A Structure-aware and Motion-adaptive Framework for 3D Human Pose Estimation with Mamba
(Supplementary Material)

Ye Lu[1*] Jie Wang [2*] Jianjun Gao[1] Rui Gong[1] Chen Cai [1] Kim-Hui Yap[1]

[1] Nanyang Technological University [2] Beijing Institute of Technology

{lu0001ye@e.,gaoj0018@e.,gong0084@e.,e190210@e.,ekhyap@}ntu.edu.sg {jwang991020}@gmail.com

A. Additional Experiment Results

A.1. Additional Comparison on Human3.6M

We evaluate our model's effectiveness across different input sequence lengths. In Tab. 5, we present results using a 243-frame input. Despite the shorter sequence length compared to 351 frames, our model retains strong performance, demonstrating its robustness. Our approach surpasses the state-of-the-art PoseMamba-X [5] in direct comparison with SAMA-L, achieving a $0.4mm$ reduction in MPJPE with a sequence length of 243. Across model variants, our method consistently delivers high accuracy, reporting MPJPE values of $40.6mm$ and $37.7mm$ for SAMA-S and SAMA-B, respectively, compared to $41.8mm$ and $40.8mm$ for PoseMamba-S and PoseMamba-B. This results in reductions of $1.2mm$ and $3.1mm$ for SAMA-S and SAMA-B, respectively, with comparable model size and MACs. For SAMA-L, aligning the estimated poses yields a P-MPJPE of 31.3, reflecting an advanced performance level. Regardless of model size, our approach consistently outperforms PoseMamba. With ground truth 2D poses as input, SAMA-L further achieves an MPJPE of $11.9mm$, marking a substantial improvement over PoseMamba (11.9 *v.s.* 14.8).

A.2. Per-Action Performance Comparison

We present per-action pose estimation results in Tab. 6, with detected 2D poses as inputs. The experimental results show that our method outperforms previous models in most action categories. We attribute the superior performance of our model across different action types to the design of SSI and MSM. In various actions, the joint topology relationships differ. In SSI, the proposed learnable adjacency matrix dynamically captures these varying relationships. Additionally, the motion characteristics across different action types are distinct. Our MSM can dynamically recognize these differences and regulate the timescale in the SSM, thereby capturing more joint motion features.

Table 1. MPJPE comparison by varying number of SSI and MSM blocks and number of channels on Human3.6M with detected 2D inputs from SHnet. D: Number of channels.

K (Depth)	D	Param (M)	MACs/frame (M)	MPJPE
2	128	1.1 M	18 M	40.2
5	128	2.8 M	45 M	37.8
6	128	3.3 M	54 M	37.4
7	128	3.8 M	63 M	37.5
7	256	15.1 M	207 M	36.7
8	256	17.3 M	234 M	36.5
9	256	19.5 M	263 M	36.5

B. Additional ablation study

B.1. Hyperparameter Setting Analysis

The network has two key hyperparameters: the depth of our SAMA (K) and the model dimension (D). We organize the configurations into two groups, with each group evaluating one hyperparameter by varying its value while keeping the other fixed, as shown in Tab. 1. This allows us to assess the impact and selection of each hyperparameter configuration.

B.2. Effect of Position Embedding

Unlike previous methods, we remove spatial and temporal embeddings. As shown in Tab. 3, adding these embeddings does not improve accuracy. We attribute this to the strong sequence modeling ability of Mamba-based models, which inherently capture spatial and temporal positions. Extra embeddings may introduce redundancy and hinder learning.

B.3. Model Varients

We introduce three configurations for our SAMA, as detailed in Tab. 4. The SAMA-B serves as the base model, offering a balance between accuracy and computational efficiency. The remaining variants are named according to

Table 2. Comparable architecture varients. *PM* denotes PoseMamba. N: Number of layers. D: Dimension of model.

Method	T	N	D	Param (M)	Infer speed (samples/s)
PM-S	243	20	64	0.9	4667
PM-B	243	20	128	3.4	2805
PM-X	243	40	256	26.5	908
SAMA-S	243	8	128	1.3	10411
SAMA-B	243	24	128	3.3	3658
SAMA-L	243	32	256	17.3	1298

Table 3. Ablation study for spatial and temporal embeddings.

Spa. Embedding	Temp. Embedding	MPJPE↓
-	-	37.4
✓	-	37.4
✓	✓	37.5

Table 4. SAMA model variants. M denotes Number of layers. d means the dimension of model. Param represents the number of model parameters. MACs/frame represents multiply-accumulate operations per output frame.

