This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

VIMCAN: Visual-Inertial 3D Human Pose Estimation with Hybrid Mamba-Cross-Attention Network

Zepeng Yang[1,2*] Junxuan Bai[3*] Hao Li[1,2] Ju Dai[2†] Junjun Pan[1,2] Yongfeng Yin[1†] Bin Li[4]
[1]Beihang University [2]Peng Cheng Laboratory
[3]Capital University of Physical Education and Sports
[4]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

Abstract

The rapid advances in deep learning have significantly enhanced the accuracy of multimodal 3D human pose estimation (HPE). However, the state-of-the-art (SOTA) HPE pipelines still rely on Transformers, whose quadratic complexity makes real-time processing for long sequences impractical. Mamba addresses this issue through selective state-space modeling, enabling efficient sequence processing without sacrificing representational power. Nevertheless, it struggles to capture complex spatial dependencies in multimodal settings. To bridge this gap, we propose VIMCAN, a hybrid architecture that combines the efficient sequence modeling of Mamba with the spatial reasoning of Cross-Attention, and performs robust visual–inertial fusion and human pose estimation between RGB keypoints and wearable IMU data. By leveraging Mamba's dynamic parameterization for temporal modeling and Attention for spatial dependency extraction, VIMCAN achieves superior accuracy, with mean per-joint position errors (MPJPE) of 17.2 mm on TotalCapture and 45.3 mm on 3DPW. VIMCAN outperforms prior Transformer-based and other SOTA approaches while supporting real-time inference at over 60 frames per second on consumer-grade hardware. The source code is available on https://github.com/Eddieyzp/VIMCAN.

1. Introduction

3D human pose estimation (HPE) is a noteworthy research topic in computer vision, encompassing both vision-based and multimodal approaches. In recent years, the development of deep neural networks has significantly improved the accuracy of vision-based HPE, with prominent examples including Transformer [15, 21, 26, 38, 39, 41], Graph Convolutional Network (GCN) [33, 37], and Temporal Convolutional Network (TCN) [18, 23]. However, vision-based HPE still faces two critical issues: (1) depth ambiguity when lifting 2D keypoints to 3D poses, and (2) inefficient computation when processing long sequences. To address the depth ambiguity problem, researchers have explored multimodal approaches by incorporating additional modalities alongside RGB images, such as depth data [13, 16, 24], inertial measurement unit (IMU) data [9, 11, 14, 27, 32], and Light Detection and Ranging (LiDAR) data [1, 31]. Thanks to its strong representation power, Transformer [26] has become a cornerstone for multimodal learning by leveraging Cross-Attention mechanism. Nevertheless, as shown in Fig. 1, the quadratic complexity $\mathcal{O}(L^2)$ of Attention with respect to sequence length L, leads to high computational and memory demands, thereby limiting the scalability and

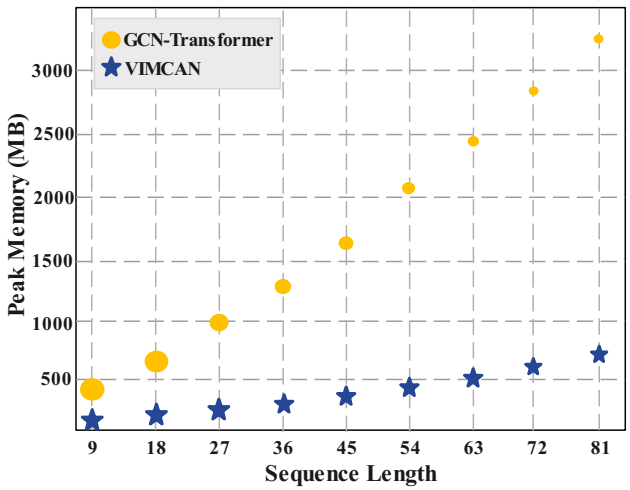


Figure 1. The comparison of peak memory usage during inference for a GCN-Transformer-based model [29] and the proposed VIMCAN. The peak memory usage at different lengths of input sequence. The x-axis denotes sequence length, and the y-axis represents peak memory usage (lower is better, in MB). The circles indicate the GCN-Transformer-based model, while the stars denote VIMCAN. The size of Symbol reflects the memory I/O throughput, illustrating computational efficiency during on-device inference.

[1]* Zepeng Yang and Junxuan Bai contributed equally to this work.
[2]† Corresponding authors: daij@pcl.ac.cn and yyf@buaa.edu.cn

applicability of these models in resource-constrained environments and high-throughput scenarios.

Mamba [4, 5] addresses the efficiency bottleneck by casting sequential data as a time-varying system and compressing historical context into a dynamically evolving hidden state, enabling recursive updates with linear complexity $\mathcal{O}(L)$. This feature facilitates efficient processing for long sequences, positioning Mamba as an alternative to Transformers. However, its spatial-reasoning capacity still lags behind that of convolution- or Attention-based approaches [34], limiting its utility in multimodal tasks that require precise spatial modeling.

To overcome the above limitations, we propose VIMCAN, a **V**isual-**I**nertial **M**amba-**C**ross-**A**ttention **N**etwork for human pose estimation. The hybrid design takes RGB keypoints and IMU data as input, leveraging Mamba's efficient sequence processing and Cross-Attention's aptitude for cross-modal fusion to achieve superior performance. Our approach strikes a balance between accuracy, robustness, and efficiency, enabling real-time inference exceeding 60 frames per second on consumer-grade hardware. Extensive experiments on TotalCapture and 3DPW validate the superiority of VIMCAN, with MPJPE reaching 17.2 mm and 45.3 mm, respectively, outperforming previous Transformer-based methods and other state-of-the-art (SOTA) approaches.

The main contributions include:

- We propose VIMCAN, a Mamba-Cross-Attention architecture that combines Mamba's efficient sequence processing with Cross-Attention's robust multimodal fusion, enabling real-time variable-length inference.
- To the best of our knowledge, VIMCAN is the first work to apply Mamba to vision-inertial multimodal 3D human pose estimation, addressing depth ambiguity and computational complexity.
- Extensive experiments demonstrate the superior accuracy, robustness, and efficiency of VIMCAN over SOTA methods.

