KineST: A Kinematics-guided Spatiotemporal State Space Model for Human Motion Tracking from Sparse Signals

Shuting Zhao[1, 3*], **Zeyu Xiao**[2*], **Xinrong Chen**[1, 3†]
[1]College of Biomedical Engineering, Fudan University
[2]College of Intelligent Robotics and Advanced Manufacturing, Fudan University
[3]Shanghai Key Laboratory of Medical Image Computing and Computer Assisted Intervention
zhaoshuting@fudan.edu.cn, xiaozy23@m.fudan.edu.cn, chenxinrong@fudan.edu.cn

Abstract

Full-body motion tracking plays an essential role in AR/VR applications, bridging physical and virtual interactions. However, it is challenging to reconstruct realistic and diverse full-body poses based on sparse signals obtained by head-mounted displays, which are the main devices in AR/VR scenarios. Existing methods for pose reconstruction often incur high computational costs or rely on separately modeling spatial and temporal dependencies, making it difficult to balance accuracy, temporal coherence, and efficiency. To address this problem, we propose KineST, a novel kinematics-guided state space model, which effectively extracts spatiotemporal dependencies while integrating local and global pose perception. The innovation comes from two core ideas. Firstly, in order to better capture intricate joint relationships, the scanning strategy within the State Space Duality framework is reformulated into kinematics-guided bidirectional scanning, which embeds kinematic priors. Secondly, a mixed spatiotemporal representation learning approach is employed to tightly couple spatial and temporal contexts, balancing accuracy and smoothness. Additionally, a geometric angular velocity loss is introduced to impose physically meaningful constraints on rotational variations for further improving motion stability. Extensive experiments demonstrate that KineST has superior performance in both accuracy and temporal consistency within a lightweight framework.

Introduction

Full-body pose reconstruction based on Head-Mounted Displays (HMDs) facilitates a diverse array of AR/VR applications, including patient rehabilitation, realistic avatar generation, and the control of teleoperated humanoid robots (He et al. 2024; Dai et al. 2024c). However, due to the sparsity of signals captured by HMDs, inferring accurate and natural full-body motion remains a challenging problem.

Previous works have demonstrated the feasibility of reconstructing realistic full-body motion, but often at the expense of substantial computational cost and large parameter counts, which limits its application. For example, AvatarJLM (Zheng et al. 2023) improves performance by stacking multiple Transformer blocks (Vaswani 2017) in a high-dimensional space, while SAGE (Feng et al. 2024) leverages large generative models such as VQ-VAE (Van Den Oord, Vinyals et al. 2017) and diffusion models (Rombach et al. 2022). Their high deployment costs underline the need for more efficient solutions that can achieve high accuracy with a compact framework.

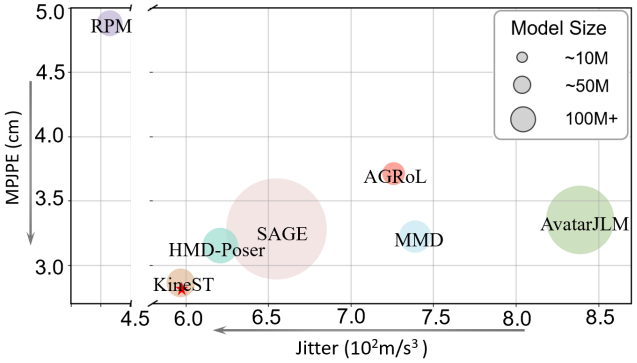


Figure 1: Comparison of our approach with state-of-the-art methods in terms of overall performance. Our method achieves the smallest average position error and smoother motion, and maintains a lightweight model architecture.

To build lightweight and robust models for realistic full-body pose estimation from sparse inputs, recent works have explored various solutions. For example, RPM (Barquero et al. 2025) introduces a prediction consistency anchor to reduce sudden motion drift, which improves smoothness but sacrifices pose accuracy. To improve pose accuracy, separate temporal and spatial modules are adopted to better capture the complex dependencies of human motion (Dai et al. 2024a; Dong et al. 2024). Although this dual-module design enhances joint interaction modeling, some modeling capacities are shifted to single-frame spatial features, which can compromise motion smoothness. Therefore, a key question is raised: **how can we design a model that remains lightweight yet achieves both high accuracy and motion smoothness?**

Recently, the State Space Duality (SSD) framework (Dao and Gu 2024) introduces a special and robust scanning strategy and shows great promise for efficient time-series modeling, making it a strong candidate for our task. However,

*These authors contributed equally.
†Corresponding author

directly applying SSD to human motion tracking yields unsatisfactory performance, primarily due to its unidirectional scanning and the absence of specific designs for full-body pose reconstruction.

To tackle this challenge, we propose KineST, a lightweight yet effective model, to fully extract spatiotemporal dependencies while integrating local and global pose information. Specifically, we first design a Temporal Flow Module (TFM) to learn inter-frame dynamics. The main components in TFM are SSD blocks in which a bidirectional scanning strategy is employed to initially capture motion features. These are followed by a Local Motion Aggregator (LMA) and a Global Motion Aggregator (GMA), which progressively refine local and global motion dependencies.

Secondly, we introduce a robust Spatiotemporal Kinematic Flow Module (SKFM), which employs a Spatiotemporal Mixing Mechanism (STMM) to tightly couple spatial and temporal contexts and maintain a balance between accuracy and temporal continuity. Moreover, a novel Kinematic Tree Scanning Strategy (KTSS) is employed to incorporate kinematic priors into spatial feature capture and fully capture intricate joint relations. To further improve motion continuity, a geometric angular velocity loss is proposed, jointly constraining both the magnitude and direction of rotation changes in a geometrically consistent way. The contributions of this work are summarized as follows:

- A kinematics-guided state space model, KineST, is proposed, which can not only fully extract spatiotemporal information but also integrate local-global perception.
- A robust Spatiotemporal Kinematic Flow Module (SKFM) is designed, which applies the Spatiotemporal Mixing Mechanism (STMM) to tightly couple spatial and temporal contexts and employs a novel Kinematic Tree Scanning Strategy (KTSS) to fully capture intricate joint relations.
- To further improve motion continuity, a geometric angular velocity loss is proposed to impose physically meaningful constraints on rotational variations.
- Extensive experiments demonstrate KineST's superiority over current state-of-the-art methods. Furthermore, comprehensive ablation studies confirm the contribution of each carefully designed component.

