HiPoser: 3D Human Pose Estimation with Hierarchical Shared Learning at Parts-Level Using Inertial Measurement Units

Guorui Liao[1], Chunyuan Zheng[2], Li Cheng[3], Haoyu Xie[1], Shanshan Huang[1], Jun Liao[1], Haoxuan Li[4], Li Liu[1,*]

[1]School of Big Data & Software Engineering, Chongqing University, China
[2]School of Mathematical Sciences, Peking University, China
[3]Department of Electrical and Computer Engineering, University of Alberta, Canada
[4]Center for Data Science, Peking University, China
{guoruiliao,haoyuxie}@stu.cqu.edu.cn, {cyzheng,hxli}@stu.pku.edu.cn, lcheng5@ualberta.ca,
{shanshanhuang,liaojun,dcsliuli}@cqu.edu.cn

Abstract

This paper considers the challenging problem of 3D Human Pose Estimation (HPE) from a sparse set of Inertial Measurement Units (IMUs). Existing efforts typically reconstruct a pose sequence by either directly tackling whole-body motions or focusing on distinctive spatio-temporal features of local body parts. Unfortunately, these methods ignore existing interdependent motor synergies amongst body parts, which may lead to pose estimation with ambiguous local parts. This observation motivates us to propose a hierarchical learning-based approach, HiPoser, which utilizes a hierarchical shared structure using Mamba blocks as the backbone to focus on the following estimation tasks, involving: 1) torso pose, 2) lower limbs pose, 3) upper limbs pose, and finally 4) global translation. These tasks selectively incorporate body motion states and are to be carried out sequentially in reconstructing part-based poses, which are amalgamated to estimate the final full-body pose with the global translation that satisfies inter-part consistencies. Our hierarchical structure allows HiPoser the flexibility in prioritizing different aspects of pose estimation, to emphasize more on detail or stability. Empirical evaluations over three benchmark datasets demonstrate the superiority of HiPoser over existing state-of-the-art models, suggesting that analyzing the synergistic movement of body parts is indeed important for advancing IMU-based 3D HPE.

Introduction

3D Human Pose Estimation (HPE) is the process of determining the human joint positions in a three-dimensional coordinate system using motion signals. It is crucial in various real-world applications, including somatosensory gaming (Lai, Lu, and Bi 2024), competitive sports (Tanaka et al. 2023), medical rehabilitation (Gu et al. 2023), and emergency rescue (Yogesh et al. 2023). These applications require the analysis of human pose transformations (Zheng et al. 2023; Desmarais et al. 2021). In capturing motion signals, compared to vision sensors (Li et al. 2023; Li, Liu, and Wu 2023; Jiang et al. 2025) and environmental sensors (Lee et al. 2023; Zhou et al. 2023), wearable sensors, like Inertial Measurement Units (IMUs) (Mollyn et al. 2023), are virtually unaffected by environmental factors like object occlusion, poor lighting, and space constraints. Additionally, they provide privacy protection for the user.

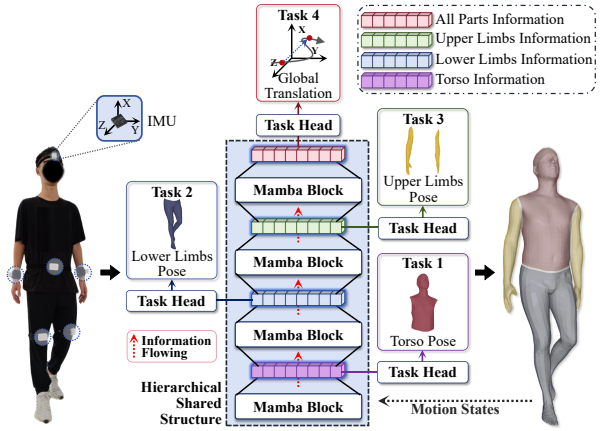

Figure 1: An overview of our Mamba-based hierarchical shared learning approach, which decomposes 3D HPE into four tasks, each focusing on a specific body part motion. By facilitating the flowing of body parts motion information among different tasks and incorporating motion states, our approach is empirically shown to produce stable results.

Early works (Von Marcard et al. 2017; Huang et al. 2018; Jiang et al. 2022a) attempted to achieve whole-body pose reconstruction using a sparse IMUs configuration to capture dynamic human motion information have proven to be challenging. Other complex network models (Yi, Zhou, and Xu 2021; Yi et al. 2022) achieved acceptable performance of whole-body pose estimation and found that the information on the leaf joints is beneficial for reconstructing the human pose. However, the contribution of articulated leaf joints located in different limbs remains ambiguous.

The latest work (Zhang et al. 2024a) aimed to address these ambiguities by classifying the leaf joints into three parts: torso, lower limbs, and upper limbs, and achieved state-of-the-art (SOTA) results. Nevertheless, their modeling may lead to negative motion information transfer among

*Corresponding author

different parts poses reconstruction (Tang and Wu 2019). Specifically, this work used a shared layer that extracted whole-body information to estimate body parts poses in parallel. However, different parts have different motion features, which may potentially lead to conflicts among other parts in pose reconstruction. Besides, this approach treated each task as equally important, inherently repeating smaller HPE tasks, and did not make explicit the kinematic correlations among different parts in a single HPE. A more reasonable approach should maintain balance during movement, the torso, lower limbs, and upper limbs will synergize with each other to counteract the effects of inertia caused by the movement of any single part of the body.