2. Related Work

2.1. RGB-based 3D HPE

3D human pose estimation from RGB images or videos has been revolutionized by novel deep architectures, including GCN [33, 37], TCN [18, 23], and Transformer [15, 21, 26, 38, 39, 41]. GCN-based models effectively capture spatial dependencies and temporal dynamics for motion sequences on a skeletal graph, while TCN-based approaches utilize strided convolutions to extract long-range dependencies. Transformer-based networks, on the other hand, exploit attention mechanisms to learn global context relationships. For instance, PoseFormer [39] employs separate Transformer modules to encode intra-frame joint relations and inter-frame temporal dependencies. Similarly, KD-Former [3] divides the human body into kinematic groups and integrates kinematic and dynamic features via a non-autoregressive Transformer, enhancing overall representational capacity. These advances underscore the pivotal role of attention mechanisms and structural priors in achieving superior accuracy and generalization. Nevertheless, unimodal approaches are inherently limited by depth ambiguity in the 2D-to-3D lifting process.

2.2. Multimodal 3D HPE

Recovering 3D human pose from 2D observations is inherently ill-posed. Multimodal fusion methods with different sensors have enhanced the accuracy and robustness of 3D HPE, including RGB-D cameras [13, 16, 24], IMUs [9, 11, 27, 32], and LiDAR data [1, 31]. Such fusion exploits the complementary properties of sensors to overcome the limitations of single modalities. For instance, RGB cameras provide detailed visual context, while IMUs deliver high-frequency, low-latency dynamic data, thereby overcoming occlusion and blur issues inherent in visual systems [14, 36]. When performing multimodal fusion, Cross-Attention is frequently employed due to its ability to model intricate relationships between heterogeneous inputs [30]. Wang *et al*. [29] propose a GCN-Transformer hybrid architecture with Cross-Attention to fuse visual and IMU data, using kinematic priors for modality grouping and robust integration. Similarly, Liu *et al*. [17] utilize a pre-trained CNN to extract 2D image features as keys and values, with IMU features as queries, to regress Skinned Multi-Person Linear (SMPL) [20] parameters and root joint translation. However, their studies found that to achieve real-time inference with Attention, the temporal window length must be constrained. Limitations in computational efficiency, memory usage, and hardware compatibility hinder its application in scenarios involving long sequences and on resource-constrained devices.

2.3. Mamba Architectures

Mamba [5] emerges as a promising alternative to Transformer architectures, due to its linear time complexity in sequence modeling. Its efficacy in language processing has motivated its extension to visual representation learning. Zhu *et al*. [40] introduce Vision Mamba, which integrates bidirectional State Space Models (SSMs) with positional embeddings to overcome limitations in spatial modeling. Liu *et al*. [19] propose VMamba, featuring efficient spatio-temporal context encoding via a Cross-Scan module. Hu *et al*. [8] develope a multimodal video object tracking framework that leverages Mamba's temporal modeling via a dedicated temporal state generator, enabling cross-modal spatio-temporal feature extraction. In 3D HPE, Huang *et al*. [10] adopt a global-local spatio-temporal scanning mechanism

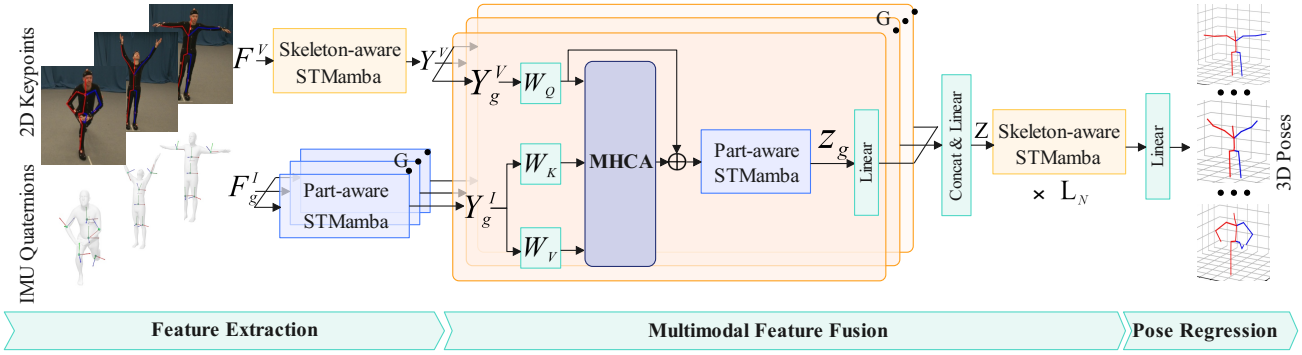


Figure 2. The framework of VIMCAN.

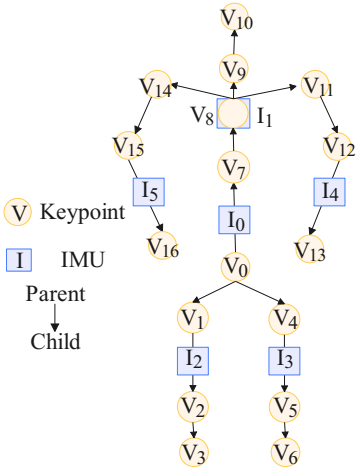
with linear complexity to exploit human skeletal geometry. Moreover, hybrid architectures combining Mamba with structural priors have been developed. Zhang *et al*. [35] introduce Pose Magic, a Mamba-GCN model which enhances local joint relationships while maintaining efficient long-range modeling and supports both bidirectional and causal reasoning. Despite these advances, an empirical study by Yu *et al*. [34] reveals that pure Mamba falls short of convolutional or attention-based methods in capturing intricate spatial relationships. To mitigate this, Hatamizadeh *et al*. [6] propose a Mamba-Transformer hybrid architecture that combines the efficiency of Mamba with the long-range dependency modeling capability of Attention, thereby establishing a benchmark for efficient multimodal 3D HPE.

3. Methodology

VIMCAN takes RGB-based keypoints and corresponding IMU data to predict 3D human poses. It integrates Mamba and Cross-Attention mechanisms to learn powerful features from multimodal inputs. As illustrated in Fig. 2, the framework comprises three main components: (1) Feature Extraction, (2) Multimodal Feature Fusion, and (3) Pose Regression.