Related Works

Motion Tracking from Sparse Observations

Earlier methods explore full-body tracking using 4 or 6 IMUs (Huang et al. 2018; Yi et al. 2022; Yang, Kim, and Lee 2021; Yi, Zhou, and Xu 2021; Von Marcard et al. 2017; Jiang et al. 2022b). However, in AR/VR scenarios, HMDs are more practical and widely adopted, which typically provide only 3 tracking signals from the head and hands. Based on HMD inputs, generative techniques are adopted to synthesize full-body poses. For example, in FLAG (Aliakbarian et al. 2022) and VAE-HMD (Dittadi et al. 2021), a variational auto-encoder (VAE) and a flow-based model are applied, respectively. In AGRoL (Du et al. 2023) and SAGE (Feng et al. 2024), the reconstruction of avatars is achieved by diffusion models or VQ-VAEs. Correspondingly, another type of work is based on regression-based approaches. Transformer-based architecture is adopted to predict full-body poses from these three sparse signals, such as AvatarPoser (Jiang et al. 2022a), AvatarJLM (Zheng et al. 2023), HMD-Poser (Dai et al. 2024a), and RPM (Barquero et al. 2025). KCTD (Dai et al. 2024b) designs an MLP-based model with kinematic constraints for the task. With the recent advancements in state space models (SSMs), several studies have explored their potential for this task. For instance, MMD (Dong et al. 2024) leverages the sequential modeling capability of SSMs to track full-body poses in the temporal and spatial dimensions, respectively. SSD-Poser (Zhao et al. 2025) further introduces a hybrid architecture combining SSD and Transformers to efficiently capture contextual motion features.

Although promising results are achieved, they struggle to balance pose accuracy and motion smoothness within a lightweight framework. To address this, a kinematics-guided spatiotemporal state space model, KineST, is designed to improve both accuracy and continuity.

Spatial Modeling in Human Pose Estimation

Spatial modeling plays a vital role in full-body motion reconstruction from sparse signals, where accurately capturing inter-joint relationships within a single frame is crucial for reconstructing plausible poses. While spatial modeling has been widely explored in related domains such as action recognition and image-based pose estimation (Zheng et al. 2024; Jiao et al. 2025; Feng et al. 2023; Tang et al. 2023; Wu, Zhang, and Zou 2023; Qian et al. 2023; Zhao et al. 2023), its application under sparse input constraints presents unique challenges. In most existing methods (Zheng et al. 2023; Aliakbarian et al. 2023; Dai et al. 2024a; Qian et al. 2024), self-attention mechanisms are employed to treat the 22 body joints as independent tokens and learn pairwise dependencies through similarity weights(Fig. 3(b)). In contrast, BPG (Yao, Wu, and Yi 2024) employs a GCN-based structure to explicitly encode local joint connectivity via an adjacency matrix, while MMD (Dong et al. 2024) leverages SSMs to process concatenated joint features as a whole, implicitly modeling global joint correlations.

Despite their effectiveness, these approaches focus on either local or global relations, and often fail to tightly integrate spatial and temporal contexts. To address this limitation, we introduce KTSS to enhance the spatial modeling capacity based on body topology, and further propose STMM to jointly encode spatial-temporal dependencies, striking a better balance between accuracy and smoothness.

Preliminary

State Space Duality The SSD framework (Dao and Gu 2024) is an advanced and efficient variant of SSMs, offering enhanced capabilities in both computation and inference. This framework introduces a novel matrix-based computation structure, which integrates the linear recurrence properties of SSMs with a quadratic dual formulation. Specifically,

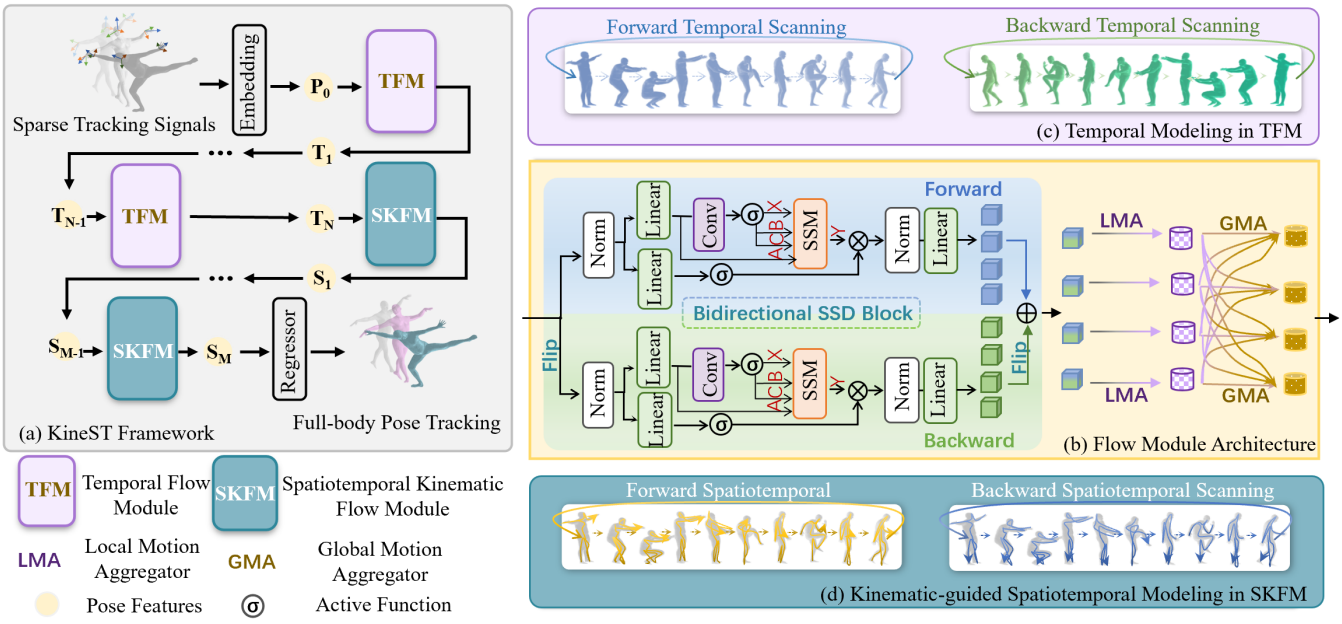


Figure 2: Overall architecture. (a) The architecture of the proposed KineST model, whose main components are the temporal flow module (TFM) and spatiotemporal kinematic flow module (SKFM). (b) The shared structure of the flow module used in both TFM and SKFM, which comprises a bidirectional SSD block, a local motion aggregator (LMA), and a global motion aggregator (GMA). (c) Temporal modeling within the TFM. (d) Kinematics-guided spatiotemporal modeling within the SKFM.