To this end, we propose a hierarchical shared learning approach for 3D HPE, HiPoser, which is shown in Figure 1. It learns the motion information and characteristics of the different parts sequentially, and finally performs local parts poses estimation as well as global translation estimation on the human body through different heads. Specifically, we decompose the 3D HPE process into four estimation tasks: 1) torso pose, 2) lower limbs pose, 3) upper limbs pose, and 4) global translation, and design a hierarchical shared structure to facilitate the learning and sharing of features among the different body parts. In addition, we utilize Mamba (Dao and Gu 2024) as the backbone, which allows us to selectively extract or omit motion information based on the current input motion deep-level features, resulting in improved performance during long action processes. We also incorporate the motion states (*i.e.*, joints poses and body position), into the hierarchical shared structure for each action window, enhancing the stable performance of HiPoser during the movement. Considering the motion synergies of body parts, HiPoser is able to adjust the task sequence to achieve varying levels of prioritized performance (*e.g.*, detail priority and stability priority).

Overall, our contributions are as follows:

- We propose a novel hierarchical shared learning-based approach, HiPoser, which extracts and shares the underlying motion features of different body parts (*i.e.*, torso, lower limbs, and upper limbs), sequentially reconstructing the body parts poses and estimating whole-body translation in the global consistency, achieving different priorities on 3D HPE based on different task sequences.
- We are the first to adopt a Mamba-based network in 3D human pose estimation. By incorporating motion states, our HiPoser can effectively store the history information of different body parts and extract potential motion features for more fine-grained pose reconstruction.
- Extensive experiments on AMASS, DIP-IMU, and TotalCapture show that our HiPoser is significantly better than the competitors. Further, we have conducted sufficient experiments on various details of HiPoser to confirm its superiority.

Related Work

IMU-based 3D Human Pose Estimation

The deployment position and number of different IMUs may affect human pose reconstruction. For commercial purposes, Schepers et al. (2018) used 17 IMUs to reconstruct the whole-body pose. However, such a dense IMUs configuration can disturb the wearing sensations of users and interfere with mobility. To evaluate the effectiveness of the sparse configuration, DIP (Huang et al. 2018) regressed 6 IMU measurements to the pose parameters of human joints. Nevertheless, DIP remains ambiguous in describing the movement relationships of the human body. Underlying the same configuration, TransPose (Yi, Zhou, and Xu 2021) used a multi-stage framework based on Bidirectional Recurrent Neural Networks (Bi-RNN), which consciously estimated the position of leaf joints before regressing to the full body pose and achieved better performance. The fact that TransPose performed pose estimation in the sequence of the human joints chain illustrated that the motion of the leaf joints contributes to the global human pose. Unfortunately, this utility remained ambiguous for different body parts in which the leaf joints are located. PIP (Yi et al. 2022), based on TransPose, considered inertial effects under physical constraints to improve the stability of the model. TIP (Jiang et al. 2022b) introduced the concept of stationary body points with zero velocity and used a Transformer structure to alleviate the estimation drift problem. However, none of these works have effectively utilized the motion correlation of different body parts, which may lead to difficulties in reconstructing the movement details of body parts. In this paper, we decompose the whole body into different body parts for pose reconstruction and consider using possible synergistic relationships among body parts to achieve more stable results with global consistency.

Multi-Task Learning on Human Pose Estimation

We present the design of whole-body pose in a hierarchical approach as a task decomposition in Multi-Task Learning (MTL), which allows the model to learn and share basic features extracted from multiple related tasks, facilitating the model performance. Luvizon, Picard, and Tabia (2020) used videos to perform 2D/3D HPE and action recognition, both of which obtained performance improvements compared to performing a single task. Burgermeister and Curio (2022) designed an MTL-based method for 3D human pose estimation and orientation estimation, achieving competitive performance. Shi et al. (2022) explored the kinematic contribution of the joints while estimating the whole-body pose, which greatly improved the performance. The success of these works may be due to the sets of tasks mutually reinforcing the optimization direction of their models. The latest work, DynaIP (Zhang et al. 2024a), can also be considered as an MTL-based approach. They utilized RNN to extract global motion information from IMUs and shared them across three body parts pose estimations, achieving SOTA performance. However, different parts have different motion paradigms, which may cause the tasks to conflict with each other and degrade the model performance. For example, when driving lower limbs to run, upper limbs will move in concert to maintain balance, whereas it can also move independently of the lower limbs. Simultaneously, different parts are synergistic during the movement, so simply sharing body information to reconstruct parts poses may lead to

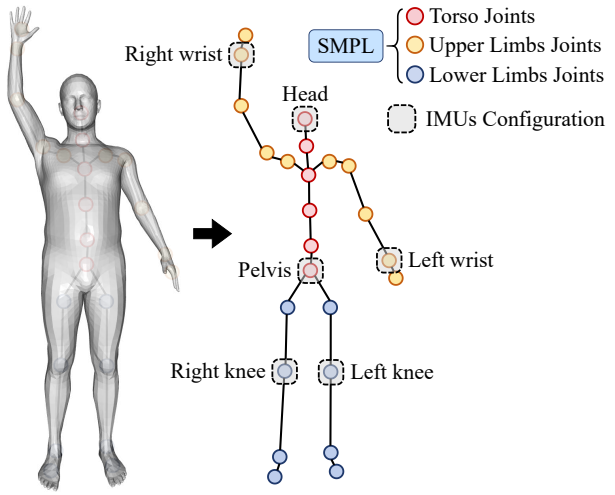

Figure 2: Our IMUs configuration is aligned with the SMPL joint positions to visualize the human pose.

unstable performance. Therefore, we consider utilizing the hierarchical structure to implement the sequential motion information sharing of body parts, which avoids simultaneous influence, effectively alleviating the problem of task conflict. Moreover, the different estimation task sequences of body parts allow HiPoser to achieve different priorities on human pose reconstruction.