3.1. Feature Extraction

As shown in Fig. 2, during feature extraction, VIMCAN learns features from visual images and IMU inputs. Specifically, VIMCAN receives J=17 keypoints coordinates from images and unit quaternions from I=6 IMUs as inputs. Drawing from kinematic priors [3, 29], the IMUs are partitioned into G=5 groups (torso, left arm, right arm, left leg, and right leg, shown in Fig. 3), each containing I_g IMUs. Both modalities are first lifted to a common embedding dimension D_e via linear layers, yielding visual feature $F^V \in \mathbb{R}^{T\times J\times D_e}$ and inertial features $F_g^I \in \mathbb{R}^{T\times I_g\times D_e}$. To efficiently capture spatial and temporal information, each modality is processed independently using Spatio-Temporal Mamba (STMamba) modules. Specifically, F^V is processed via skeleton-aware STMamba, while F_g^I from each



Group	Components
Torso	$V_{0,\,7,\,8,\,9,\,10}$
	$I_{0,\,1}$
Left Arm	$V_{0,\,7,\,8,\,11,\,12,\,13}$
	$I_{0,\,1,\,4}$
Right Arm	$V_{0,\,7,\,8,\,14,\,15,\,16}$
	$I_{0,\,1,\,5}$
Left Leg	$V_{0,\,4,\,5,\,6}$
	$I_{0,\,3}$
Right Leg	$V_{0,\,1,\,2,\,3}$
	$I_{0,\,2}$

Figure 3. Illustration of skeleton topology and group components.

body part g is handled by dedicated part-aware STMamba.

As shown in Fig. 4, each STMamba comprises two Bidirectional Spatio-Temporal State Space Model (BiSTSSM) blocks. The input is first processed spatially by a BiSTSSM to encode intra-frame topological relationships, followed by a temporal BiSTSSM to model inter-frame dynamics. The primary distinction between part-aware and skeleton-aware STMamba lies in the BiSTSSM scanning mechanism, tailored to each modality.

For the inertial branch, a bidirectional spatio-temporal selective scan (SS2D) [19] aggregates information across four directions: spatial forward/reverse and temporal forward/reverse. Regarding the visual branch, inspired by [10], keypoints are reordered according to parent-child joint relationships (defined in Fig. 3) during spatial forward scanning to incorporate kinematic priors.

Prior to spatial feature extraction, F^V and F_g^I are augmented with learnable spatial positional encodings and layer normalization (LN). Within each BiSTSSM, the input features F (either F^V or F_g^I) are projected to an internal SSM dimension using a linear layer (FC) to facilitate mod-

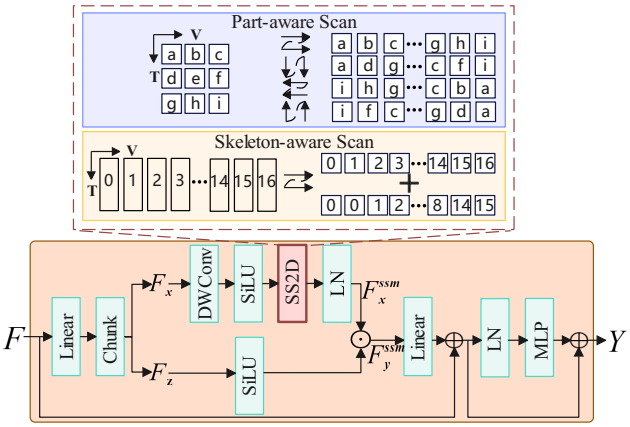


Figure 4. The architecture of BiSTSSM module.

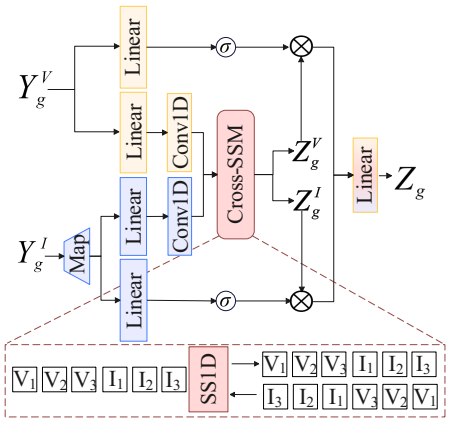


Figure 5. The architecture of Cross-Mamba module.

eling of complex spatio-temporal dependencies. The projected tensor is then divided (Chunk) into two components:

$$F_x, F_z = \text{Chunk}(\text{FC}(F)). \tag{1}$$

For F_x, a Depthwise Convolution (DWConv) layer is leveraged to capture local patterns, followed by an activation function (SiLU) to enhance local feature interactions. The scanning yields an intermediate tensor:

$$F_x^{ssm} = \text{LN}(\text{SS2D}(\sigma(\text{DWConv}(F_x)))). \tag{2}$$

Concurrently, F_z serves as a gating signal. The gated output is computed as:

$$F_y^{ssm} = F_x^{ssm} \cdot \sigma(F_z), \tag{3}$$

which is then projected back to D_e, combined with a residual connection, and processed by an MLP:

$$\begin{aligned} Y^{ssm} &= \text{FC}(F_y^{ssm}) + F, \\ Y &= \text{MLP}(\text{LN}(Y^{ssm})) + Y^{ssm}. \end{aligned} \tag{4}$$

Here, Y denotes the extracted visual feature Y^V or inertial feature Y_g^I for group g.

Subsequently, to model joint-wise spatio-temporal correlations within sliding windows, temporal positional encodings are added to Y, followed by a temporal BiSTSSM block that comprehensively captures inter-frame dependencies. The temporal scanning mechanism aligns with the spatial one, adapted per modality.

3.2. Multimodal Feature Fusion

Although Mamba's state-space model attains linear complexity $\mathcal{O}(L)$, its selective mechanism compromises its expressive capacity for efficiency, resulting in information loss that hinders cross-modal interactions in tasks that demand precise spatial reasoning [34]. Conversely, Attention mechanisms are utilized to aggregate information across all tokens, rendering cross-attention well-suited for capturing intricate cross-modal spatial dependencies [30].

To facilitate robust fusion of visual and inertial features, we introduce a Cross-Attention module as the backbone. The architecture of the module is depicted in the appendix.