the SSD framework is denoted as follows:

$$\begin{aligned} y_t &= \sum_{i=0}^{t} C_t^T A_{t:i}^X B_i x_i, \\ y &= SSM(A,B,C)(x) = Mx, \end{aligned} \tag{1}$$

where $A_{t:i}^X$ denotes the product of A terms from $i+1$ to t, and M is defined as:$M_{ji} := C_j^\top A_j \cdots A_{i+1} B_i$. When A_i is reduced to a scalar, Eq.(1) can be reformulated as:

$$\begin{aligned} y &= Mx = F \cdot \left(C^T\ B\right) x, \\ where F_{ji} &= \begin{cases} A_j A_{j-1} \cdots A_{i+1} & i < j \\ 1 & i = j \\ 0 & i > j \end{cases}. \end{aligned} \tag{2}$$

In SSD framework, the original time-series recurrence is reformulated into an equivalent product-sum matrix form, which can be efficiently parallelized by associative scan algorithms (Smith, Warrington, and Linderman 2022; Martin and Cundy 2017). In this work, we extend this scan mechanism by embedding kinematic priors and bidirectional constraints to enhance realistic full-body pose reconstruction.

Method

Problem Formulation

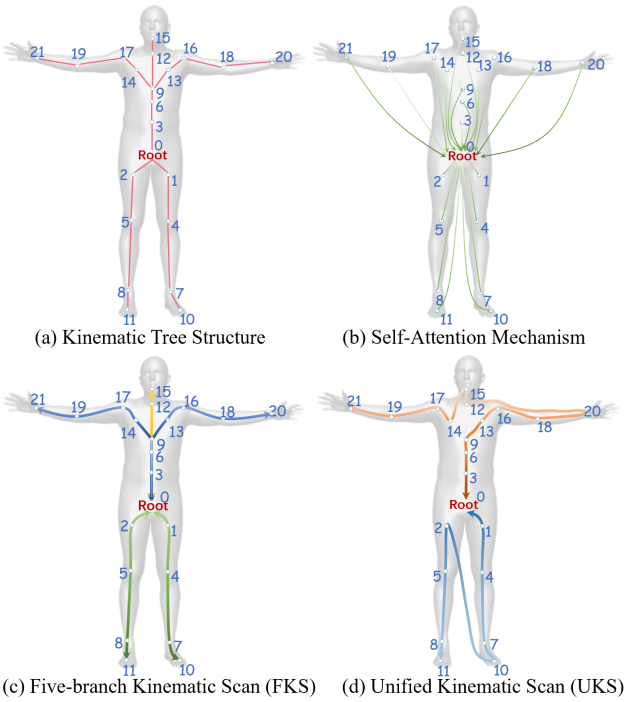
The purpose of this task is to achieve realistic full-body motion prediction $Y = \{y_i\}_{i=1}^{L} \in \mathbb{R}^{L \times V}$ from sparse IMU signals $X = \{x_i\}_{i=1}^{L} \in \mathbb{R}^{L \times C}$ captured from the head and hands over L time frames, where C and V denote the input and output dimensions, respectively. The ground-truth full-body motion is denoted as $Z = \{z_i\}_{i=1}^{L} \in \mathbb{R}^{L \times V}$. Following (Jiang et al. 2022a), each $\mathbf{x}_i$ contains a 3D position, 6D rotation, linear velocity, and angular velocity for each of the three tracked parts. We adopt pose parameters of the first 22 joints of the SMPL model (Loper et al. 2023) to represent the output. As a result, the input and output dimensions are $C = 3 \times (3+6+3+6)$ and $V = 22 \times 6$.

Overall Architecture

The overall architecture of the proposed KineST model consists of Temporal Flow Modules (TFMs) and Spatiotemporal Kinematic Flow Modules (SKFMs), as shown in Fig. 2 (a). Given the sparse tracking signals $X \in \mathbb{R}^{L \times C}$, we first use a single linear layer to obtain embedded pose features $P_0 \in \mathbb{R}^{L \times E}$. Here C and E represent the feature dimensions of the original and embedded signals, respectively. Subsequently, the $P_0 \in \mathbb{R}^{L \times E}$ are processed through a stack of N TFMs, producing deep temporal motion features $T_n \in \mathbb{R}^{L \times E}$ at each stage, where $n \in \{1, 2, ..., N\}$. To further learn kinematics-guided spatiotemporal information, the output $T_N \in \mathbb{R}^{L \times E}$ is passed through a sequence of M SKFM blocks. The S_m is produced at each stage, where $m \in \{1, 2, ..., M\}$. Finally, the full-body motion poses are estimated by a linear regressor.