Method	K (Depth)	*D*	Param(M)	MACs/frames (M)
SAMA-S	2	128	1.2	18
SAMA-B	6	128	3.3	54
SAmA-L	8	256	17.1	234

their parameters and computational demands. The choice of each variant depends on the specific requirements of the application, such as real-time performance or accuracy in estimations.

B.4. Comparable architecture varients.

The SAMA variants (S/B/L) correspond to the PoseMamba variants (S/B/X) in model scale. Detailed architecture layers, parameters, and inference speed are shown in Tab. 2.

C. Additional Visualization

C.1. Effect of Structure-aware State Integrator

As mentioned in the Method section, the SSD mixer family of Mamba-2 has been shown to be equivalent to sequentially-semi-separable matrices. The SSD can be formulated as:

$$y_t = \mathbf{F}x_t = \mathbf{P} \cdot (\mathbf{C}^T\mathbf{B})x, \tag{1}$$

where $\mathbf{P}_{ij}$ is defined as follows: $\mathbf{P}_{ij} = \overline{A}i + 1 \times \cdots \times \overline{A}j$ for $i > j$, $\mathbf{P}_{ij} = 1$ when $i = j$, and $\mathbf{P}_{ij} = 0$ for $i < j$. Consequently, the Mamba-2 network can be interpreted as a causal linear attention mechanism with a learnable causal mask. In Fig. 1, we visualize the matrix F from Eq. (1) . It

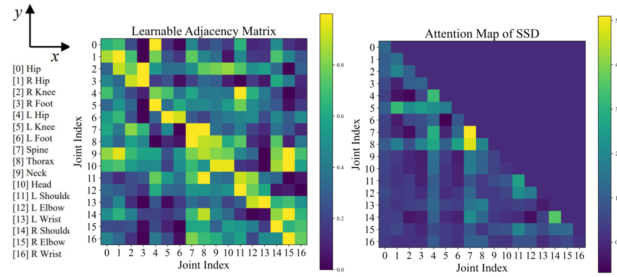

Figure 1. Visualization of SSM map among body joints and frames.

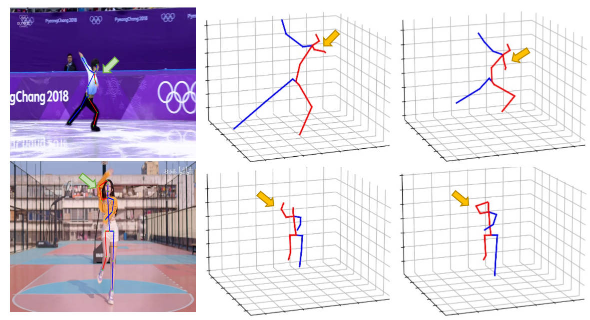

Figure 2. Qualitative comparisons of PoseMamba and our method in in-the-wild scenarios. We highlight the deviated 2D detection results with green arrows and corresponding 3D pose estimations with orange arrows.

reveals that the attention map of SSD forms a lower triangular matrix, meaning each joint can only be influenced by those with a smaller joint index. In contrast, our learnable adjacency matrix provides a global perspective, allowing all joints to exchange information while preserving the original joint topology effectively.

C.2. Additional visualization of estimated poses

Fig. 5 illustrates the 3D pose predictions of MotionBERT, PoseMamba, and our method, where ground truth poses are shown in blue and estimated poses in orange. Additional examples further support our findings in the main text: our approach achieves higher accuracy than PoseMamba and MotionBERT, especially in highly dynamic limb regions. This underscores the effectiveness of our SSI and MSM, which enhances joint connections and motion capture precision and improves overall performance.

C.3. Generalization to in-the-wild scenarios.

We add qualitative results under in-the-wild conditions. In Fig. 2, our method obtains more reliable 3D human pose, even in cases where the human actions are complex and rare.

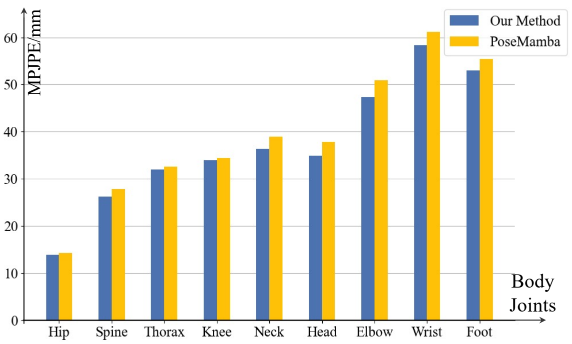

Figure 3. MPJPE of individual joints

C.4. Joints with high degrees of freedom.