The extracted visual features Y^V are partitioned into groups Y_g^V, each containing J_g joints. In the Cross-Attention module, Multi-Head Cross-Attention (MHCA) integrates the multimodal inputs. For each group g, visual features serve as queries Q_g^V, while inertial features yield keys K_g^I and values V_g^I. The outputs from each head h are concatenated and normalized, with a residual connection applied solely to the visual queries to retain skeletal information:

$$\begin{aligned} \text{MHCA} &= \text{Concat}\left[\text{Softmax}\left(\frac{Q_g^V {K_g^I}^\top}{\sqrt{d_k}}\right) V_g^I\right]_h, \\ Z_g &= \text{LN}(\text{MHCA}) + Q_g^V, \end{aligned} \tag{5}$$

where d_k denotes the key dimension.

As illustrated in Fig. 2, the initial fused features $Z_g \in \mathbb{R}^{T\times(J_g\cdot D_e)}$ for each group are obtained via a part-aware STMamba module. A fully connected (FC) layer then reduces the feature dimension from $J_g \cdot D_e$ to D_g. The resulting group-wise fusion features are concatenated to construct the global fused representation $Z \in \mathbb{R}^{T\times(G\cdot D_g)}$. Subsequently, another FC layer transforms the feature dimension per frame from $G \cdot D_g$ to $J \cdot D_g$. Finally, a stack of L_N skeleton-aware STMamba performs global modeling to capture comprehensive spatio-temporal dependencies.

3.3. Cross-Mamba: An Alternative for Fusion

To verify that our design overcomes Mamba's limitations, we conduct a comparative evaluation. We further propose a Cross-Mamba module as an alternative, which is illustrated in Fig. 5. A linear layer adaptively maps inertial features Y_g^I to visual features Y_g^V, with weights learned during training to capture inter-modal significance. Both modalities undergo linear projections and 1D convolutions to refine local patterns. Deviating from the four-directional scanning in SS2D, the Cross State Space Model (Cross-SSM) concatenates visual and inertial features along the spatial axis and

applies bidirectional scanning (SS1D) for alignment with cross-attention. The resulting fused features are segmented into visual z_g^V and inertial z_g^I components, modulated via activated mapped features (SiLU), and then concatenated and projected to a unified group embedding dimension D_g, producing Z_g.

3.4. Pose Regression

The regression generates root-relative 3D poses for the entire sequence via layer normalization and a linear layer, transforming fused multimodal features into output pose $P \in \mathbb{R}^{T \times J \times 3}$ (3D coordinates of J joints across T frames).

3.5. Loss Function

To train VIMCAN, we utilize a composite loss function to promote accurate 3D pose estimation while enforcing temporal coherence. The loss function comprises four components: mean per-joint position error (MPJPE), normalized MPJPE (N-MPJPE), mean per-joint velocity error (MPJVE), and temporal consistency (TC) loss.

The MPJPE loss ($\mathcal{L}_{\text{MPJPE}}$) quantifies the L2 distance between predicted 3D poses $\hat{\mathbf{J}}$ and ground-truth (GT) coordinates $\mathbf{J}$ across batches B, time steps T, and joints J:

$$\mathcal{L}_{\text{MPJPE}} = \frac{1}{B \cdot T \cdot J} \sum_{b=1}^{B} \sum_{t=1}^{T} \sum_{j=1}^{J} \|\mathbf{J}_{b,t,j} - \hat{\mathbf{J}}_{b,t,j}\|^2. \quad (6)$$

To mitigate potential global-scale drift between the predicted and ground-truth sequences, we first use the least-squares method to compute a scale factor s, and then compute the normalized MPJPE loss ($\mathcal{L}_{\text{N-MPJPE}}$) between the scaled prediction and the target:

$$\mathcal{L}_{\text{N-MPJPE}} = \frac{1}{B \cdot T \cdot J} \sum_{b=1}^{B} \sum_{t=1}^{T} \sum_{j=1}^{J} \|s\mathbf{J}_{b,t,j} - \hat{\mathbf{J}}_{b,t,j}\|^2. \quad (7)$$

For temporal smoothness, the MPJVE loss [23] ($\mathcal{L}_{\text{V}}$) measures the L2 distance between the first-order derivatives of predicted and GT joint sequences, ensuring consistent motion velocities across frames:

$$\begin{aligned} \text{PJVE} &= \|(\mathbf{J}_{b,t+1,j} - \mathbf{J}_{b,t,j}) - (\hat{\mathbf{J}}_{b,t+1,j} - \hat{\mathbf{J}}_{b,t,j})\|^2, \\ \mathcal{L}_{\text{V}} &= \frac{1}{B \cdot (T-1) \cdot J} \sum_{b=1}^{B} \sum_{t=1}^{T-1} \sum_{j=1}^{J} \text{PJVE}. \end{aligned} \quad (8)$$

Additionally, the TC loss [7] ($\mathcal{L}_{\text{TC}}$) imposes a weighted L2 penalty on first-order differences in predicted poses to suppress abrupt fluctuations:

$$\mathcal{L}_{\text{TC}} = \frac{1}{B \cdot (T-1) \cdot J} \sum_{b=1}^{B} \sum_{t=1}^{T-1} \sum_{j=1}^{J} w_j \left\|\hat{\mathbf{J}}_{b,t+1,j} - \hat{\mathbf{J}}_{b,t,j}\right\|^2, \quad (9)$$

where w_j assigns higher penalties to perceptually critical joints, such as distal limbs.

The total loss is a weighted sum of these terms, enabling end-to-end training:

$$\begin{aligned} \mathcal{L}_{\text{Total}} = \lambda_{\text{MPJPE}} \cdot \mathcal{L}_{\text{MPJPE}} + \lambda_{\text{N-MPJPE}} \cdot \mathcal{L}_{\text{N-MPJPE}} \\ + \lambda_{\text{V}} \cdot \mathcal{L}_{\text{V}} + \lambda_{\text{TC}} \cdot \mathcal{L}_{\text{TC}}, \end{aligned} \quad (10)$$

where λ_{MPJPE}, $\lambda_{\text{N-MPJPE}}$, λ_{V}, and λ_{TC} are weighting coefficients optimized during training. Detailed parameters can be found in Sec. 4.