Mamba and Selective State Space Models

Mamba (Dao and Gu 2024) is a linear time series model that can be computed in parallel at runtime, reducing memory requirements and improving efficiency. It possesses Selective State Space Models (SSM) that more efficiently capture relevant information over long periods. Recently, extensive research has been done on Mamba to explore its possibilities in various fields such as vision (Chen et al. 2024; Zhu et al. 2024), speech (Jiang, Han, and Mesgarani 2024), large language models (Zhao et al. 2024) and context generation (Zhang et al. 2024b). In this paper, we use Mamba blocks as the backbone for HiPoser, which may have advantages in reconstructing human poses during long-term movement.

Preliminaries

The Skinned Multi-Person Linear (SMPL) (Loper et al. 2023) model is capable of rendering realistic human shape due to its linear hybrid skeletal skinning technique, and describes joint motions with rotation matrices to represent dynamic human pose effectively. Our IMUs configuration follows the works of Huang et al. (2018); Yi, Zhou, and Xu (2021); Yi et al. (2022), aligning with the corresponding joint positions of the SMPL to efficiently capture the poses of body parts during the movement, as shown in Figure 2.

Problem Formulation

Given a dataset $\mathcal{D}$ containing $\mathcal{N}$ samples of action events to $\mathcal{S}$ IMUs recording motion changes in human pose. Each sample is a sequence of T frames collected at a uniform time scale. Formally, we align the IMU measurements to the spine root joint and perform normalization and splicing operations to obtain the input signal $X \in \mathbb{R}^{\mathcal{S}\times 12} = \{x_A^{(1)}, x_A^{(2)}, \ldots, x_A^{(\mathcal{S})}, x_R^{(1)}, x_R^{(2)}, \ldots, x_R^{(\mathcal{S})}\}$, where $x_A \in \mathbb{R}^3$ denotes the 3-axis acceleration measurements, $x_R \in \mathbb{R}^{3\times 3}$ denotes the rotation matrix of the selected joints in three-dimensional space. The output $Y = \{Y_{Pose}, Y_{Trans}\}$, where $Y_{Pose} \in \mathbb{R}^{24\times 3\times 3}$ is the rotation matrix integrating the 24 joints of SMPL, $Y_{Trans} \in \mathbb{R}^3$ denotes the position under the global 3-axis coordinate system. Our model estimates $X(t)$ to get the output $Y(t)$ of t-th frame, where $t = \{1, \ldots, T\}$.

Methodology

The overall architecture of HiPoser is shown in Figure 3. We design a hierarchical shared structure so that motion features from the torso, lower limbs, and upper limbs can be shared at different hierarchies. Mamba blocks as the backbone of the hierarchical shared structure are able to infer the correct pose during motion based on the input of the current motion information and the motion states, which is crucial for reducing the pose accumulation error in IMU-based 3D HPE problems. Lastly, simple linear functions acting as task heads can efficiently reconstruct the pose of each body part and the translation of the whole body. HiPoser with a hierarchical shared structure allows flexible setting of task sequence, achieving different priorities of 3D HPE.

Mamba on Action Process

Previous works have chosen RNN (Zhang et al. 2024a), Bi-RNN (Huang et al. 2018; Yi, Zhou, and Xu 2021; Yi et al. 2022), or Transformer (Jiang et al. 2022b) to capture motion-related feature. However, these methods may not be effective solutions for IMU-based 3D HPE in terms of resource consumption and historical inference.

To satisfy the requirement of both efficient motion pose inference capability and capturing all the necessary information from context, we consider employing Mamba (Dao and Gu 2024) blocks based on selective SSM (S-SSM) as the backbone of a hierarchical shared structure. Specifically, S-SSM allows selective processing of motion features and efficiently memorizes information about the movement over a long period. We set S_t to be the state variable of frame t in the shared layer, which is formulated as follows:

$$S_t = A_{(\Delta)} S_{t-1} + B_{(\Delta)} h_t, \tag{1}$$

$$h_{t+1} = C S_t, \tag{2}$$

where h denotes the hidden feature in S-SSM. Δ is the time step and is computed on a continuous time scale for the discretized matrices $A_{(\Delta)}$ and $B_{(\Delta)}$. The input matrix $B_{(\Delta)}$ acts directly on h_t, and $A_{(\Delta)}$ is a state matrix, which stores all historical motion information. The output matrix C defines a linear mapping relation from S_t to h_{t+1}, which does not require Δ for discretization.

Motion Information Store. Notably, $A_{(\Delta)}$ is significant in the modeling of action processes due to the importance of action state updates depending on the historical motion