We evaluate the per-joint MPJPE on Human3.6M (Fig. 3). Our method consistently outperforms PoseMamba, with particularly notable improvements on joints with high degrees of freedom, such like wrist, indicating our model effectively captures complex joint dynamics.

D. Additional Related Work

D.1. Mamba-based Models in Human-Centric Tasks

Gu et al. [4] first introduce the Linear State Space Layer (LSSL) to effectively manage long-range dependencies in extensive sequences. Motion Mamba [16] consists of two modules: Hierarchical Temporal Mamba (HTM), which enhances motion consistency across frames, and Bidirectional Spatial Mamba (BSM), which captures the bidirectional flow of channel-wise hidden information. Hamba [2] first incorporates graph learning and state space modeling for reconstructing a robust 3D hand mesh. It proposes a simple yet effective Graph-guided State Space (GSS) block to capture structured relations between hand joints.

D.2. 2D-to-3D Pose Lifting

Peng et al. [10] propose KTPFormer, incorporating Kinematics Prior Attention (KPA) and Trajectory Prior Attention (TPA) to leverage human anatomical structure and motion trajectory, enhancing global dependency learning in multi-head self-attention. PoseFormerV2 [17] makes slight modifications to PoseFormer, leveraging transformers to effectively aggregate temporal and frequency domain information, significantly improving computational speed while maintaining strong performance. Unlike prior works such as MotionAGFormer [9], POT [6], and KTPFormer [10] that use GCNs or complex attention for spatial modeling, we adopt a lightweight learnable matrix for joint feature aggregation in the state space. Temporally, rather than treating joints across frames uniformly, we leverage Mamba's timescale to capture joint-specific dynamics.

D.3. Diverse scanning method of Mamba

To achieve this goal, rather than adopting repeated scanning strategies as in prior works, we incorporate a structure-aware state transition into the original Mamba formulation. As shown in Fig. 4, we illustrate the different scan mechanisms used in previous methods, including PoseMamba and PoseMagic.

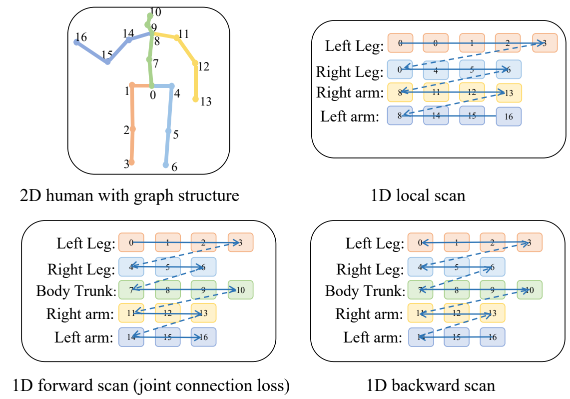

Figure 4. Diverse scanning method of Mamba

Table 5. Quantitative comparisons on Human3.6M. T: Number of input frames. CE: Estimating center frame only. MACs/frame: multiply-accumulate operations per output frame. P1: MPJPE (mm). P2: P-MPJPE (mm). P1†: P1 on 2D ground truth. (*) denotes using HRNet for 2D pose estimation. The best and second-best scores are in bold and underlined, respectively.

Method	T	CE	Param(M)	MACs(G)	MACs/frame(M)	P1↓ /P2↓	P1†↓
*MHFormer [CVPR2022] [7]	351	✓	30.9	7.0	20	43.0/34.4	30.5
MixSTE [CVPR2022] [14]	243	×	33.6	139.0	572	40.9/32.6	21.6
P-STMO [ECCV2022] [11]	243	✓	6.2	0.7	3	42.8/34.4	29.3
Stridedformer [TMM2022] [8]	351	✓	4.0	0.8	2	43.7/35.2	28.5
Einfalt *et al.* [WACV2023] [3]	351	✓	10.4	0.5	1	44.2/35.7	-
STCFormer [CVPR2023] [12]	243	×	4.7	19.6	80	41.0/32.0	21.3
STCFormer-L [CVPR2023] [12]	243	×	18.9	78.2	321	40.5/<u>31.8</u>	-
PoseFormerV2 [CVPR23] [17]	243	✓	14.4	4.8	20	45.2/35.6	-
GLA-GCN [ICCV2023] [13]	243	✓	1.3	1.5	6	44.4/34.8	21.0
MotionBERT [ICCV2023] [18]	243	×	42.3	174.8	719	39.2/32.9	17.8
HDFormer [IJCAI2023] [1]	96	×	3.7	0.6	6	42.6/33.1	21.6
MotionAGFormer-L [WACV2024] [9]	243	×	19.0	78.3	322	38.4/32.5	17.3
KTPFormer [CVPR2024] [10]	243	×	35.2	76.1	313	40.1/31.9	19.0
PoseMagic [Arxiv2024] [15]	243	×	14.4	20.29	84	37.5/-	-
PoseMamba-S [AAAI2025] [5]	243	×	0.9	3.6	15	41.8/35.0	20.0
PoseMamba-B [AAAI2025] [5]	243	×	3.4	13.9	57	40.8/34.3	16.8
PoseMamba-L [AAAI2025] [5]	243	×	6.7	27.9	115	38.1/32.5	15.6
PoseMamba-X [AAAI2025] [5]	243	×	26.5	109.9	452	37.1/31.5	14.8
SAMA-S (Ours)	243	×	1.1	3.9	16	40.6/34.0	20.2
SAMA-S (Ours)	351	×	1.1	6.3	18	40.2/33.8	19.5
SAMA-B (Ours)	243	×	3.3	11.7	48	37.7/32.0	13.6
SAMA-B (Ours)	351	×	3.3	18.9	54	37.4/31.7	12.4
SAMA-L (Ours)	243	×	17.3	53.2	219	<u>36.9/31.3</u>	<u>11.9</u>
SAMA-L (Ours)	351	×	17.3	82.1	234	**36.5/31.0**	**11.4**
vs. prev. SoTA	-	-	***↓11.2***		***↓218***	***↓0.6/↓0.5***	***↓3.4***