3.6. Training strategy

VIMCAN inherently supports inference on variable-length sequences, obviating the need for padding, masking, or fixed-length constraints prevalent in conventional approaches. To utilize this flexibility, we set a maximum sequence length of T = 81 and randomly sample sequences of lengths {9, 18, 27, 36, 45, 54, 63, 72, 81} within each training batch.

4. Experiments

4.1. Datasets and Evaluation Metrics

We assess VIMCAN on two benchmark datasets for 3D HPE: TotalCapture [25] and 3DPW [28]. The TotalCapture dataset consists of 1.9 million frames with synchronized multi-view videos from eight cameras, data from 13 IMUs on body segments, and Vicon-derived ground-truth annotations, making it suitable for visual-inertial fusion evaluation. It encompasses five subjects performing four actions—Range of Motion (ROM), Walking, Acting, and Freestyle—each in three trials. Consistent with prior protocols, the train set includes ROM (1, 2, 3), Walking (1, 3), Acting (1, 2), and Freestyle (1, 2) from subjects 1, 2, and 3. The test set comprises Walking (2), Acting (3), and Freestyle (3) across all subjects, facilitating evaluation of generalization to both seen and unseen subjects.

Similarly, the 3DPW dataset captures multiple actors engaging in diverse natural activities and contains various actions, such as playing basketball, walking, and freestyling. The images are sampled at 30 frames per second, and the poses are collected at 60 HZ, with SMPL parameters and 3D joints location ground truths. Due to the unavailability of direct IMU data, we generate synthetic IMU readings following [22] and adopt the official train-valid-test split [28].

Performance is evaluated using MPJPE (denoted as P1) in millimeters for absolute positional accuracy and Procrustes-aligned MPJPE (P-MPJPE, denoted as P2), where the estimated poses are aligned with GT through a rigid transformation. Memory I/O determines the computational efficiency of the model on the device during inference [12]. We measure throughput in frames per second (FPS), peak memory usage, and parameter count.

4.2. Experiments Setup

The VIMCAN model is implemented in PyTorch and trained on a single NVIDIA RTX 3090 GPU. Training is

Table 1. The performance comparison on TotalCapture dataset with different visual-inertial fusion methods. #IMUs: Number of IMUs used. 2D: Type of 2D pose detector (MP: MediaPipe, SN: SimpleNet, GT: Ground-Truth). S: Seen subject. U: Unseen subject. W: Walking. A: Acting. FS: Freestyle. P1: Average MPJPE (mm) across all test sets (lower is better). P2: Procrustes-aligned MPJPE (mm, lower is better).

Method	#IMUs	2D	SW2	SFS3	SA3	UW2	UFS3	UA3	P1 ↓	P2 ↓
Pan's [22]	6	MP	39.3	62.0	45.6	38.2	**43.6**	62.1	48.1	32.9
Wang's [29]	6	MP	**17.7**	48.3	29.1	39.6	64.1	50.2	39.0	28.8
VIMCAN (Ours)	6	MP	18.5	**39.8**	**24.1**	**34.8**	58.6	**44.3**	**33.2**	**25.7**
Wang's [29]	6	SN	**16.4**	41.7	25.2	35.4	62.0	44.8	34.9	26.9
VIMCAN (Ours)	6	SN	17.6	**39.1**	**22.3**	**33.8**	**58.7**	**42.6**	**31.2**	**23.6**
Bao's [2]	8	SN	**15.7**	47.3	20.1	**24.5**	60.2	50.5	33.7	-
Wang's [29]	8	SN	17.8	40.4	23.2	33.7	58.4	41.6	33.4	25.1
VIMCAN (Ours)	8	SN	17.0	**33.8**	**19.6**	33.2	**56.5**	**40.5**	**28.9**	**21.3**
Wang's [29]	6	GT	12.9	29.7	18.7	37.2	46.9	41.9	28.6	17.6
VIMCAN (Ours)	6	GT	**8.0**	**23.7**	**14.2**	**25.7**	**42.3**	**36.3**	**17.2**	**13.8**

Table 2. The performance comparison on 3DPW testing set using MediaPipe 2D pose detector.

Methods	Liu's	Pan's	Wang's	Ours
P1	60.3	55.0	53.9	**45.3**
Refs.	[17]	[22]	[29]	-

conducted with a batch size of 16 over 20 epochs, utilizing the AdamW optimizer with a weight decay of 0.01. The learning rate is initialized at 2×10^{-4} and decays exponentially with a decay factor of 0.99. The loss weights are set as λ_{MPJPE} =1, $\lambda_{\text{N-MPJPE}}$=0.5, λ_{V}=20, and λ_{TC}=0.5. Network hyper-parameters include an embedding dimension D_e=64, group dimension D_g=256. With a stack of L_N=5 skeleton-aware STMamba layers for global modeling.

4.3. Data Pre-processing

Visual inputs comprise 2D keypoint coordinates for J joints, extracted from monocular RGB videos and normalized to $[-1, 1]$ to ensure scale invariance across frames. Inertial inputs comprise unit-quaternion measurements from 6 IMUs mounted on the pelvis, sternum, and limbs, capturing rotations in the sensors' local coordinate systems. To align these heterogeneous modalities spatially, we compute the bone orientation in the camera frame as follows:

$$\mathbf{R}_B^C = \mathbf{R}_B^I \cdot \mathbf{R}_I^S \cdot \mathbf{R}_S^G \cdot \mathbf{R}_G^C, \tag{11}$$

where $\mathbf{R}_B^I$ denotes the relative rotation between the IMU and the corresponding skeletal bone, $\mathbf{R}_I^S$ is the measured orientation of the IMU, $\mathbf{R}_S^G$ transforms the IMU frame to a global coordinate system (*i.e.*, Vicon), and $\mathbf{R}_G^C$ accounts for extrinsic camera parameters. The terms $\mathbf{R}_B^I$ and $\mathbf{R}_S^G$ are derived via standardized T-pose calibration [9, 25].

4.4. Performance Comparison on Accuracy

We compare VIMCAN against the state-of-the-art visual-inertial fusion methods on TotalCapture and 3DPW datasets, including Pan's RNN-based approach [22] and Bao's TCN-based method [2], Liu's CNN-Transformer hybrid framework [17], and Wang's GCN-Transformer hybrid architecture [29].