Temporal Flow Module

To learn inter-frame motion dynamics, we propose the Temporal Flow Module (TFM) whose architecture is shown in Fig. 2(b). It adopts a bidirectional SSD block (Bi-SSD) with parallel forward and backward branches to enhance temporal modeling. Each SSD block comprises layer normal-



(a) Kinematic Tree Structure (b) Self-Attention Mechanism

(c) Five-branch Kinematic Scan (FKS) (d) Unified Kinematic Scan (UKS)

Figure 3: Comparison of different scanning strategies.

ization (LN), linear layers, convolutional layers, the SiLU activation function, and the core state space model (SSM). Given embedded features $P_0 \in \mathbb{R}^{L\times E}$, the forward branch is formulated as:

$$\begin{aligned} \mathrm{X, B, C} &= \mathrm{SiLU(Conv(Linear(LN(P_0))))}, \\ \mathrm{A} &= \mathrm{Linear(LN(P_0))}, \\ f_1 &= \mathrm{SiLU(Linear(LN(P_0)))}, \\ \mathrm{F}_f^t &= \mathrm{Linear(LN}(f_1 \odot \mathrm{SSM}(X, A, B, C))), \end{aligned} \tag{3}$$

where X, A, B and C denote the state vector, state transition matrix, input matrix, and output matrix, respectively, while f_1 serves as an adaptive gating vector and F_f^t represents forward temporal motion features. The backward branch shares the same structure, except that P_0 is time-reversed to obtain P_0^{flip}, passed through the same operations, and flipped back to yield backward features F_b^t. To enhance representation, we further apply a Local Motion Aggregator (LMA) and a Global Motion Aggregator (GMA) for modeling local dependencies and global motion periodicity, respectively:

$$\mathrm{T}_1 = \mathrm{GMA(LMA}(F_f^t + F_b^t)). \tag{4}$$

LMA is implemented via convolution, while GMA uses a lightweight transformer.

Spatiotemporal Kinematic Flow Module

To enhance inter-joint dependencies while preserving motion smoothness, we propose the Spatiotemporal Kinematic Flow Module (SKFM), which performs kinematics-guided spatiotemporal modeling. Specifically, a Kinematic Tree Scanning Strategy (KTSS) is introduced to inject kinematic priors into spatial feature extraction enhancing joint interactions, while a Spatiotemporal Mixing Mechanism (STMM) is incorporated to tightly couple spatial and temporal features to balance pose accuracy and continuity. It is worth noting that SKFM shares the same overall structure as TFM, as shown in Fig. 2(b).

Algorithm 1: Spatiotemporal Mixing Mechanism

Input: Final temporal motion features $\boldsymbol{T}_N \in \mathbb{R}^{L\times E}$
Output: Final spatiotemporal motion features $\boldsymbol{S}_M \in \mathbb{R}^{L\times E}$

for i in M **do**
  if $i = 0$ **then** $\boldsymbol{S}_l \leftarrow \mathrm{Linear}(\boldsymbol{T}_N)$
  else $\boldsymbol{S}_l \leftarrow \mathrm{Linear}(\boldsymbol{S}_i)$
  end if
  $\boldsymbol{S}'_l \leftarrow \mathrm{rearrange}(\boldsymbol{S}_l,\ L\ F \rightarrow L\ (J\ D))$
  $\boldsymbol{S}_f,\ \boldsymbol{S}_b \leftarrow \mathrm{KTSS}(\boldsymbol{S}'_l)$
  $\boldsymbol{S}'_f \leftarrow \mathrm{rearrange}\ (\boldsymbol{S}_f,\ L\ J_f\ D \rightarrow (L\ J_f)\ D)$
  $\boldsymbol{S}'_b \leftarrow \mathrm{rearrange}\ (\boldsymbol{S}_b,\ L\ J_b\ D \rightarrow (L\ J_b)\ D)$
  $\boldsymbol{S}''_f,\ \boldsymbol{S}''_b \leftarrow \mathrm{Bi\text{-}SSD}(\boldsymbol{S}'_f,\ \boldsymbol{S}'_b)$
  $\boldsymbol{S}_{i+1} \leftarrow \mathrm{GMA(LMA(Linear}(\boldsymbol{S}''_f + \boldsymbol{S}''_b)))$
end for
return $\boldsymbol{S}_M$

Kinematic Tree Scanning Strategy In the SMPL model, the human skeleton is represented as a hierarchical structure (Fig. 3 (a)), where a kinematic tree encodes the parent-child relationship between joints. Such kinematic representations are essential for realistic motion tracking and allow efficient control over joint transformations. However, existing methods overlook these important kinematic priors, leading to suboptimal performance. In this paper, we leverage the sequential nature of SSD, which allows each joint feature to be inferred from the previous joint state. By reformulating the original unidirectional scan into a kinematics-guided bidirectional scan, the interactions between the parent-child joints are effectively captured, enabling features to flow forward and backward along the kinematic hierarchy. Within this design, we introduce two distinct scanning variants under this framework for efficient full-body reconstruction.

First, we introduce the Five-branch Kinematic Scan (FKS), which strictly follows the kinematic tree structure, as shown in Fig. 3(c). The forward scanning order is [0,1,4,7,10,0,2,5,8,11,0,3,6,9,13,16,18,20, 0,3,6,9,12,15,0,3,6,9,14,17,19,21]. Compare to the index-order scan in SMPL (sequentially from 0 to 21), FKS allows better perception of local kinematic dependencies. However, its branch-wise design limits the model's ability to capture integral body movement dynamics (Table 4). To address these issues, we propose the Unified Kinematic Scan (UKS), as shown in Fig. 3(d). By positioning the root joint centrally, UKS effectively couples upper and lower body motion, enhancing global motion coherence and overall reconstruction robustness. The specific forward scanning order is [21,19,17,14,15,12,20,18,16,13,9,6,3,0,1,4,7,10,2,5,8,11]. Therefore, the proposed KTSS mainly adopts UKS to more effectively capture full-body joint dependencies.

