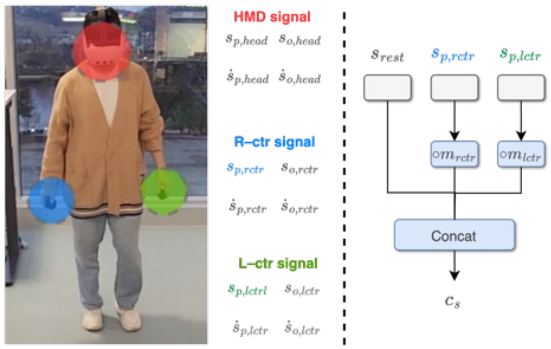

Figure 1. Illustration of controllers' positional signal masking.

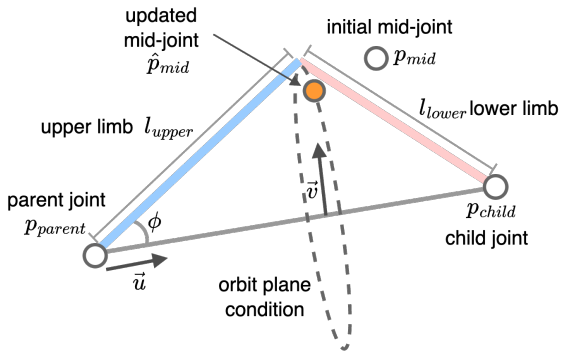

Figure 2. Illustration of limb length condition. See text.

As promised in our main paper, this supplemental document provides detailed explanation of our input data processing and evaluation metrics. In addition, we demonstrate more qualitative results here.

1. Input signal processing

Head Mounted Displays (HMD) estimate their 3D position and orientation using a combination of inside-out cameras with SLAM algorithms and inertial measurement unit (IMU) sensors for precise motion tracking. Conversely, most hand controllers lack inside-out cameras, requiring their position to be tracked by the HMD, while their orientation is estimated using onboard IMU sensors. This limitation makes the hand controllers' positional signal unreliable when the controllers are outside the HMD's field of view or occluded, as they can no longer be accurately tracked by HMD. As a result, the estimation of 3D human motion is impacted, since existing models [1–3, 5, 6, 13] inherently learn to predict users' hand positions based on the controllers' positional signals, making their predictions vulnerable to tracking errors.

To enhance robustness against this issue, we incorporate random masking of the controllers' positions during training. Figure 1 illustrates our approach to input signal processing. The input signals include the 3D position s_p, 3D orientation s_o, linear velocity $\dot{s}_p$, and angular velocity $\dot{s}_o$ of each device as input signal. We divide these signals into three components: (1) the position of the right controller $s_{p,rctr}$, (2) the position of the left controller $s_{p,lctr}$, and (3) all remaining signals s_{rest}. Each component is independently projected into an embedding space, where we apply masking to the position embeddings of the left and right hand controllers using the tracking mask m:

$$m = \begin{cases} 1 & \text{if controller is tracked,} \\ 0 & \text{otherwise.} \end{cases}$$

To generate synthetic masking during training, we applied the following approach: with a 10% probability, both controllers were masked out across all frames. For 60% of cases, no masking was applied to either controller. In the remaining 30%, each controller was individually masked for randomly selected consecutive frames, where the length of the masked frames followed a Gaussian distribution with a standard deviation of 10.

When testing our system on real data captured by the Meta Quest 3, we directly utilized the position flag provided by the OpenXR API. Empirical evaluations showed that training EgoMDM with synthetic masking significantly enhances the system's robustness to tracking errors in the controllers.

2. Analytical limb IK

Following HybrIK's [7, 8] insight that 3D joint angles can be analytically computed from the body joints' 3D position and twist angle, we reconstruct joint angles from $\hat{P}_{global}$ and θ_{twist}. However, HybrIK solves the IK problem from the pelvis outward, leading to cumulative position errors along the limb joints due to slight mismatches between the predicted motion and decoding avatar's limb lengths. To address these errors, our analytical IK module first refines the positions of mid-joints (*i.e.*, knees and elbows) to match limb length conditions precisely. Consider a mid-joint p_{mid} with its corresponding child joint p_{child} and parent joint p_{parent} as well as bone lengths l_{lower} and l_{uper} connecting these three joints. For example, in case of the left knee as the mid-joint, p_c is the left ankle, p_{parent} is the left hip, l_{lower} and l_{upper} are the lengths of left shank and thigh. As introduced by MANIKIN [6], the set of mid-joint positions satisfying bone-length condition can be formulated as an orbit condition (see Fig 2):

$$\{\tilde{p}_{mid}\} = \{p_{parent} + l_{upper} \cdot (\vec{u}\cos\phi + \vec{v}\sin\phi)\},$$

$$s.t.\ \vec{u} = \frac{p_{child} - p_{parent}}{||p_{child} - p_{parent}||},\ \ \vec{v} \perp \vec{u} \ \text{ and } \ \|\vec{v}\| = 1.$$