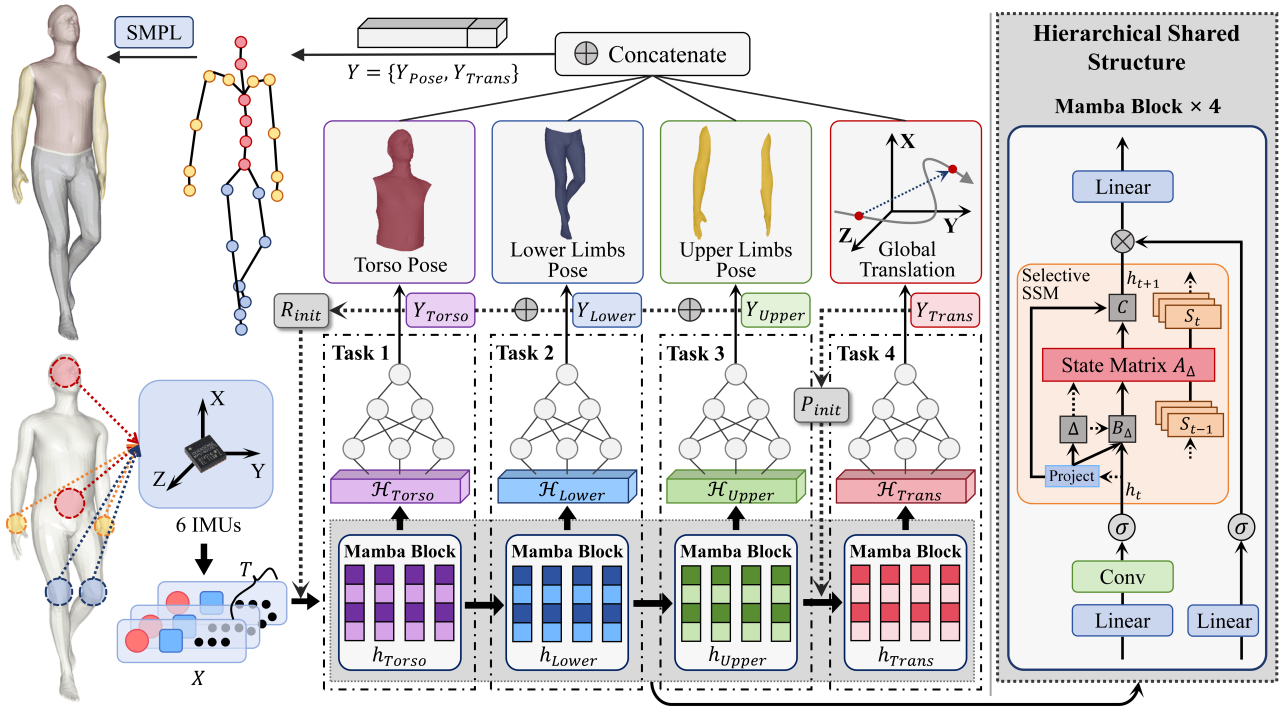

Figure 3: Overall architecture of HiPoser with task sequence: *Torso→ Lower limbs→ Upper limbs→ Translation*, which corresponds to the kinematic force generation and has a more stable effect in reconstructing the body parts poses. The hierarchical shared structure with Mamba blocks as the backbone is used to extract the motion information of each part. Different tasks and motion states are hierarchically learned and shared to finally reconstruct the whole human pose.

information it stores. To efficiently solve the long-range dependency problem of action process modeling in limited storage space, $A_{(\Delta)}$ uses a High-order Polynomial Projection Operator to compress all the current input information from S_{t-1}, which is defined as follows:

$$A_{(\Delta)}(i,j) = \begin{cases} 0, & \text{if } i < j; \\ i+1, & \text{if } i = j; \\ 2i+1, & \text{if } i > j, \end{cases} \tag{3}$$

where $A_{(\Delta)}$ is a matrix vector of coefficients. Remote dependencies can be handled by computing $A_{(\Delta)}S_{t-1}$ to infinitely approximate all historical states up to moment t.

Selective Mechanism. Practically, the model parameters will remain unchanged during the motion inference, which will lead to the fact that different motion features will combine with the same state matrix $A_{(\Delta)}$ for computation. Therefore, the model will lose the ability to make targeted inferences about the current human pose and historical action information. To address the problem of the irrelevance of the motion input to the state space, we use a simple selection mechanism that dynamically computes the parameters p of $B_{(\Delta)}$, C, and Δ, which is calculated as follows:

$$p_{B_{(\Delta)}}(t) = Linear_{B_{(\Delta)}}(h_t), \tag{4}$$
$$p_C(t) = Linear_C(h_t), \tag{5}$$
$$p_\Delta(t) = Linear_\Delta(h_t), \tag{6}$$

where h_t is projected through $Linear$, and the parameters $p_{B_{(\Delta)}}$, p_C and p_Δ are determined during training. Indeed, the state matrix $A_{(\Delta)}$, storing historical motion information, is discretized by Δ and obtains the new state S_t with $B_{(\Delta)}$, thus $A_{(\Delta)}$ achieves data dependent in a parameter efficient way as well, *i.e.*, $A_{(\Delta)}$ can generate new states S_t with h_t, enabling our model to implement pose inference selectively.

Mamba Block. To capture as much necessary information from the context as possible, a complete Mamba Block (MB) can be formulated as follows:

$$Cap(\mathcal{X}) = S\text{-}SSM(\sigma(Conv(Linear(\mathcal{X})))), \tag{7}$$
$$MB(\mathcal{X}) = Linear(Cap(\mathcal{X}) \times \sigma(Linear(\mathcal{X}))), \tag{8}$$

where $\mathcal{X}$ is the input of the current hierarchy, and Cap captures motion feature from S-SSM. σ denotes the $SiLU$ function. $Linear$ function increases the dimensionality of $\mathcal{X}$ to capture more detailed and complex motion features in a higher dimensional solution space. Convolution operation $Conv$ enhances MB to capture localized motion features at short distances, complementing the S-SSM to capture long-term dependencies, forming a complex representation of the motion information.

Hierarchical Shared Learning

In our hierarchical shared structure, implementing shallow tasks can enhance our model to utilize more basic motion information of body parts related to motion when performing

deeper tasks. Compared to the direct regression of whole-body pose using IMU measurements (Huang et al. 2018) or using the information from body parts-level movement (Yi, Zhou, and Xu 2021; Yi et al. 2022; Zhang et al. 2024a), we address the motion ambiguity of human body parts and sequentially exploit the correlation of part-level tasks in a more fine-grained way of hierarchical shared learning, which can further prevent the problem of conflicting estimation tasks and improve the performance of 3D HPE in a targeted manner based on the underlying features of different body parts. In addition, we introduce the motion states so that the model has a more stabilizing effect in reconstructing the parts poses during random and complex motions.