Table 6. Quantitative Per-action performance comparisons on Human3.6M using detected 2D pose as input. The best result is marked in blue in each column.

Protocol 1	Dir.	Disc.	Eat	Great	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
MixSTE [14]	36.7	39.0	36.5	39.4	40.2	44.9	39.8	36.9	47.9	54.8	39.6	37.8	39.3	29.7	30.6	39.8
MHFormer [7]	39.2	43.1	40.1	40.9	44.9	51.2	40.6	41.3	53.5	60.3	43.7	41.1	43.8	29.8	30.6	43.0
P-STMO [11]	38.4	42.1	39.8	40.2	45.2	48.9	40.4	38.3	53.8	57.3	43.9	41.6	42.2	29.3	29.3	42.1
STCFormer [12]	38.4	41.2	36.8	38.0	42.7	50.5	38.7	38.2	52.5	56.8	41.8	38.4	40.2	26.2	27.7	40.5
MotionBERT [18]	36.3	38.7	38.6	33.6	42.1	50.1	36.2	35.7	50.1	56.6	41.3	37.4	37.7	25.6	26.5	39.2
GLA-GCN [13]	41.3	44.3	40.8	41.8	45.9	54.1	42.1	41.5	57.8	62.9	45.0	42.8	45.9	29.4	29.9	44.4
KTPFormer [10]	37.9	39.8	35.9	37.6	42.5	48.2	38.6	39.0	51.4	55.9	41.6	39.0	40.0	27.0	27.4	40.1
SAMA-S (Ours) (T=351)	37.3	40.9	39.5	34.4	42.1	50.0	37.7	37.0	51.9	57.5	41.7	38.0	39.1	27.3	27.8	40.2
SAMA-B (Ours) (T=351)	34.9	38.9	35.5	32.2	39.9	47.4	35.9	35.0	47.1	52.6	38.9	36.0	36.9	25.2	25.5	37.4
SAMA-L (Ours) (T=351)	34.2	37.2	34.7	31.2	39.3	46.0	34.3	33.5	46.4	52.8	37.7	35.2	34.9	24.6	25.5	36.5