Here, the angle ϕ can be computed as:

$$\phi = \arccos\left(\frac{l_{upper}^2 + ||p_{child} - p_{parent}||^2 - l_{lower}^2}{2\, l_{upper}\, ||p_{child} - p_{parent}||}\right).$$

Then, we can choose the refined mid-joint position $\hat{p}_{mid}$ among the orbit circle that is closest to the initial prediction p_{mid},

$$\hat{p}_{mid} = \arg\min_{\tilde{p}_{mid}} ||\tilde{p}_{mid} - p_{mid}||.$$

The final step is to compute the rotation of each limb segment R, following HybrIK's analytical solution. Let $\vec{l}$ and $\vec{l}_0$ be the 3D vectors of the current limb pose and the neutral pose, where $\vec{l}_0$ is the rotation axis of the twist angle. Then, the swing angle θ_{swing} that rotates around $\vec{w} = (\vec{l}_0 \times \vec{l})$ satisfies:

$$\cos\theta_{swing} = \frac{\vec{l}_0 \cdot \vec{l}}{||\vec{l}_0||\,||\vec{l}||}, \quad \sin\theta_{swing} = \frac{||\vec{l}_0 \times \vec{l}||}{||\vec{l}_0||\,||\vec{l}||}.$$

We then compute the limb rotation $R = R_{swing}R_{twist}$, where the rotation matrices, R_{swing} and R_{twist}, can be derived by the *Rodrigues formula*.

3. Evaluation metrics

We use the AMASS dataset [9], a large-scale human motion capture dataset, to train and evaluate EgoMDM. To thoroughly assess our model's performance and compare it with existing methods, we employ a variety of metrics that measure both tracking accuracy and the quality of the generated motion.

3.1. Tracking accuracy metrics.

To evaluate motion tracking accuracy, we adopt widely-used metrics that quantify the geometric similarity between predicted and ground-truth motion. These include the Mean Per Joint Position Error (*MPJPE*, in cm), which calculates the average Euclidean distance between predictions and ground-truth labels across all joints and time frames. Notably, no alignment was applied when computing *MPJPE*. Additionally, we decompose the position error (*PE*) into separate components: upper-body position error (*UPE*), lower-body position error (*LPE*), hand position error (*HPE*), and root position error (*RPE*). Following prior works [1, 2, 13], we use 14 joints—including spines, shoulders, elbows, wrists, neck, and head—for *UPE*, and 8 joints—including hips, knees, ankles, and toes—for *LPE*. We further use temporal coherence metrics: Mean Per Joint Velocity Error (*MPJVE*, in cm/s) and *Jitter* (in $10^2\ m/s^{-3}$). *MPJVE* measures the L_2 distance between the first-order time derivatives of the predicted and ground-truth joint positions, capturing the similarity of motion flows. *Jitter*, on the other hand, is computed as the L_2 norm of the third-order time derivatives of the predicted joint positions, indicating the smoothness of the predicted motion.

3.2. Motion quality metrics.

We use *Skate* metric (in cm) to quantify the average horizontal displacement of the foot while it is in contact with the ground. Ground contact labels are computed based on the ankle and toe velocities of the ground truth motion, using a threshold of $1cm$/frame. To evaluate the feasibility of the predicted motion relative to the ground plane, we use the *Ground* metric (in cm). *Ground* represents the average of floor penetration and floating, where both are calculated using the distance between the lowest joint of the predicted motion and the ground plane. Moreover, inspired by text-to-motion synthesis methods [4, 11, 12], we evaluate the synthesized motion using the Fréchet Inception Distance (*FID*) score and *Diversity* (in cm). *FID* compares the distribution of the latent spaces of predicted and ground-truth motion, using a pretrained motion encoder [4] to project 3D human motion into the latent space after aligning the motion framerate. *Diversity* measures the variation in lower-body joint positions between motion samples generated from the same input signal. We compare *Diversity* exclusively with generative models [2, 3], randomly sampling 16 initial noise vectors to produce different motion samples.