Task Determination. To enable hierarchical shared learning of parts motion information at different levels, we categorize 3D HPE into four tasks, which are torso (6 joints) pose estimation $Y_{Torse} \in \mathbb{R}^{6\times9}$, lower limbs (8 joints) pose estimation $Y_{Lower} \in \mathbb{R}^{8\times9}$, upper limbs (10 joints) pose estimation $Y_{Upper} \in \mathbb{R}^{10\times9}$ (*i.e.*, $Y_{Pose} = \{Y_{Torse}, Y_{Lower}, Y_{Upper}\}$); and 3-axis body translation estimation $Y_{Trans} \in \mathbb{R}^3$. These tasks can sequentially acquire information from the front task and pass backward in a hierarchical shared structure, enabling flow sharing.

Motion States. Determining the starting position and the magnitude of pose changes of the human body can effectively improve the stability of the model on subsequent motion pose estimation tasks (Zhang et al. 2024a; Yi, Zhou, and Xu 2021). To obtain more details of the human body in the global coordinate system, we extract the rotation matrix R_{init} and global position P_{init} of the human body to represent the pose state and position state of the human body, which are defined as follows:

$$R_{init}(t) = \begin{cases} 0, & \text{if } \ 0 \le t \le T; \\ Y_{Pose}(kT-1), & \text{if } kT < t \le (k+1)T, \end{cases} \tag{9}$$

$$P_{init}(t) = \begin{cases} 0, & \text{if } \ 0 \le t \le T; \\ Y_{Trans}(kT-1), & \text{if } kT < t \le (k+1)T, \end{cases} \tag{10}$$

where $k \in \mathbb{N}^+ = \{1, 2, \ldots, \mathcal{N}-1\}$. All joints are aligned to the spine joint and P_{init} and R_{init} are set to the value of 0 at first. After that P_{init} and R_{init} are determined by Y_{Pose} and Y_{Trans} of the last frame of sequences, respectively.

Hierarchical Shared Structure. Considering that recent works (Jiang et al. 2022b; Yi et al. 2022; Zhang et al. 2024a) have focused more on pose reconstruction, we tentatively use task sequence *Torso*→ *Lower*→ *Upper*→ *Trans* as an example that can better capture the motion details of body parts. Therefore, our pipeline of hierarchical shared structure can be formulated as follows:

$$\mathcal{H}_{Torso} = MB(X \oplus R_{init}), \tag{11}$$

$$\mathcal{H}_{Lower} = MB(\mathcal{H}_{Torso}), \tag{12}$$

$$\mathcal{H}_{Upper} = MB(\mathcal{H}_{Lower}), \tag{13}$$

$$\mathcal{H}_{Trans} = MB(\mathcal{H}_{Upper} \oplus P_{init}), \tag{14}$$

where $\mathcal{H}_n$ is the motion feature at the shared hierarchy of task $n \in \{Torso, Lower, Upper, Trans\}$, and $\oplus$ is concatenation. Since R_{init} represents the local movement variations of the human joints, and P_{init} represents the relative global position of the human body, we selectively place R_{init} before the parts poses estimation and P_{init} before the translation estimation, which facilitates our model to maximize the extraction of critical information and does not have local or global impacts.

Loss Function. Lastly, the estimation Y'_n of each task n is calculated as follows:

$$Y'_n(t) = TaskHead(\mathcal{H}_n(t)), \tag{15}$$

where $TaskHead$ is just a $Linear$ layer, though it is capable of competently regressing the parts pose Y_{Pose} and global translation Y_{Trans} from $\mathcal{H}$. Therefore, the global loss $\mathcal{L}_{global}$ can be formulated as:

$$\mathcal{L}_{global} = \frac{1}{T}\sum_{n}\sum_{t=1}^{T} \|Y'_n(t) - Y_n(t)\|_2. \tag{16}$$

Experiments

Experimental Setup

Datasets. We use AMASS (Mahmood et al. 2019), DIP-IMU (Huang et al. 2018), and TotalCapture (Trumble et al. 2017) datasets to evaluate the effectiveness of models.

Baselines. We use DIP (Huang et al. 2018), TransPose (Yi, Zhou, and Xu 2021), TIP (Jiang et al. 2022b), PIP (Yi et al. 2022), DynaIP (Zhang et al. 2024a) as baselines to demonstrate the superiority of HiPoser.

Metrics. Following TransPose (Yi, Zhou, and Xu 2021), PIP (Yi et al. 2022), and DynaIP (Zhang et al. 2024a), we evaluate the estimation performance by metrics: 1) *SIP error*(°): the mean global rotation error of upper arms and upper legs; 2) *Ang error*(°): the mean global rotation error of all human joints; 3) *Pos error*(cm): the mean Euclidean distance error of all joints; 4) *Mesh error*(cm): the mean Euclidean distance error of all vertices of body mesh; 5) *Dist error*(cm): the mean Euclidean distance error of translation of whole body.

Implement Details. Following TransPose (Yi, Zhou, and Xu 2021), PIP (Yi et al. 2022), and IMUPoser (Mollyn et al. 2023), we fix the sampling frequency to 60Hz for all the datasets and set frame $T = 300$. The translation estimation is calculated by combining the velocity and the contact-ground probability of feet in the offline setting. We build our model using PyTorch and PyTorch Lightning and use the $Adam$ optimizer with a learning rate of 1e-3 to update the parameters. The training epoch is 500. The batch size is 256. The training and validation process is implemented on an NVIDIA GeForce RTX 4090 GPU.