Spatiotemporal Mixing Mechanism To capture inter-joint relations and maintain smoothness over frames, the Spatiotemporal Mixing Mechanism (STMM) is employed through mixed spatiotemporal representation learning. The process of STMM is presented in Algorithm 1. Specially, we start from the final temporal motion features $T_N \in \mathbb{R}^{L\times E}$, which are projected into a latent joint space $S_l \in \mathbb{R}^{L\times H}$ and then reshaped into detailed joint features $S_l' \in \mathbb{R}^{L\times J\times D}$, where $H = J\times D$. Following KTSS, the joints are reordered along both forward and backward directions to align features according to the kinematic chain, resulting in two new tensors: $S_f \in \mathbb{R}^{L\times J_f\times D}$ and $S_b \in \mathbb{R}^{L\times J_b\times D}$, where J_f and J_b denote the forward and backward joint sequences, respectively. To perceive spatial and temporal features simultaneously, we rearrange the sequence and joint dimensions into a unified axis, obtaining two mixed tensors: $S_f' \in \mathbb{R}^{(LJ_f)\times D}$ and $S_b' \in \mathbb{R}^{(LJ_b)\times D}$. These tensors are then processed by the Bi-SSD to enhance spatiotemporal dependencies. Finally, the bidirectional features are summed and linearly projected, followed by sequential aggregation via LMA and GMA, to produce the refined spatiotemporal features $S_1 \in \mathbb{R}^{L\times E}$. After M iterations of processing, the final spatiotemporal motion features $S_M \in \mathbb{R}^{L\times E}$ are obtained, preparing for regression.

Loss Function

We retain two commonly used supervision terms, the L1 loss on body orientation and the L1 loss on joint rotations (Jiang et al. 2022a). To further enhance motion continuity, a geometric angular velocity loss $\mathcal{L}_{\text{angvel}}^{\text{geo}}$ is proposed to physically capture the true geometric relationship between rotations. Unlike existing works (Barquero et al. 2025; Aliakbarian et al. 2023) estimating angular velocity by applying first-order finite differences to rotation representations, we compute angular velocity within its tangent space, namely the Lie algebra $\mathfrak{so}(3)$ (Murray, Li, and Sastry 2017).

Specifically, we begin by defining the angular velocity at time t as the geodesic rotational difference between two consecutive frames, formulated as $V_t = R_{t-1}^{-1}R_t$. Here, R_t is the rotation matrix derived from z_t via Gram-Schmidt orthogonalization. To accurately measure the difference between two angular velocities, we impose constraints from two perspectives: rotational magnitude and rotational direction. The rotational magnitude is measured using a natural Riemannian metric defined on the compact Lie group $\mathrm{SO}(3)$, while the rotational direction is extracted from the skew-symmetric part of the relative rotation matrix. Each relative rotation is eventually mapped to its corresponding axis-angle representation, and the formulation is defined as follows:

$$\begin{aligned}\theta_V &= \arccos\left(\frac{\mathrm{Tr}(V)-1}{2}\right),\\ \log V &= \theta_V \cdot \frac{1}{2\sin\theta_V}\begin{bmatrix}V_{32}-V_{23}\\ V_{13}-V_{31}\\ V_{21}-V_{12}\end{bmatrix},\\ \mathcal{L}_{\text{angvel}}^{\text{geo}} &= \sum_{t=1}^{T-1}\left\|\log(V_t)-\log(\hat{V}_t)\right\|_1.\end{aligned} \tag{5}$$

Here, V_t denotes the GT angular velocity at frame t, $\hat{V}_t$ represents the corresponding predicted angular velocity computed from y_t, θ_V denotes the rotational magnitude of V.

In summary, the final loss function is defined as:

$$\mathcal{L} = \alpha\mathcal{L}_{\text{rot}} + \beta\mathcal{L}_{\text{ori}} + \delta\mathcal{L}_{\text{angvel}}^{\text{geo}}.$$

The weights for each component are empirically set to $\alpha = 1$, $\beta = 0.02$, and $\delta = 1$, respectively.

Experiments

Dataset and Implementation Details

Our method is trained and evaluated on the AMASS (Mahmood et al. 2019) and a real-captured dataset (Zheng et al. 2023), both represented using SMPL parameters. The TFMs and SKFMs each contain 2 modules, with an embedded feature dimension $E = 256$. The number of full-body joints J is 22, each with a latent dimension $D = 64$, and the input sequence length $L = 96$. Training is conducted on an NVIDIA 4090 with a batch size of 256. We adopt the Adam optimizer (Kingma 2014) with a learning rate of 3e-4 (decayed to 3e-5 after 200000 iterations) and a weight decay of 1e-5. Inference is efficient, requiring only 12.9 ms to process 96 frames on the same GPU.

Evaluation Metrics

Following (Du et al. 2023), we assess model performance across three distinct metric categories. The first category measures rotational accuracy and is represented by the Mean Per Joint Rotation Error (MPJRE) [degrees]. The second category focuses on motion smoothness, which includes the Mean Per Joint Velocity Error (MPJVE) [cm/s] and Jitter ($10^2 m/s^3$) (Yi et al. 2022). The third category evaluates positional accuracy and includes the Mean Per Joint Position Error (MPJPE) [cm], Root PE, Hand PE, Upper PE, and Lower PE.

Evaluation Results

We follow the recent common practice (Dai et al. 2024a; Barquero et al. 2025) of using AMASS dataset with two different protocols. **In the first protocol**, the subsets of the AMASS dataset, including CMU (Carnegie Mellon University 2000), BMLrub (Troje 2002) and HDM05 (Müller et al. 2007), are split into 90% training data and 10% testing data. **In the second protocol**, a larger benchmark from AMASS (Mahmood et al. 2019) is utilized, including 12 subsets for training and the HumanEva (Sigal, Balan, and Black 2010) and the Transition (Mahmood et al. 2019) subsets for testing. To further assess the real-world applicability of our method, we introduce **a third protocol**, where the model is evaluated on real headset-and-controller data collected by AvatarJLM (Zheng et al. 2023) in an online setting, and the training setup remains consistent with Protocol 2.

