab. E, the smooth loss $\mathcal{L}_{smooth}$ has a positive impact on reducing the *MPJVE* and *Jitter* as expected. The global pose loss $\mathcal{L}_{grot}$ reduces the *MPJPE* and *H-PE*, which may be attributed to its role in reducing the accumulating error of pose estimation along the kinematic chain.

D. Real-Captured Data

To investigate the model's performance gap between synthetic data and real-captured sensor data and evaluate our HMD-Poser's performance running on HMDs, we built an additional dataset of real-captured data with HMD+2IMUs. As shown in Fig. A, each subject wears PICO 4 (including HMD and two hand controllers) and 2 PICO motion trackers on his/her lower legs and dances freely with music. Meanwhile, we use a synchronized marker-based motion capture system, OptiTrack [1], to track body markers and attain ground-truth SMPL parameters using Mosh++ [4]. A total of 74 free-dancing motions from 8 subjects are recorded. The duration of each motion sequence is set to 120 seconds. We will release this dataset for research soon.

Calibration. Since the raw IMU measurements are in the sensor-local coordinate system, we need to transform the raw IMU data into the same coordinate frame, which is referred to as calibration. Although we have specified the rough wear positions of IMU sensors, i.e., the pelvis, and the left and right lower legs, there could still be differences in the precise wear position and orientation of each subject. Our calibration method could automatically compute the transition matrices for each IMU sensor by requiring the subject to perform three specified actions: (1) stand straight for more than 5 seconds, (2) bend the knees forward and hold for 5 seconds, (3) lift the left and right legs in sequence.

Synchronization. Since we jointly capture ground-truth motions and the sensor data of HMD and IMUs with separate devices, our records must be accurately synchronized in the absence of a genlock signal. To this end, we add a rigid body on top of the HMD headset, as shown in Fig. A. Subjects are asked to perform simple control movements at the beginning of each capture motion, consisting of turning their heads clockwise and counterclockwise, nodding, and shaking their heads. This enables matching the orientations measured by IMUs on the HMD device with the rigid

Method	MPJRE↓	MPJPE↓	MPJVE↓	Jitter↓	H-PE↓	U-PE↓	L-PE↓	R-PE↓
$N = 1$	2.40	3.31	22.85	11.77	1.74	1.75	5.58	3.09
$N = 2$	**2.28**	3.19	17.47	6.07	1.65	1.67	5.40	**3.02**
$N = 3$	**2.28**	**3.18**	**16.97**	**5.20**	**1.60**	**1.66**	**5.38**	3.06

Table D. Evaluating the effect of the number of blocks N in the lightweight TSFL network.

Method	MPJRE↓	MPJPE↓	MPJVE↓	Jitter↓	H-PE↓	U-PE↓	L-PE↓	R-PE↓
w/o $\mathcal{L}_{grot}$	2.29	3.25	17.76	**6.06**	1.86	1.74	5.41	3.08
w/o $\mathcal{L}_{smooth}$	**2.27**	**3.16**	17.89	6.71	1.69	1.68	**5.30**	**3.00**
with all loss terms	2.28	3.19	**17.47**	6.07	**1.65**	**1.67**	5.40	3.02

Table E. Evaluating the effect of each loss term.

body orientations measured by OptiTrack. The frame rates of OptiTrack (120Hz) and IMUs (500Hz) are sufficiently high and the synchronization error is negligible.

Downsampling. Our HMD-Poser can reach a frequency of 90.0Hz on PICO 4 HMD. To align with the setup in training, we set the *FPS* of our HMD-Poser to a fixed frequency of 60Hz on HMD devices. Hence, we also downsample the ground-truth motions from 120Hz to 60Hz.

References

[1] Optitrack motion systems. https://optitrack.com/. 3

[2] 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. 1, 2

[3] 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 *Proceedings of the European Conference on Computer Vision*, pages 443–460, 2022. 1, 2

[4] Matthew Loper, Naureen Mahmood, and Michael J Black. Mosh: motion and shape capture from sparse markers. *ACM Transactions on Graphics.*, 33(6):220–1, 2014. 3

[5] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6): 1–16, 2015. 1

[6] 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. 1

[7] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics*, 40(4):1–13, 2021. 1, 2, 3

[8] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13167–13178, 2022. 1, 2, 3

[9] 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. 1, 2

[10] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 1