Realistic Full-Body Tracking from Sparse Observations via Joint-Level Modeling Supplementary Materials

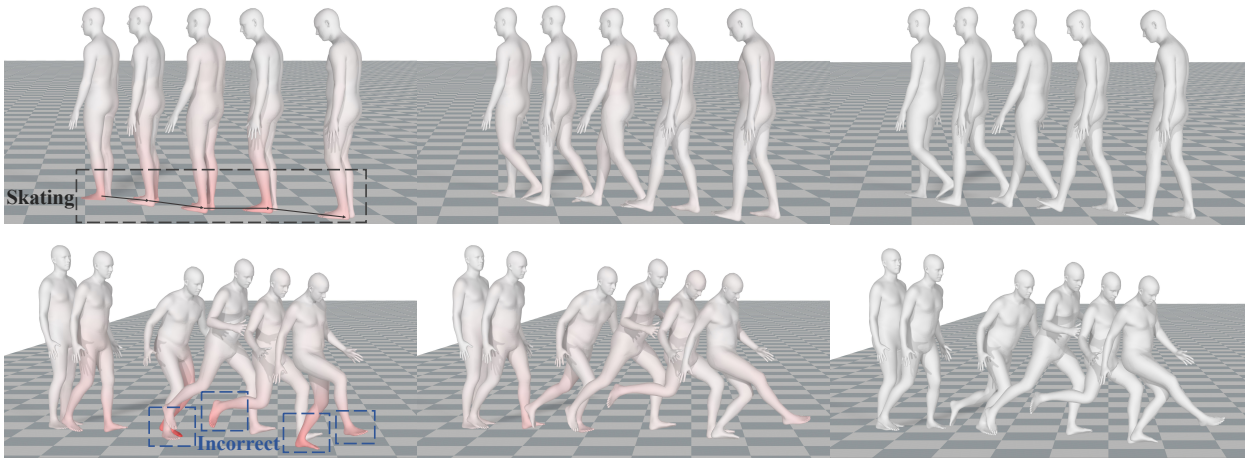

Figure A. Visual comparisons between various methods and ground truth.

A. Qualitative Comparisons

A.1. Error Color Coding

For visualization, we use error color coding to show the difference between our reconstructed motion and GT. The coded RGB value of a vertex is calculated as $[r, g, b] = (1 - e) \times [204, 204, 204] + e \times [255, 0, 0]$, where e is a vertex's error (m) clipped to 0$\sim$1.

A.2. Comparison with GT

To go beyond error color-coding in showcasing distinctions among various methods and Ground Truth (GT), we have also integrated GT into Fig. A. The results illustrate that our model can capture realistic motions that are close to real motions. For more results, please refer to the supplementary video.

A.3. Video Result

The most effective way to qualitatively showcase motion tracking is through video results. Therefore, in addition to the highlighted figure results in the main manuscript, we also provide sequential qualitative comparison results for our proposed approach with AvatarPoser [3], both in the AMASS test set and real application scenarios. Please kindly refer to the accompanying supplementary video.

A.4. User study.

Based on the qualitative results, we also evaluate the subjective quality by conducting a user study. We randomly selected 25 participants from various schools and grades who were unfamiliar with the system. Participants rated the naturalness and realism of given motion sequences on a 5-level Likert scale. We computed mean opinion scores for each method, randomly selecting 12 motion samples in the test set and shuffling the list. Our method achieved 3.69 scores while AvatarPoser gained 1.98 scores only.

B. Real-Captured Data

To evaluate the model performance in real scenarios, we capture several sequences of real data with the corresponding ground truth. The head and hands tracking signals are captured from the PICO 4 VR device, including both HMD and two controllers. Besides, we also use a synchronized marker-based motion capture system, OptiTrack [2]. Ground-truth SMPL parameters were then obtained from the MoCap data using MoSh++ [4]. To alleviate jittering, we apply the temporal filter to the ground-truth sequences. For effective testing with real-world data, using the 6DoF (six degrees of freedom) inputs directly from the HMD and controllers might not be suitable. This is because practi-

cal wearing configurations can lead to gaps between the devices and the body joints. To fill the sim-to-real gap, we apply empirically derived rigid transformations to convert the devices' 6-DoF data to joint-based representations for evaluation. We will release our code with the above evaluation samples to facilitate future research in this field [1].

C. Ablation Study

Regarding the ablation study, we also conduct additional experiments on transformer design, more loss combinations, and shorter input sequence lengths.

C.1. Transformer Design

As shown in Tab. A, removing STB or TTB both leads to significantly worse performance, indicating the importance of modeling spatial and temporal correlation simultaneously.

C.2. More Loss Combinations

Our loss design substantially contributes to achieving accurate motion with temporal consistency. Furthermore, it confers benefits not only to our method but also to other related approaches, such as AvatarPoser [3], as indicated in the main manuscript. To better substantiate the efficacy of each individual loss term and assess the performance of different combinations of loss types, we conduct a comprehensive set of experiments. As detailed in the main manuscript, our total loss function consists of *hand alignment* loss, *motion* loss (velocity-short loss, velocity-long loss, and foot contact loss), and *physical* loss (penetration loss and foot height loss).