Quantitative Evaluation As shown in Table 1, Table 2, and Table 3, our method achieves the best overall performance compared to existing approaches, demonstrating both high accuracy and robustness across key metrics. Specifically, in Table 1, KineST achieves a 2.59% reduction in

Method	MPJRE↓	MPJPE↓	MPJVE↓	Hand PE↓	Upper PE↓	Lower PE↓	Root PE↓	Jitter↓	Param.
AvatarPoser (Jiang et al. 2022a)	3.08	4.18	27.70	2.12	1.81	7.59	3.34	14.49	4M
AvatarJLM (Zheng et al. 2023)	2.90	3.35	20.79	1.24	1.72	6.20	2.94	8.39	63M
AGRoL (Du et al. 2023)	2.66	3.71	18.59	1.31	1.55	6.84	3.36	7.26	7M
SAGE (Feng et al. 2024)	2.53	3.28	20.62	1.18	1.39	6.01	2.95	6.55	137M
HMD-Poser* (Dai et al. 2024a)	2.32	3.15	18.15	1.35	1.34	5.76	2.76	6.21	17M
MMD (Dong et al. 2024)	2.31	3.22	17.88	**0.94**	1.29	6.01	2.88	7.39	14M
RPM* (Barquero et al. 2025)	3.69	4.88	21.99	5.90	2.94	7.69	4.01	**4.43**	9M
KineST (Ours)	**2.25**	**2.86**	**15.26**	1.04	**1.24**	**5.20**	**2.65**	5.97	11M
GT	0	0	0	0	0	0	0	4.00	——

Table 1: Evaluation results on three subsets of AMASS under Protocol 1. The best results are in bold, and the second-best results are underlined. For fair comparison, HMD-Poser and RPM are retrained using their public code, as they originally use different body shape parameters or FPS, and are denoted with *.

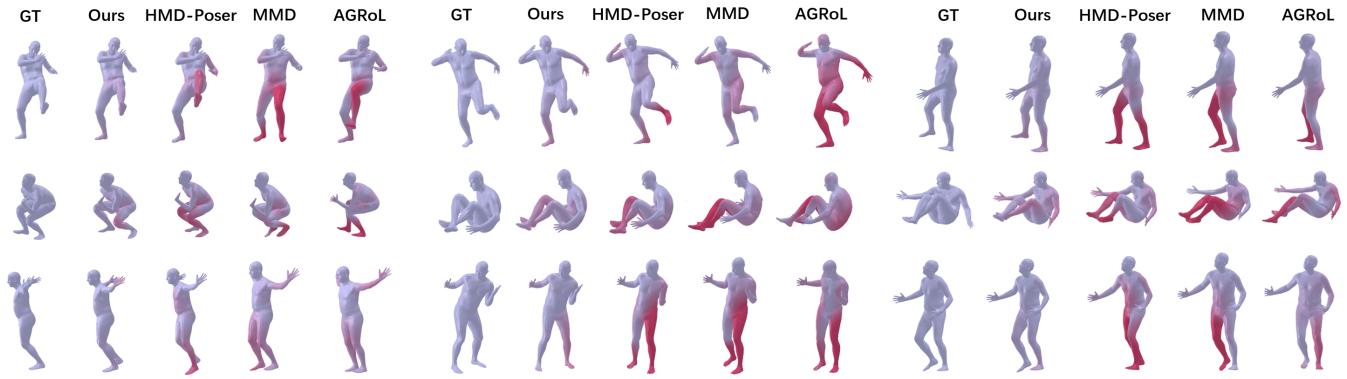


Figure 4: Visualization results of different actions compared with other methods. The joint error degrees are indicated by red shading, allowing a comparative assessment of reconstruction accuracy across various poses for each method. These visuals confirm the robustness and enhancements of the proposed model, particularly in lower body predictions.

Method	MPJRE↓	MPJPE↓	MPJVE↓	Jitter↓
AvatarPoser	4.70	6.38	34.05	10.21
AvatarJLM	4.30	**4.93**	26.17	7.19
AGRoL	4.30	6.17	24.40	8.32
SAGE	4.62	5.86	33.54	7.13
HMD-Poser*	4.36	5.60	29.32	7.46
RPM*	5.37	7.19	29.27	**3.48**
KineST (Ours)	**4.28**	5.17	**24.08**	7.36
GT	0	0	0	2.93

Table 2: Evaluation results on AMASS under Protocol 2.

MPJRE, an 11.18% decrease in MPJPE, and a 14.65% improvement in MPJVE compared to MMD. Furthermore, as shown in Table 2, KineST achieves the lowest MPJRE and MPJVE, indicating its ability to construct precise and smooth motions in various actions. Although RPM achieves the best result in jitter, it exhibits significantly reduced pose accuracy. Notably, in Table 2, AvatarJLM achieves competitive MPJPE, and SAGE reports slightly lower jitter. However, they both rely on complex architectures with large parameter counts (63M and 137M respectively), which limits their practicality for lightweight AR/VR deployment. As shown in Table 3, our method outperforms all other methods across all metrics, demonstrating the practical applicability in real-world scenarios.

Method	MPJRE↓	MPJPE↓	MPJVE↓	Jitter↓
AvatarPoser	7.28	11.22	31.67	12.87
AvatarJLM	7.01	9.72	27.59	13.10
KineST (Ours)	**6.91**	**9.68**	**25.16**	**9.49**

Table 3: Evaluation results on the real-captured data under Protocol 3.

Qualitative Evaluation Visualization results are presented in Figure 4 and Figure 5. In Figure 4, we compare the reconstruction errors of a single pose with other SOTA methods. In Figure 5, continuous motion sequence results are shown, with samples taken every 120 frames. The visualization results demonstrate that the proposed method achieves performance closest to the ground truth (GT).

Ablation Study

In this section, the ablation studies are conducted on the AMASS dataset following the protocol 1 setting to verify the impact of each component and parameter in our designed model. The results are shown in Table 4-6.