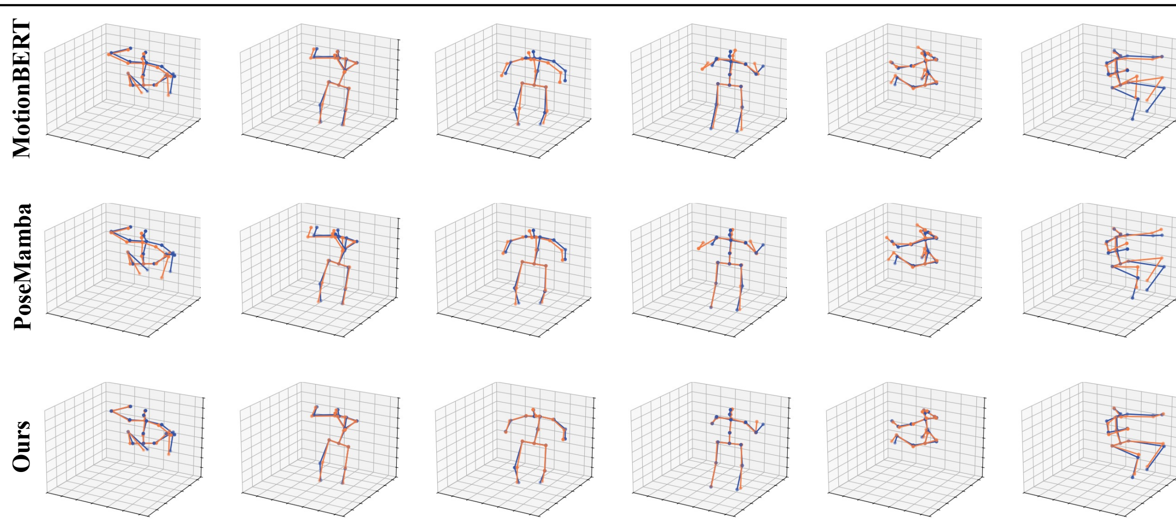

Figure 5. Additional visual comparable results of estimated 3D poses with MotionBERT and PoseMamba.

References

[1] Hanyuan Chen, Jun-Yan He, Wangmeng Xiang, Zhi-Qi Cheng, Wei Liu, Hanbing Liu, Bin Luo, Yifeng Geng, and Xuansong Xie. Hdformer: High-order directed transformer for 3d human pose estimation. arXiv preprint arXiv:2302.01825, 2023.

[2] Haoye Dong, Aviral Chharia, Wenbo Gou, Francisco Vicente Carrasco, and Fernando De la Torre. Hamba: Single-view 3d hand reconstruction with graph-guided bi-scanning mamba. In Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024, 2024.

[3] Moritz Einfalt, Katja Ludwig, and Rainer Lienhart. Uplift and upsample: Efficient 3d human pose estimation with uplifting transformers. In IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023, pages 2902–2912. IEEE, 2023.

[4] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state-space layers. CoRR.

[5] Yunlong Huang, Junshuo Liu, Ke Xian, and Robert Caiming Qiu. Posemamba: Monocular 3d human pose estimation with bidirectional global-local spatio-temporal state space model. arXiv preprint arXiv:2408.03540, 2024.

[6] Han Li, Bowen Shi, Wenrui Dai, Hongwei Zheng, Botao Wang, Yu Sun, Min Guo, Chenglin Li, Junni Zou, and Hongkai Xiong. Pose-oriented transformer with uncertainty-guided refinement for 2d-to-3d human pose estimation. In Proceedings of the AAAI conference on artificial intelligence, 2023.

[7] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, pages 13137–13146. IEEE, 2022.

[8] Wenhao Li, Hong Liu, Runwei Ding, Mengyuan Liu, Pichao Wang, and Wenming Yang. Exploiting temporal contexts with strided transformer for 3d human pose estimation. IEEE Trans. Multim., 25:1282–1293, 2023.

[9] Soroush Mehraban, Vida Adeli, and Babak Taati. Motionagformer: Enhancing 3d human pose estimation with a transformer-gcnformer network. In IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024, pages 6905–6915. IEEE, 2024.

[10] Jihua Peng, Yanghong Zhou, and P. Y. Mok. Ktpformer: Kinematics and trajectory prior knowledge-enhanced transformer for 3d human pose estimation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024, pages 1123–1132. IEEE, 2024.

[11] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. P-STMO: pre-trained spatial temporal many-to-one model for 3d human pose estimation. In Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part V, pages 461–478. Springer, 2022.

[12] Zhenhua Tang, Zhaofan Qiu, Yanbin Hao, Richang Hong, and Ting Yao. 3d human pose estimation with spatio-temporal criss-cross attention. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, pages 4790–4799. IEEE, 2023.

[13] Bruce X. B. Yu, Zhi Zhang, Yongxu Liu, Sheng-Hua Zhong, Yan Liu, and Chang Wen Chen. GLA-GCN: global-local adaptive graph convolutional network for 3d human pose estimation from monocular video. In IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023, pages 8784–8795. IEEE, 2023.

[14] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, pages 13222–13232. IEEE, 2022.

[15] Xinyi Zhang, Qiqi Bao, Qinpeng Cui, Wenming Yang, and Qingmin Liao. Pose magic: Efficient and temporally consistent human pose estimation with a hybrid mamba-gcn network. CoRR, abs/2408.02922, 2024.

[16] Zeyu Zhang, Akide Liu, Ian D. Reid, Richard I. Hartley, Bohan Zhuang, and Hao Tang. Motion mamba: Efficient and long sequence motion generation. In Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part I, pages 265–282. Springer, 2024.

[17] Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, pages 8877–8886. IEEE, 2023.

[18] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023, pages 15039–15053. IEEE, 2023.