Quantitative Results

Since DIP-IMU and TotalCapture are relatively small, we only train on AMASS and validate on DIP-IMU and TotalCapture to evaluate the generalization performance of

Methods	DIP-IMU					TotalCapture				
	SIP(°)	*Ang*(°)	*Pos*(cm)	*Mesh*(cm)	*Dist*(cm)	*SIP*(°)	*Ang*(°)	*Pos*(cm)	*Mesh*(cm)	*Dist*(cm)
DIP	18.35	10.04	7.55	8.35	55.74	18.95	10.76	7.79	9.07	59.35
TransPose	17.24	9.45	6.82	7.61	43.12	17.81	9.83	7.31	8.29	45.76
TIP	16.58	9.31	6.60	7.42	34.50	16.66	9.39	7.08	7.89	41.37
PIP	15.72	8.84	6.45	7.28	33.86	16.25	9.28	7.03	8.11	39.11
DynaIP	14.97	8.57	5.69	6.50	33.95	15.17	8.55	5.96	6.78	40.63
HiPoser (ours)	**14.11**	**8.16**	**5.44**	**6.30**	**32.48**	**14.53**	**8.35**	**5.63**	**6.61**	**36.61**

Table 1: Performance comparison on the state-of-the-art models on DIP-IMU and TotalCapture when trained with AMASS.

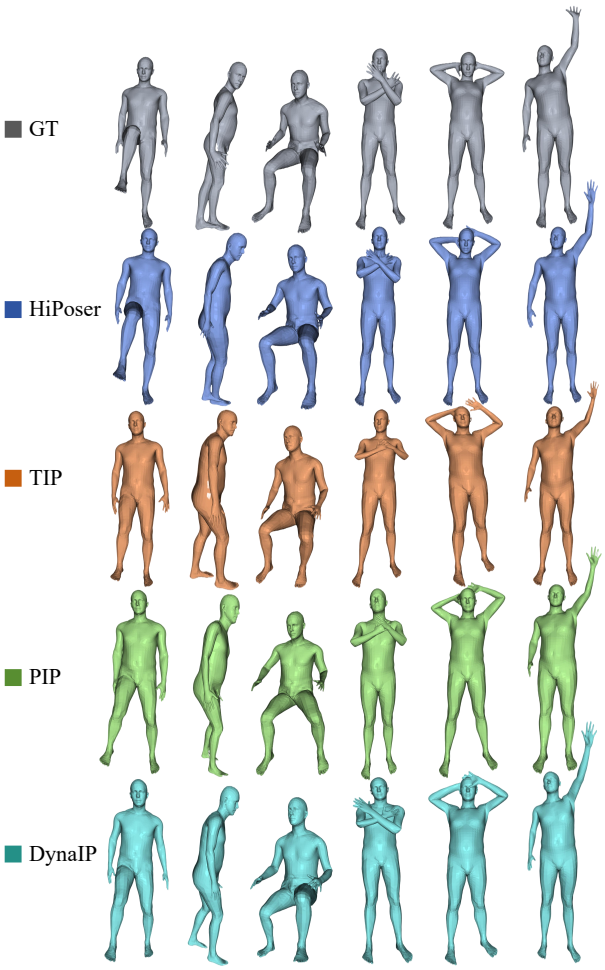

Figure 4: Qualitative comparison of HiPoser with SOTAs.

HiPoser. We compare the mean values of baselines on each metric. As shown in Table 1, our HiPoser performs better than other methods. We argue that the superiority of HiPoser lies in the hierarchical shared learning structure, which can effectively utilize the information underlying the movement of different body parts to estimate joint rotations better. Moreover, the involvement of motion states helps to maintain the continuity of motion patterns. Figure 4 shows the estimation results of HiPoser and SOTAs.

Methods	*SIP*(°)	*Ang*(°)	*Pos*(cm)	*Mesh*(cm)	*Dist*(cm)
HiPoser	**14.11**	**8.16**	**5.44**	**6.30**	**32.48**
w/o Hier	14.23	8.20	5.51	6.57	37.81
w/o Parts	14.42	8.24	5.81	6.62	35.49
Baseline	14.90	8.41	6.02	6.91	42.48

Table 2: Comparison of with&w/o hierarchical shared learning and body parts on HiPoser on DIP-IMU.

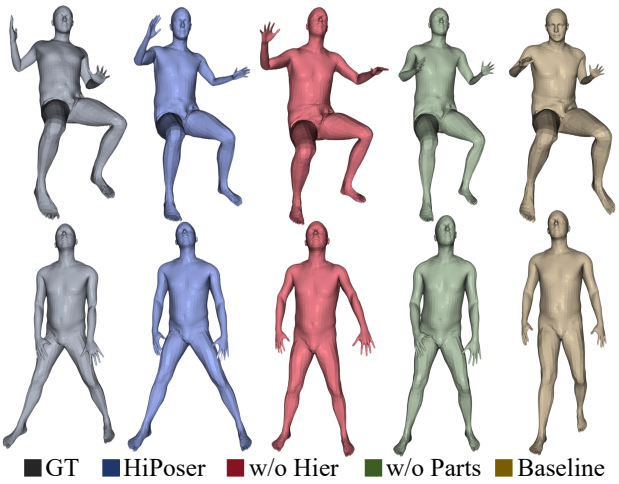

Figure 5: Visualization of with&w/o hierarchical shared learning and body parts of HiPoser on DIP-IMU.