Effects of different scanning strategies We compare three distinct strategies, including the index-order scan in SMPL, which sequentially scans from index 0 to 21, and the two proposed kinematics-guided bidirectional scans, FKS

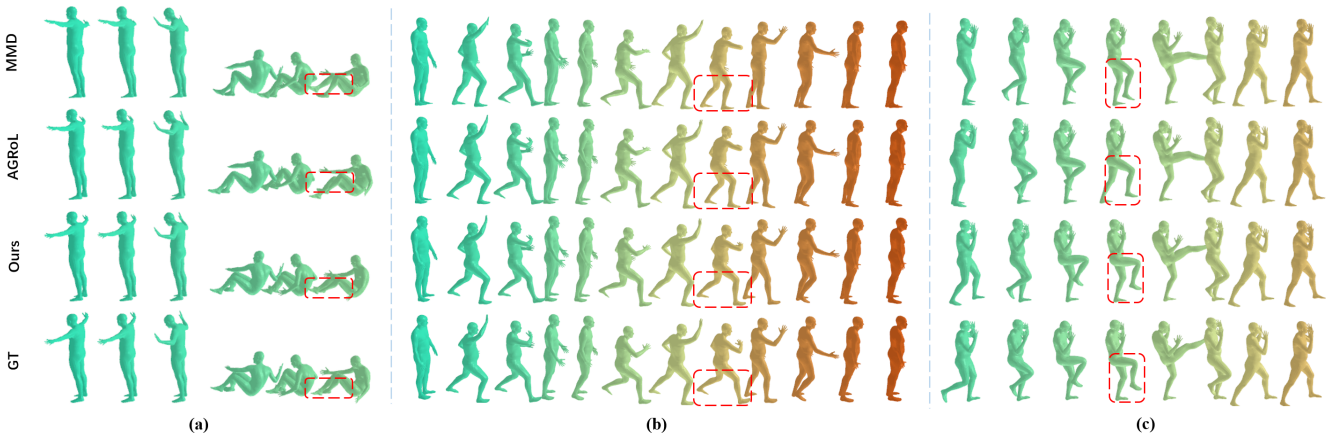


Figure 5: Visualization results of continuous pose sequences compared with other methods. The visualization illustrates that the proposed model delivers smoother and more realistic body motion tracking. Notably, the proposed model provides refined reconstruction highlighted by red dashed boxes.

Method	MPJRE↓	MPJPE↓	MPJVE↓	Jitter↓
Index-order in SMPL	2.32	3.11	17.81	8.27
FKS (Ours)	2.28	3.00	16.25	7.01
UKS (Ours)	**2.25**	**2.86**	**15.26**	**5.97**

Table 4: Evaluating the effect of different scanning strategies.

and UKS. As shown in Table 4, by embedding kinematic priors, FKS and UKS can effectively improve performance in joint relationship extraction. Additionally, due to the branch-wise design that harms the integrity of the human body, FKS cannot achieve better motion smoothness compared to UKS.

Effects of different modeling mechanisms in SKFM To evaluate the effectiveness of the proposed STMM in SKFM, we compare it with two other modeling mechanisms: pure temporal modeling and pure spatial modeling. Note that prior works have adopted two representative types of pure spatial modeling. One approach adopts a holistic modeling method that processes the concatenated features of all joints as a whole (Dong et al. 2024), while the other follows a token-wise method that treats each of the 22 joints as an independent token to extract spatial features (Zheng et al. 2023; Dai et al. 2024a). As shown in Table 5, the proposed STMM achieves a favorable balance between pose accuracy and motion smoothness compared to the other methods. While the token-wise spatial modeling yields the lowest MPJRE, independently extracting features for each joint inevitably overlooks the temporal continuity.

Effects of the loss function We employ three distinct loss functions to train KineST. Since L_{rot} and L_{ori} are validated in previous works (Jiang et al. 2022a; Zheng et al. 2023), we take them as the baseline and focus on assessing the role of the proposed geometric angular velocity loss L^{geo}_{angvel}. Unlike calculating angular velocity using first-order finite differences (L^{diff}_{angvel}) (Barquero et al. 2025), our proposed

Method	MPJRE↓	MPJPE↓	MPJVE↓	Jitter↓
Pure Temporal	2.27	2.97	16.84	7.83
Pure Spatial (holistic)	2.41	3.10	16.77	7.72
Pure Spatial (token-wise)	**2.23**	2.93	17.85	9.31
STMM (Ours)	2.25	**2.86**	**15.26**	**5.97**

Table 5: Evaluating the effect of different modeling mechanisms in SKFM.

Method	MPJRE↓	MPJPE↓	MPJVE↓	Jitter↓
Baseline	2.25	2.87	16.10	6.75
with L^{diff}_{angvel}	2.29	3.03	15.91	6.44
with L^{geo}_{angvel}(Ours)	**2.25**	**2.86**	**15.26**	**5.97**

Table 6: Evaluating the effect of the loss function.

L^{geo}_{angvel} operates it on the Lie group SO(3) in a geometrically consistent manner. As shown in Table 6, although L^{diff}_{angvel} helps reduce MPJVE and Jitter, it significantly increases the errors in rotation and position. In contrast, the proposed L^{geo}_{angvel} achieves smoother motion while preserving the performance on MPJRE and MPJPE, ensuring a good balance between accuracy and smoothness.

Conclusion

A novel kinematics-guided state space model, KineST, is proposed for full-body motion tracking from sparse signals, which mainly relies on two key innovations, the kinematics-guided bidirectional scanning strategy and the mixed spatiotemporal representation learning mechanism. To further improve motion continuity, a geometric angular velocity loss is designed to regulate rotational variations in manifold space. Extensive experiments show that KineST achieves superior accuracy and temporal consistency within a lightweight framework. We believe that the proposed model can significantly contribute to an immersive and seamless user experience in AR/VR applications.

Acknowledgments

This work was partly supported by the National Natural Science Foundation of China (Grant No.82472116) and the Natural Science Foundation of Shanghai (Grant No.24ZR1404100).

References

Aliakbarian, S.; Cameron, P.; Bogo, F.; Fitzgibbon, A.; and Cashman, T. J. 2022. Flag: Flow-based 3d avatar generation from sparse observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13253–13262.

Aliakbarian, S.; Saleh, F.; Collier, D.; Cameron, P.; and Cosker, D. 2023. Hmd-nemo: Online 3d avatar motion generation from sparse observations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9622–9631.

Barquero, G.; Bertsch, N.; Marramreddy, M.; Chacón, C.; Arcadu, F.; Rigual, F.; He, N. S.; Palmero, C.; Escalera, S.; Ye, Y.; et al. 2025. From Sparse Signal to Smooth Motion: Real-Time Motion Generation with Rolling Prediction Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 1850–1860.

Carnegie Mellon University. 2000. Cmu graphics lab motion capture database.

Dai, P.; Zhang, Y.; Liu, T.; Fan, Z.; Du, T.; Su, Z.; Zheng, X.; and Li, Z. 2024a. 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*, 874–884.