On TotalCapture, we simultaneously employ MediaPipe and SimpleNet as 2D pose detectors. The ground-truth 2D keypoints are also used to isolate the impact of 2D detection noises and evaluate upper-bound performance. Table 1 reports the P1 performance on subsets and the P1/P2 results on the entire dataset. With 6 IMUs and MediaPipe, VIMCAN attains P1 of 33.2 mm and P2 of 25.7 mm, outperforming Pan's [22] and Wang's [29]. Using SimpleNet with 6 IMUs, VIMCAN achieves P1 of 31.2 mm and P2 of 23.6 mm, exceeding the results of Wang's [29]. The correlation between the 2D detection quality and the final 3D pose estimation accuracy is evident, which is discussed in the appendix. For 8 IMUs with SimpleNet, where our configuration emphasizes the upper-body activity via additional upper-arm IMUs, VIMCAN yields P1 of 28.9 mm and P2 of 21.3 mm, surpassing the results of Bao's [2] and Wang's [29]. Although Bao's [2] performs better on Walking subset, VIMCAN exhibits stronger generalization to complex actions such as Freestyle 3 (FS3) and Acting 3 (A3) for unseen subjects. With the ground-truth keypoints and 6 IMUs, VIMCAN further reduces errors, improving over Wang's [29] by 11.4 mm on P1 and 3.8 mm on P2. The qualitative result is visualized in Fig. 6.

On 3DPW dataset, following Pan *et al*. [22] and Wang *et al*. [29], we use MediaPipe as the 2D pose detector and fine-tune VIMCAN on its training set. Specifically, the model was initially trained on TotalCapture dataset, and we fine-tuned the feature extractors and regressor. The learning rate was set to 1×10^{-4} with the same decay strategy as above. As presented in Table 2, VIMCAN achieves 45.3 mm on P1, surpassing prior methods. These results highlight the robust accuracy and generalization capacity across diverse datasets and scenarios for VIMCAN.

Table 3. The ablation study for variable-length training strategy using the ground-truth 2D inputs on TotalCapture testing set. T: Fixed lengths. V: Variable-length. P1: Average MPJPE (mm, lower is better).

T	SW2	SFS3	SA3	UW2	UFS3	UA3	P1 ↓
9	9.7	26.6	14.9	28.3	44.7	39.0	21.1
27	8.6	24.7	14.5	26.4	43.0	37.7	19.2
81	**8.0**	**23.7**	**14.2**	**25.7**	**42.3**	**36.3**	**17.2**
V	8.4	24.2	14.5	26.4	42.8	37.3	18.9

Table 4. The ablation study for grouping and skeleton-aware scanning on TotalCapture testing set. #G: Number of groups for body parts. Skel.: Whether to use a skeleton-aware scanning schema or not. P1: Average MPJPE. #Params.: Number of parameters. Peak: Peak Memory (MB).

#G	Skel.	P1 ↓	#Parms.	Peak
0	✓	25.6	6.8M	248.6
3	✓	20.5	9.6M	265.1
5	✗	25.7	12.7M	290.7
5	✓	**17.2**	12.3M	282.5

Table 5. The ablation study for fusion strategies on TotalCapture test set. M: Methods. PM: PoseMamba (vision-only) [10], SA: Self-Attention (vision-only with Self-Attention module), CM: Cross-Mamba (visual-inertial Mamba-based fusion), CA: Cross-Attention (visual-inertial fusion with Cross-Attention module). P1: Average MPJPE.

M	SW2	SFS3	SA3	UW2	UFS3	UA3	P1 ↓
PM	15.6	38.4	16.4	38.0	62.9	53.3	28.1
SA	15.4	35.5	17.1	41.2	58.8	51.7	26.9
CM	12.3	26.3	15.4	31.2	48.0	41.9	24.3
CA	**8.0**	**23.7**	**14.2**	**25.7**	**42.3**	**36.3**	**17.2**

4.5. Ablation Studies

4.5.1. Stability of Variable-Length Training Strategy

To evaluate the robustness of the variable-length training strategy, we compare the performance using fixed-length inputs (9, 27, and 81 frames) against a dynamic approach, where the sequence lengths are randomly sampled from $9, 18, \ldots, 81$ during training. Table 3 presents the results with the ground-truth 2D keypoints, assessed via P1. The variable-length configuration (denoted as V) yields 18.9 mm on P1, approaching the optimal fixed-length result (17.2 mm at $T = 81$). Importantly, this approach maintains consistent accuracy across varying lengths, ensuring stability while accommodating flexible sequence processing.

4.5.2. Effectiveness of Feature Extractor

To evaluate the contributions of part-aware and skeleton-aware feature extraction, we conduct an ablation study on TotalCapture dataset using the ground-truth 2D keypoints. As detailed in Table 4, applying STMamba directly to ungrouped visual and inertial inputs (#G = 0) reduces the model size to 6.8 million parameters and peak memory to 248.6 MB, yet incurs a substantial accuracy degradation of 25.6 mm MPJPE. Introducing a three-part segmentation (trunk, upper limbs, lower limbs; see appendix for details) improves performance by over 5 mm, reducing the error to 20.5 mm. This gain underscores the benefit of the grouping method in enhancing robustness. Further increasing the number of groups to five yields an additional improvement, achieving the lowest error of 17.2 mm, while maintaining a moderate increase in model parameters. This progression confirms that finer anatomical decomposition improves the quality of estimation. Moreover, removing the skeleton-aware mechanism and relying solely on the part-aware STMamba for visual feature extraction leads to a sharp accuracy degradation (25.7 mm). These findings validate the essential integration of kinematic priors via skeleton-aware modeling and underscore the synergistic benefits of anatomical grouping and modality-specific scanning in VIMCAN.

4.5.3. Effectiveness of Multimodal Inputs

To evaluate the contribution of multimodal inputs, we compare VIMCAN against a vision-only PoseMamba (PM) [10], and then substitute the Cross-Attention module with Self-Attention module for visual inputs while maintaining the same grouping strategy (SA). As summarized in Table 5, incorporating Self-Attention module in PoseMamba reduces the P1 from 28.1 mm to 26.9 mm, a modest improvement of 1.2 mm. In contrast, VIMCAN's Cross-Attention integration of inertial data (CA) yields 17.2 mm on P1, a substantial 9.7 mm improvement over the vision-only Self-Attention baseline.