In-Depth Study

Effect on Hierarchical Sharing. To evaluate the effectiveness of the hierarchical shared structure, we compare the following variants to reconstruct human pose: 1) w/o Hierarchical: using global and body parts information with parallel partition modeling; 2) w/o Parts: using global and body parts information but w/o partition modeling; 3) Baseline: a Mamba-based network using body information. The comparison results are shown in Table 2 which indicates reconstructing parts poses based on motion information is effective, probably because the model can obtain stable results on smaller estimations. Besides, sequentially sharing allows the model to better reconstruct poses for different movements, which shows the rationality of setting up different task se-

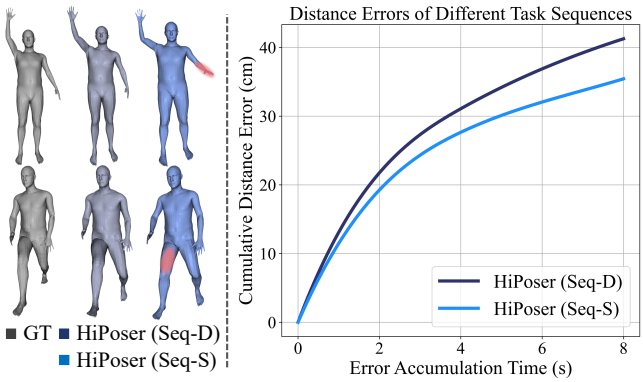

Figure 6: Comparison of task sequences Seq-D and Seq-S on DIP-IMU. The red region indicates that some details of Seq-S are weaker than Seq-D on pose reconstruction.

Methods	*SIP*(°)	*Ang*(°)	*Pos*(cm)	*Mesh*(cm)	*Dist*(cm)
Seq-D	**14.11**	**8.16**	5.44	6.30	32.48
Seq-S	14.42	8.24	**5.26**	**6.13**	**29.75**

Table 3: Performance of HiPoser with different task sequences Seq-D and Seq-S on DIP-IMU.

quences consistent with motion associations. The visualization is shown in Figure 5 which indicates that these variants can collectively improve the performance of HiPoser.

Effect on Task Sequence. We set two representative task sequences, *i.e.*, Seq-D: *Torso*→ *Lower*→ *Upper*→ *Trans*, which denotes detail priority, and Seq-S: *Upper*→ *Lower*→ *Torso*→ *Trans*, which denotes stability priority. The translation estimation task is placed at the deepest level because it measures the change in the position of the human body in global coordinates, laying the groundwork for pose reconstruction. Figure 6 illustrates the visualization of pose reconstruction as well as translation estimation errors for Seq-D and Seq-S. Table 3 and Figure 6 indicate that Seq-D is more advantageous for limbs pose reconstruction, though having a weaker performance of translation estimation than Seq-S in long-term movement. This may be due to the fact that Seq-D possesses more information from the torso movement in predicting upper and lower limbs pose changes, which facilitates the delimitation of the range of parts motion. While Seq-S simplifies the complexity of the estimation task by characterizing the range of body limbs movement, making the body position more stable in the global representation.

Effect on Motion States. To evaluate the effectiveness of the motion states, we set R_{init} of different numbers of joints and P_{init} in our HiPoser by the following variants: 1) All: R_{init} of all joints and P_{init}; 2) Config: R_{init} of Configured joints and P_{init}; 3) Single: only P_{init}; 4) None: without any motion state. The results are shown in Table 4, which illustrates that as the number of joint points within the state increases, more fine-grained motion information can be captured by our model, achieving a better reconstruction of the human pose. Without inputting any state information into the variants, the model is greatly limited in its ability to reason, possibly due to the memory of movement being important for the human body regarding motor coherence.

Methods	*SIP*(°)	*Ang*(°)	*Pos*(cm)	*Mesh*(cm)	*Dist*(cm)
All	**14.11**	**8.16**	**5.44**	**6.30**	**32.48**
Config	14.75	8.48	5.61	6.54	34.12
Single	15.21	8.62	5.87	6.62	39.28
None	15.60	8.76	6.32	7.22	47.61

Table 4: Performance of HiPoser (Seq-D) with&w/o R_{init} of different numbers of joints and P_{init} on DIP-IMU.

Metrics		Mamba	RNN	Bi-RNN	Transformer
5s	*SIP*(°)	**14.11**	14.52	14.40	14.90
	Pos(cm)	**5.44**	5.64	5.72	5.79
	Dist(cm)	**32.48**	33.97	36.32	41.27
	RT(s)	**0.004**	0.011	0.012	0.006
20s	*SIP*(°)	**13.66**	14.12	13.98	14.45
	Pos(cm)	**5.08**	5.16	5.32	5.86
	Dist(cm)	**62.43**	75.68	84.79	89.61
	RT(s)	**0.012**	0.024	0.029	0.015
FLOPs(G)		5.77	58.80	141.21	25.65
Parameters		116.76K	1.17M	2.82M	519.20K

Table 5: Comparison performance of different basic models as the backbone of HiPoser (Seq-D) on DIP-IMU. RT: Inference runtime. FLOPs: Floating points per second. Parameters: Number of model parameters.

Effect on Mamba Block. To evaluate the effectiveness of Mamba, we replace the backbone of the hierarchical shared structure with different basic models (RNN, Bi-RNN, and Transformer) and set the same hidden dimensions as the MBs. Table 5 shows the performance of different backbones of HiPoser during long-time pose inference, and the corresponding runtime and space occupation. The results indicate that Mamba is able to perform more stably during long-time action, which may be due to its unique selective mechanism that efficiently utilizes historical motion information. The small resource consumption of Mamba is also a great advantage in its efficient inference ability.

Conclusion

In this paper, we propose a hierarchical shared learning-based approach, HiPoser, which addresses 3D HPE from a body parts level. It considers implementing the sequential sharing of the motion information among different body parts and incorporating motion states, achieving a stable performance in long-term movement. This flexible hierarchical shared structure enables our HiPoser to achieve different prioritized requirements of human pose reconstruction. Extensive experiments demonstrate that HiPoser significantly outperforms existing methods in all performance metrics.