	Protocol 2				
# Sampling Steps	MPJPE	MPJVE	Jitter	Skate	Ground
2	<u>4.92</u>	21.30	**1.49**	0.21	<u>1.27</u>
5	**4.89**	<u>20.74</u>	<u>1.51</u>	0.19	**1.27**
10	4.95	**20.58**	1.55	0.19	1.28
50	5.15	20.97	1.56	<u>0.17</u>	1.33
100	5.24	21.25	1.57	**0.17**	1.33

Table 1. Ablation experiments of the number of DDIM [10] sampling steps during inference. The best and second-best results are in **bold** and <u>underline</u>.

4. Number of sampling steps during inference

We conducted a quantitative comparison of different sampling steps to determine an optimal balance between motion tracking accuracy (MPJPE, MPJVE, Jitter) and physical feasibility (Skate, Ground). Table 1 suggests that 5 DDIM [10] sampling steps are provide an optimal balance, achieving strong performance across most evaluation metrics while maintaining real-time inference speed.

We conducted quantitative comparison of EgoMDM by changing the number of diffusion sampling steps during inference (Table 1). While EgoMDM was trained with 1,000 sampling steps, we take DDIM [10] sampling strategy with a subset of diffusion steps during inference. We found that 5 DDIM sampling steps are the optimal design as it provides not only the best tracking accuracy (*MPJPE*) but also ended up with physically plausible and smooth motion generation.

5. Qualitative results

We present more visual comparison of EgoMDM with state-of-the-art model, HMD-Poser [1] in Figure 3. We compute