4.5.4. Performance Comparison of Fusion Modules

To demonstrate VIMCAN's effectiveness in addressing fine-grained cross-modal spatial reasoning, we ablate the fusion mechanism by comparing the integrated Cross-Attention module with a Cross-Mamba alternative. Both modules utilize identical IMUs and hyperparameters, with performance evaluated on fixed-length (T=81) and ground-truth keypoints. As reported in Table 5, the Cross-Attention module (CA) consistently outperforms the Cross-Mamba module (CM), particularly on unseen subject testing sets (e.g., UFS3 and UA3), where P1 improvements exceed 5.6 mm. The P1 performance for Cross-Attention (17.2 mm) surpasses Cross-Mamba (24.3 mm), underscoring the hybrid architecture's ability to model complex cross-modal spatial dependencies.

4.5.5. The Impact of Mamba Hyperparameters

To assess the influence of Mamba hyperparameters, we ablate the embedding dimension (D_e) and group dimension

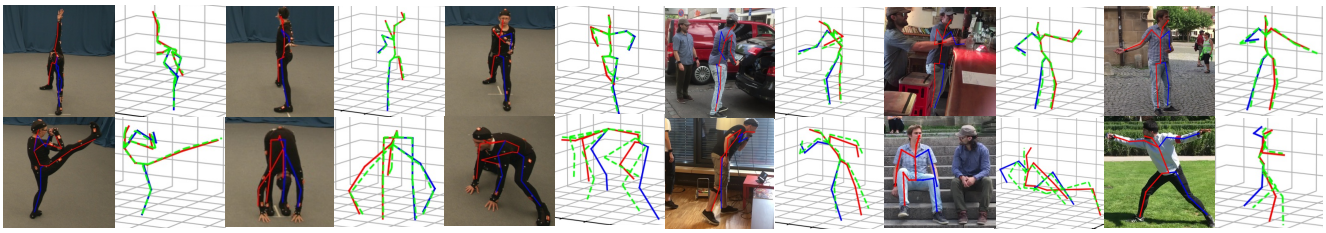


Figure 6. The qualitative analysis of VIMCAN. The green dashed lines denote the ground truth, and other colored lines represent the predictions.

Table 6. The ablation study for model hyperparameters. No.: Configuration index. D_e: Embedding dimension. D_g: Group dimension. P1: Average MPJPE. #Params.: Number of parameters. FPS: Frames per second.

No.	D_e	D_g	P1 ↓	#Parms.	FPS ↑
1	64	64	36.7	1.8M	**74.9**
2	64	128	34.5	3.9M	73.3
3	64	256	**31.2**	12.3M	61.4
4	96	256	35.2	13.7M	60.3
5	96	96	35.6	4.1M	72.7
6	128	128	34.5	7.1M	71.1

Table 7. The comparison for computational efficiency on TotalCapture testing set. P1: Average MPJPE (mm). #Params.: Number of parameters. Peak: Peak Memory (MB). FPS: Frames Per Second.

Methods	P1 ↓	#Params.	Peak ↓	FPS ↑
Wang's [29]	34.9	7.3M	969.8	45.8
CrossMamba	35.3	12.5M	**89.6**	64.6
VIMCAN-B	**31.2**	12.3M	282.5	61.4
VIMCAN-T	34.5	3.9M	156.4	**71.1**

(D_g). Experiments employ SimpleNet-derived 2D poses with fixed sequence length (T=27) on the TotalCapture testing set, measuring MPJPE, parameter count, and throughput (FPS) on an NVIDIA RTX 4060 Laptop GPU. Results in Table 6 reveal distinct trends. Configurations 1–3 show that increasing D_g from 64 to 256 (with $D_e = 64$) markedly enhances accuracy, reducing P1 from 36.7 mm to 31.2 mm, albeit at the expense of parameter count and FPS. Configuration 2 offers a strong balance of efficiency and accuracy, making it ideal for resource-constrained deployments. Configuration 3 delivers the lowest P1 (31.2 mm) while sustaining over 60 FPS, and is thus adopted as the primary setting. In contrast, raising D_e from 64 to 96 (Configuration 3 and 4) degrades performance, likely due to overfitting. Furthermore, configurations 1, 5, and 6 enforce $D_e = D_g$, achieving higher throughput (FPS $>$ 70) and lower computational cost (less parameters), but fail to match the best accuracy of the asymmetric configuration. These findings guide the selection of hyperparameters for practical deployment scenarios.

4.6. Performance Analysis

As demonstrated in Table 7, we compare VIMCAN with its Cross-Mamba variant and Wang's model (GCN-Transformer) [29] on consumer-grade hardware (NVIDIA RTX 4060 Laptop), evaluating accuracy (P1), model size (#Params.), peak memory usage (Peak), and inference speed (FPS). VIMCAN-B (Balance) corresponds to Configuration 3 in Table 6, while VIMCAN-T (Tiny) corresponds to Configuration 2. Since Wang's [29] code is not publicly available, we reimplemented their model, achieving a comparable parameter count (7.3M vs. 7.9M reported) while maintaining a reasonable computational scale. VIMCAN-B achieves the lowest MPJPE (31.2 mm), outperforming Wang's [29] (34.9 mm) and Cross-Mamba (35.3 mm), with a peak memory usage of 282.5 MB and FPS of 61.4. VIMCAN-T trades slight accuracy (34.5 mm) for a reduced model size (3.9M) and higher FPS (71.1), with a lower peak memory usage of 156.4 MB. This trade-off, as discussed above, makes VIMCAN-B preferable for accuracy-critical applications, while VIMCAN-T is suitable for resource-constrained deployments.

5. Conclusion

In this work, we introduce VIMCAN, a hybrid visual-inertial Mamba-Cross-Attention network for 3D HPE. Our approach integrates RGB keypoints and IMU data via Mamba's efficient sequence modeling and Cross-Attention's robust multimodal fusion, mitigating depth ambiguity. Extensive experiments demonstrate VIMCAN's efficacy. It significantly outperforms a leading GCN-Transformer model on TotalCapture, reducing MPJPE by 11.4 mm while requiring only 29% of its memory footprint and delivering 1.3× throughput. Furthermore, VIMCAN achieves a competitive 45.3 mm MPJPE on the challenging 3DPW dataset. Consequently, it achieves a better balance of accuracy and efficiency than prior RNN, CNN, and Transformer-based methods. VIMCAN natively supports variable-length sequences, making it well-suited for real-time applications in motion capture, rehabilitation, and human-computer interaction. One limitation is its reliance on strict sensor calibration. Future work will explore adaptive alignment techniques to enhance robustness.