Acknowledgments

This work was supported by grants from the National Natural Science Foundation of China (grant Nos. 62477004, 62377040, 62207007, 623B2002) and Chongqing Natural Science Foundation Innovation and Development Joint Fund (grant. No. CSTB2023NSCQ-LZX0109).

References

Burgermeister, D.; and Curio, C. 2022. PedRecNet: Multi-task deep neural network for full 3D human pose and orientation estimation. In *IV*. IEEE.

Chen, K.; Chen, B.; Liu, C.; Li, W.; Zou, Z.; and Shi, Z. 2024. Rsmamba: Remote sensing image classification with state space model. *IEEE Geoscience and Remote Sensing Letters*.

Dao, T.; and Gu, A. 2024. Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. In *ICML*.

Desmarais, Y.; Mottet, D.; Slangen, P.; and Montesinos, P. 2021. A review of 3D human pose estimation algorithms for markerless motion capture. *Computer Vision and Image Understanding*, 103275.

Gu, C.; Lin, W.; He, X.; Zhang, L.; and Zhang, M. 2023. IMU-based motion capture system for rehabilitation applications: A systematic review. *Biomimetic Intelligence and Robotics*, 100097.

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*, 1–15.

Jiang, J.; Streli, P.; Meier, M.; and Holz, C. 2025. EgoPoser: Robust Real-Time Egocentric Pose Estimation from Sparse and Intermittent Observations Everywhere. In *ECCV*. Springer.

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 *ECCV*. Springer.

Jiang, X.; Han, C.; and Mesgarani, N. 2024. Dual-path mamba: Short and long-term bidirectional selective structured state space models for speech separation. *arXiv:2403.18257*.

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*.

Lai, W.; Lu, T.; and Bi, H. 2024. Design and implementation of an AR sensory fruit cutting system based on monocular video. In *EIBDCT*. SPIE.

Lee, S.-P.; Kini, N. P.; Peng, W.-H.; Ma, C.-W.; and Hwang, J.-N. 2023. Hupr: A benchmark for human pose estimation using millimeter wave radar. In *WACV*.

Li, H.; Shi, B.; Dai, W.; Zheng, H.; Wang, B.; Sun, Y.; Guo, M.; Li, C.; Zou, J.; and Xiong, H. 2023. Pose-oriented transformer with uncertainty-guided refinement for 2d-to-3d human pose estimation. In *AAAI*.

Li, J.; Liu, K.; and Wu, J. 2023. Ego-body pose estimation via ego-head pose estimation. In *CVPR*.

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.

Luvizon, D. C.; Picard, D.; and Tabia, H. 2020. Multi-task deep learning for real-time 3D human pose estimation and action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2752–2764.

Mahmood, N.; Ghorbani, N.; Troje, N. F.; Pons-Moll, G.; and Black, M. J. 2019. AMASS: Archive of motion capture as surface shapes. In *ICCV*.

Mollyn, V.; Arakawa, R.; Goel, M.; Harrison, C.; and Ahuja, K. 2023. Imuposer: Full-body pose estimation using imus in phones, watches, and earbuds. In *CHI*.

Schepers, M.; Giuberti, M.; Bellusci, G.; et al. 2018. Xsens MVN: Consistent tracking of human motion using inertial sensing. *Xsens Technol*, 1–8.

Shi, D.; Wei, X.; Li, L.; Ren, Y.; and Tan, W. 2022. End-to-end multi-person pose estimation with transformers. In *CVPR*.

Tanaka, R.; Suzuki, T.; Takeda, K.; and Fujii, K. 2023. Automatic edge error judgment in figure skating using 3d pose estimation from a monocular camera and imus. In *ACM MMSports*.

Tang, W.; and Wu, Y. 2019. Does learning specific features for related parts help human pose estimation? In *CVPR*.

Trumble, M.; Gilbert, A.; Malleson, C.; Hilton, A.; and Collomosse, J. P. 2017. Total capture: 3D human pose estimation fusing video and inertial sensors. In *BMVC*. London, UK.

Von Marcard, T.; Rosenhahn, B.; Black, M. J.; and Pons-Moll, G. 2017. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. *Computer Graphics Forum*.

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 *CVPR*.

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*, 1–13.

Yogesh, V.; Buurke, J. H.; Veltink, P. H.; and Baten, C. T. 2023. Integrated UWB/MIMU Sensor System for Position Estimation towards an Accurate Analysis of Human Movement: A Technical Review. *Sensors*, 7277.

Zhang, Y.; Xia, S.; Chu, L.; Yang, J.; Wu, Q.; and Pei, L. 2024a. Dynamic Inertial Poser (DynaIP): Part-Based Motion Dynamics Learning for Enhanced Human Pose Estimation with Sparse Inertial Sensors. In *CVPR*.

Zhang, Z.; Liu, A.; Reid, I.; Hartley, R.; Zhuang, B.; and Tang, H. 2024b. Motion mamba: Efficient and long sequence motion generation with hierarchical and bidirectional selective ssm. *arXiv:2403.07487*.

Zhao, H.; Zhang, M.; Zhao, W.; Ding, P.; Huang, S.; and Wang, D. 2024. Cobra: Extending mamba to multi-modal large language model for efficient inference. *arXiv:2403.14520*.

Zheng, C.; Wu, W.; Chen, C.; Yang, T.; Zhu, S.; Shen, J.; Kehtarnavaz, N.; and Shah, M. 2023. Deep learning-based human pose estimation: A survey. *ACM Computing Surveys*, 1–37.

Zhou, Y.; Huang, H.; Yuan, S.; Zou, H.; Xie, L.; and Yang, J. 2023. Metafi++: Wifi-enabled transformer-based human pose estimation for metaverse avatar simulation. *IEEE Internet of Things Journal*, 14128–14136.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv:2401.09417*.