We first add each loss term to the basic loss one by one. The results and contribution of each individual loss term are presented at the top of Table B. As discussed in the main manuscript, the addition of hand alignment loss is crucial to render the entire framework end-to-end trainable and significantly enhances performance. Moreover, it better aligns the predicted motion with observed signals, resulting in more precise motion outputs. The widely adopted velocity-short loss ($L_v(1)$) remarkably improves motion-related metrics, such as *MPJVE* and *Jitter*, as well as the *Skate* metric. However, relying solely on $L_v(1)$ may not entirely eliminate accumulated velocity errors, necessitating the inclusion of additional velocity-long losses ($L_v(3)$ and $L_v(5)$). Our results indicate that velocity-long losses lead to further decreases in all error metrics, attesting to their effectiveness. Incorporating foot contact loss to constrain foot movement slightly enhances motion-related metrics. While the use of penetration loss can significantly reduce *Ground* errors, it can also result in performance degradation in other evaluation metrics. This is because relying solely on penetration loss induces the network to predict results above the ground to reduce the *Ground* error. Complementary foot-height loss can mitigate this issue, considerably reducing penetration errors while improving other evaluation metrics.

Subsequently, we explore the performance of various combinations of loss types, including *hand alignment*, *motion*, and *physical* losses. The bottom of Tab. B demonstrates the performance of all different combinations. Based on our experimental results, several conclusions can be drawn. Firstly, it is crucial to frame the task as a sequence-to-sequence problem and employ the motion loss function accordingly. Secondly, hand alignment loss is a complementary component that enhances the alignment of hands while simultaneously improving overall accuracy. Thirdly, the physical loss term is a potent constraint that must be applied judiciously, as it can enhance performance only when the system already attains a substantially high level of accuracy and smoothness in motion.

C.3. Shorter Input Sequence Length

Although we have shown the possibility of applying our method in real scenarios in the attached video, achieving real-time performance for the application on mobile head-mounted displays (HMDs) with limited computing power is still challenging and important. Therefore, migrating to the HMDs is one of our future directions.

For exploring the possibility of our method applying to mobile devices, we perform experiments following Protocol 1 to evaluate our performance with shorter sequence lengths. The model's ability to process shorter input sequences is crucial for two reasons. First, it enhances the model's efficiency, which is essential for practical applications. Second, the capacity to handle short sequences enables the model to leverage future information with an acceptable latency.

According to the findings presented in Tab. C, our model exhibits robustness across varying sequence lengths. Even when processing very short sequences (11), our model demonstrates superior performance compared to AvatarPoser [3]. These results suggest that our model effectively leverages temporal information.

Method	MPJRE	MPJPE	MPJVE	Jitter
Ours	5.86	6.60	23.57	4.10
w/o STB	6.03	7.19	24.34	4.21
w/o TTB	6.13	7.30	27.99	5.14

Table A. Performance comparisons between our proposed method with different transformer designs.

Method	MPJRE	MPJPE	MPJVE	Jitter	Ground	Skate	H-PE	U-PE	L-PE
Ours - Basic Loss	6.09	7.50	32.53	8.98	3.66	0.35	3.11	3.93	13.76
+ Hand Alignment	5.87	6.90	29.07	6.88	3.73	0.33	1.58	3.51	12.85
+ Velocity-Short	5.86	6.83	25.80	4.38	3.55	0.27	1.68	3.57	12.53
+ Velocity-Long	5.81	6.67	24.48	4.35	3.39	0.23	1.61	3.47	12.27
+ Foot Contact	5.81	6.74	24.25	4.22	3.22	0.22	1.60	3.56	12.32
+ Penetration	5.85	6.80	23.34	3.87	2.65	0.20	1.69	3.50	12.58
+ Foot Height	5.86	6.60	23.57	4.10	2.46	0.21	1.69	3.52	12.12
+ Hand	5.87	6.90	29.07	6.88	3.73	0.33	1.58	3.51	12.85
+ Motion	5.73	6.91	22.42	2.97	4.35	0.17	2.36	3.55	12.80
+ Physical	5.99	8.30	33.24	8.54	2.63	0.31	5.12	4.61	14.75
+ (Hand + Motion)	5.81	6.74	24.25	4.22	3.22	0.22	1.60	3.56	12.32
+ (Hand + Physical)	6.06	7.55	27.29	5.20	2.49	0.28	1.22	3.55	14.54
+ (Motion + Physical)	5.90	7.18	22.93	3.26	2.43	0.17	2.41	3.73	13.22
+ (Hand + Motion + Physical)	5.86	6.60	23.57	4.10	2.46	0.21	1.69	3.52	12.12

Table B. Performance comparisons between our proposed method with different loss functions. The different background color is used for indicating the category of the loss. Green denotes hand alignment loss; purple denotes motion loss; and blue denotes physical loss.

Method	Length	MPJRE	MPJPE	MPJVE	Jitter	Ground	Skate	H-PE	U-PE	L-PE
AvatarPoser [3]	41	3.21	4.18	29.40	-	-	-	-	-	-
Ours	11	3.19	3.76	24.67	11.39	3.37	0.20	1.31	1.84	7.13
Ours	21	3.05	3.52	21.69	9.17	3.31	0.15	1.25	1.73	6.65
Ours	41	2.90	3.35	20.79	8.39	3.30	0.13	1.24	1.72	6.20

Table C. Performance comparisons between our proposed method with different input sequence lengths.

References

[1] Avatarjlm project page. https://zxz267.github.io/AvatarJLM/. 2

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

[3] J. Jiang, P. Streli, H. Qiu, A. Fender, L. Laich, P. Snape, and C. Holz. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 443–460. Springer, 2022. 1, 2, 3

[4] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. 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