Dai, X.; Zhang, X.; Li, S.; and Chen, X. 2024b. Enhancing Motion Reconstruction From Sparse Tracking Inputs With Kinematic Constraints. *IEEE Transactions on Automation Science and Engineering*.

Dai, X.; Zhang, Z.; Zhao, S.; Liu, X.; and Chen, X. 2024c. Full-body pose reconstruction and correction in virtual reality for rehabilitation training. *Frontiers in Neuroscience*, 18: 1388742.

Dao, T.; and Gu, A. 2024. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*.

Dittadi, A.; Dziadzio, S.; Cosker, D.; Lundell, B.; Cashman, T. J.; and Shotton, J. 2021. Full-body motion from a single head-mounted device: Generating smpl poses from partial observations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11687–11697.

Dong, K.; Xue, J.; Niu, Z.; Lan, X.; Lu, K.; Liu, Q.; and Qin, X. 2024. Realistic full-body motion generation from sparse tracking with state space model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 4024–4033.

Du, Y.; Kips, R.; Pumarola, A.; Starke, S.; Thabet, A.; and Sanakoyeu, A. 2023. 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*, 481–490.

Feng, H.; Ma, W.; Gao, Q.; Zheng, X.; Xue, N.; and Xu, H. 2024. Stratified Avatar Generation from Sparse Observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 153–163.

Feng, R.; Gao, Y.; Tse, T. H. E.; Ma, X.; and Chang, H. J. 2023. Diffpose: Spatiotemporal diffusion model for video-based human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14861–14872.

He, T.; Luo, Z.; Xiao, W.; Zhang, C.; Kitani, K.; Liu, C.; and Shi, G. 2024. Learning human-to-humanoid real-time whole-body teleoperation. *arXiv preprint arXiv:2403.04436*.

Huang, Y.; Kaufmann, M.; Aksan, E.; Black, M. J.; Hilliges, O.; and Pons-Moll, G. 2018. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics (TOG)*, 37(6): 1–15.

Jiang, J.; Streli, P.; Qiu, H.; Fender, A.; Laich, L.; Snape, P.; and Holz, C. 2022a. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *European conference on computer vision*, 443–460. Springer.

Jiang, Y.; Ye, Y.; Gopinath, D.; Won, J.; Winkler, A. W.; and Liu, C. K. 2022b. Transformer inertial poser: Real-time human motion reconstruction from sparse imus with simultaneous terrain generation. In *SIGGRAPH Asia 2022 Conference Papers*, 1–9.

Jiao, Y.; Wang, Z.; Wu, S.; Fan, S.; Liu, Z.; Xu, Z.; and Wu, Z. 2025. Spatiotemporal learning for human pose estimation in sparsely-labeled videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 4093–4101.

Kingma, D. P. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 851–866.

Mahmood, N.; Ghorbani, N.; Troje, N. F.; Pons-Moll, G.; and Black, M. J. 2019. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5442–5451.

Martin, E.; and Cundy, C. 2017. Parallelizing linear recurrent neural nets over sequence length. *arXiv preprint arXiv:1709.04057*.

Müller, M.; Röder, T.; Clausen, M.; Eberhardt, B.; Krüger, B.; and Weber, A. 2007. Mocap database hdm05. *Institut für Informatik II, Universität Bonn*, 2(7).

Murray, R. M.; Li, Z.; and Sastry, S. S. 2017. *A mathematical introduction to robotic manipulation*. CRC press.

Qian, B.; Wei, Z.; Li, J.; and Wei, X. 2024. Reliaavatar: A robust real-time avatar animator with integrated motion prediction. *arXiv preprint arXiv:2407.02129*.

Qian, X.; Tang, Y.; Zhang, N.; Han, M.; Xiao, J.; Huang, M.-C.; and Lin, R.-S. 2023. Hstformer: Hierarchical spatial-temporal transformers for 3d human pose estimation. *arXiv preprint arXiv:2301.07322*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Sigal, L.; Balan, A. O.; and Black, M. J. 2010. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1): 4–27.

Smith, J. T.; Warrington, A.; and Linderman, S. W. 2022. Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933*.

Tang, Z.; Qiu, Z.; Hao, Y.; Hong, R.; and Yao, T. 2023. 3d human pose estimation with spatio-temporal criss-cross attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4790–4799.

Troje, N. F. 2002. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of vision*, 2(5): 2–2.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

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

Von Marcard, T.; Rosenhahn, B.; Black, M. J.; and Pons-Moll, G. 2017. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer graphics forum*, volume 36, 349–360. Wiley Online Library.

Wu, L.; Zhang, C.; and Zou, Y. 2023. SpatioTemporal focus for skeleton-based action recognition. *Pattern Recognition*, 136: 109231.

Yang, D.; Kim, D.; and Lee, S.-H. 2021. Lobstr: Real-time lower-body pose prediction from sparse upper-body tracking signals. In *Computer Graphics Forum*, volume 40, 265–275. Wiley Online Library.

Yao, F.; Wu, Z.; and Yi, L. 2024. Full-body motion reconstruction with sparse sensing from graph perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6612–6620.

Yi, X.; Zhou, Y.; Habermann, M.; Shimada, S.; Golyanik, V.; Theobalt, C.; and Xu, F. 2022. 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*, 13167–13178.

Yi, X.; Zhou, Y.; and Xu, F. 2021. Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions On Graphics (TOG)*, 40(4): 1–13.

Zhao, Q.; Zheng, C.; Liu, M.; Wang, P.; and Chen, C. 2023. Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8877–8886.

Zhao, S.; Bai, L.; Shao, L.; Zhang, Y.; and Chen, X. 2025. SSD-Poser: Avatar Pose Estimation with State Space Duality from Sparse Observations. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, 1849–1857.

Zheng, X.; Su, Z.; Wen, C.; Xue, Z.; and Jin, X. 2023. Realistic full-body tracking from sparse observations via joint-level modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14678–14688.

Zheng, Y.; Huang, H.; Wang, X.; Yan, X.; and Xu, L. 2024. Spatio-temporal fusion for human action recognition via joint trajectory graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 7579–7587.