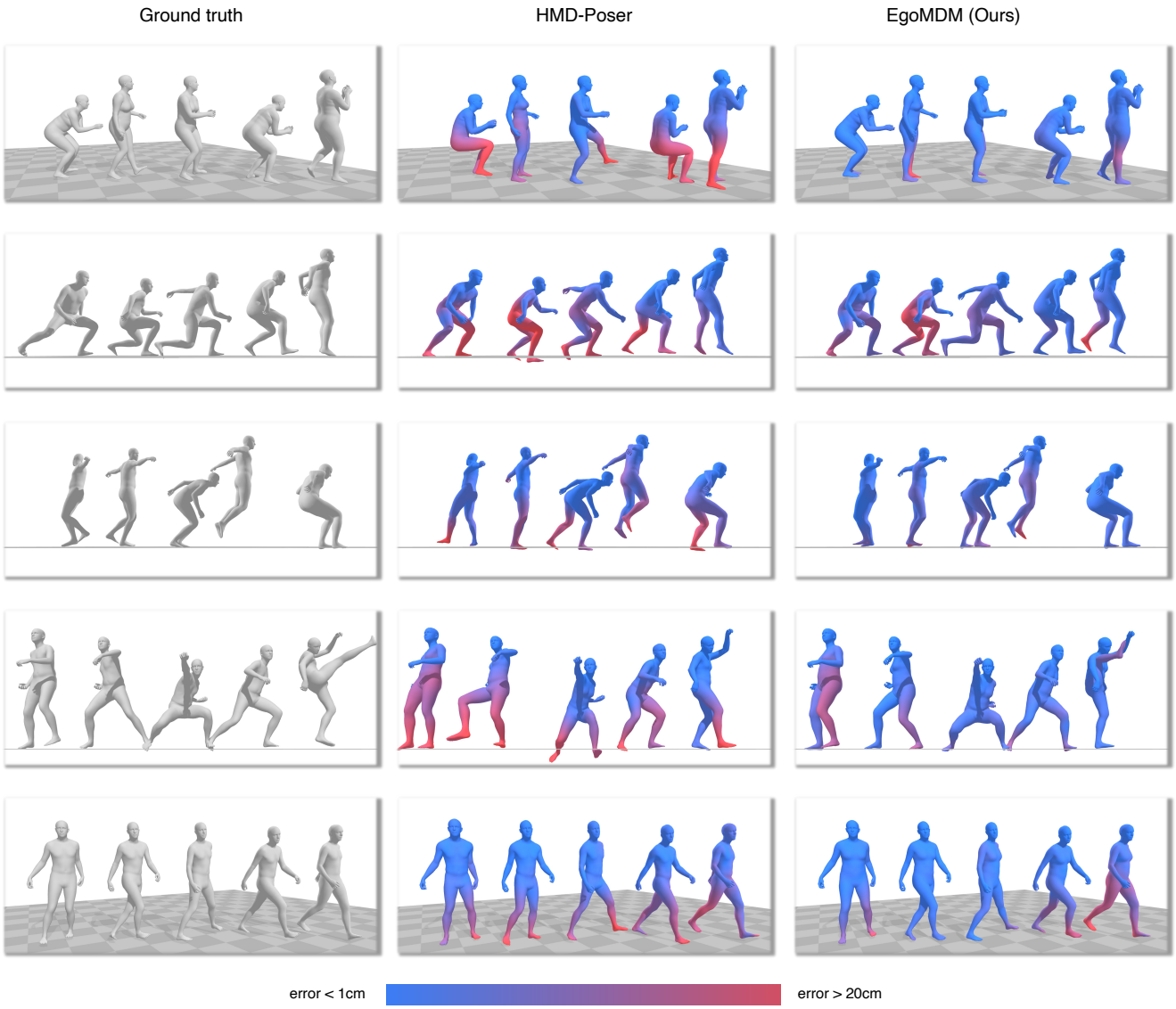

Figure 3. Additional qualitative comparison with state-of-the-art method [1].

vertex-to-vertex error and colored each vertex accordingly. The proposed method shows better geometric similarity with the ground truth motion, while exhibiting less ground penentration or floating compared to HMD-Poser.

6. Failure cases

EgoMDM synthesizes full-body human motion from sparse egocentric inputs (3-point signals), which inherently faces challenges in capturing accurate lower-body motion due to limited positional input. As illustrated in Figure 4, while upper-body joints align closely with ground truth, lower-body joints exhibit noticeable deviations. These deviations arise from the intrinsic ambiguity of reconstructing lower-body poses using limited tracking signals. Nevertheless, the predicted lower-body movements remain plausible and physically realistic, demonstrating the robustness of our approach in maintaining coherent motion predictions despite these limitations.

References

[1] Peng Dai, Yang Zhang, Tao Liu, Zhen Fan, Tianyuan Du, Zhuo Su, Xiaozheng Zheng, and Zeming Li. Hmd-poser: On-device real-time human motion tracking from scalable sparse observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1, 2, 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 *CVPR*, 2023. 2

[3] Han Feng, Wenchao Ma, Quankai Gao, Xianwei Zheng, Nan Xue, and Huijuan Xu. Stratified avatar generation from sparse observations. In *CVPR*, 2024. 1, 2

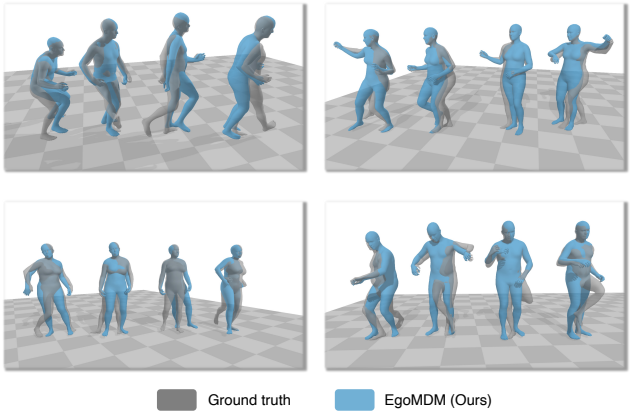

Figure 4. Visualization of inherent limitations in 3-point-based motion tracking. While our method accurately tracks upper-body movements, the predicted lower-body motion (blue) diverges from the ground truth (gray) due to limited inputs, yet maintains a realistic and physically plausible motion estimation.

[4] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161, 2022. 2

[5] 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 European Conference on Computer Vision*. Springer, 2022. 1

[6] Jiaxi Jiang, Paul Streli, Xuejing Luo, Christoph Gebhardt, and Christian Holz. Manikin biomechanically accurate neural inverse kinematics for human motion estimation. In *ECCV*, 2024. 1

[7] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3383–3393, 2021. 1

[8] Jiefeng Li, Siyuan Bian, Chao Xu, Zhicun Chen, Lixin Yang, and Cewu Lu. Hybrik-x: Hybrid analytical-neural inverse kinematics for whole-body mesh recovery. *arXiv preprint arXiv:2304.05690*, 2023. 1

[9] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, 2019. 2

[10] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 2

[11] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[12] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[13] 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 (ICCV)*, 2023. 1, 2