Acknowledgments. This work was supported by the National Key R&D Program of China (No. 2022ZD0115902), the National Science and Technology Major Project (No. 2025ZD1605703), the National Natural Science Foundation of China (No. 32571260), the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No. VRLAB2024C06), and the Fundamental Research Funds for Beijing Municipal Universities (Grant No. 154225001). The authors gratefully acknowledge the financial support.

References

[1] Xiaoqi An, Lin Zhao, Chen Gong, Jun Li, and Jian Yang. Pre-training a density-aware pose transformer for robust lidar-based 3d human pose estimation. In *AAAI*, pages 1755–1763, 2025. 1, 2

[2] Yiming Bao, Xu Zhao, and Dahong Qian. Hybrid 3d human pose estimation with monocular video and sparse imus. *CoRR*, abs/2404.17837, 2024. 6

[3] Ju Dai, Hao Li, Rui Zeng, Junxuan Bai, Feng Zhou, and Junjun Pan. Kd-former: Kinematic and dynamic coupled transformer network for 3d human motion prediction. *Pattern Recognit.*, 143:109806, 2023. 2, 3

[4] Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. In *ICML*, 2024. 2

[5] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv Preprint arXiv: 2312.00752*, 2023. 2

[6] Ali Hatamizadeh and Jan Kautz. Mambavision: A hybrid mamba-transformer vision backbone. In *CVPR*, pages 25261–25270, 2025. 3

[7] Mir Rayat Imtiaz Hossain and James J. Little. Exploiting temporal information for 3d human pose estimation. In *ECCV*, 2018. 5

[8] Xiantao Hu, Ying Tai, Xu Zhao, Chen Zhao, Zhenyu Zhang, Jun Li, Bineng Zhong, and Jian Yang. Exploiting multimodal spatial-temporal patterns for video object tracking. In *AAAI*, pages 3581–3589, 2025. 2

[9] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J. Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser: learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Trans. Graph.*, 37(6): 185, 2018. 1, 2, 6

[10] 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. In *AAAI*, pages 3842–3850, 2025. 2, 3, 7

[11] Yifeng Jiang, Yuting Ye, Deepak Gopinath, Jungdam Won, Alexander W. Winkler, and C. Karen Liu. Transformer inertial poser: Real-time human motion reconstruction from sparse imus with simultaneous terrain generation. In *SIGGRAPH*, pages 3:1–3:9. ACM, 2022. 1, 2

[12] Sanghyeok Lee, Joonmyung Choi, and Hyunwoo J. Kim. Efficientvim: Efficient vision mamba with hidden state mixer based state space duality. In *CVPR*, pages 14923–14933, 2025. 5

[13] Jianwei Li, Wei Gao, Yihong Wu, Yangdong Liu, and Yanfei Shen. High-quality indoor scene 3d reconstruction with RGB-D cameras: A brief review. *Comput. Vis. Media*, 8(3): 369–393, 2022. 1, 2

[14] Tong Li and Haoyong Yu. Visual-inertial fusion-based human pose estimation: A review. *IEEE Trans. Instrum. Meas.*, 72:1–16, 2023. 1, 2

[15] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *CVPR*, pages 13137–13146, 2022. 1, 2

[16] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. NTU RGB+D 120: A large-scale benchmark for 3d human activity understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(10):2684–2701, 2020. 1, 2

[17] Liujun Liu, Jiewen Yang, Ye Lin, Peixuan Zhang, and Lihua Zhang. 3d human pose estimation with single image and inertial measurement unit (IMU) sequence. *Pattern Recognit.*, 149:110175, 2024. 2, 6

[18] Ruixu Liu, Ju Shen, He Wang, Chen Chen, Sen-Ching S. Cheung, and Vijayan K. Asari. Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction. In *CVPR*, pages 5063–5072, 2020. 1, 2

[19] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. In *NeurIPS*, 2024. 2, 3

[20] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, 2015. 2

[21] Soroush Mehraban, Vida Adeli, and Babak Taati. Motionagformer: Enhancing 3d human pose estimation with a transformer- gcnformer network. In *WACV*, pages 6905–6915, 2024. 1, 2

[22] Shaohua Pan, Qi Ma, Xinyu Yi, Weifeng Hu, Xiong Wang, Xingkang Zhou, Jijunnan Li, and Feng Xu. Fusing monocular images and sparse IMU signals for real-time human motion capture. In *SIGGRAPH Asia Conference*, pages 116:1–116:11, 2023. 5, 6

[23] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*, pages 7753–7762, 2019. 1, 2, 5

[24] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A large scale dataset for 3d human activity analysis. In *CVPR*, pages 1010–1019, 2016. 1, 2

[25] Trumble, Matthew, Gilbert, Andrew, Malleson, Charles, Hilton, Adrian, Collomosse, and John P. Total capture: 3D human pose estimation fusing video and inertial sensors. In *British Machine Vision Conference*, 2017. 5, 6

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 1, 2

[27] Timo von Marcard, Bodo Rosenhahn, Michael J. Black, and Gerard Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. *Comput. Graph. Forum*, 36(2):349–360, 2017. 1, 2

[28] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, pages 614–631, 2018. 5

[29] Baicun Wang, Ci Song, Xingyu Li, Huiying Zhou, Huayong Yang, and Lihui Wang. A deep learning-enabled visual-inertial fusion method for human pose estimation in occluded human-robot collaborative assembly scenarios. *Robotics Comput. Integr. Manuf.*, 93:102906, 2025. 1, 2, 3, 6, 8

[30] Peng Xu, Xiatian Zhu, and David A. Clifton. Multimodal learning with transformers: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(10):12113–12132, 2023. 2, 4

[31] Ming Yan, Yan Zhang, Shuqiang Cai, Shuqi Fan, Xincheng Lin, Yudi Dai, Siqi Shen, Chenglu Wen, Lan Xu, Yuexin Ma, and Cheng Wang. RELI11D: A comprehensive multimodal human motion dataset and method. In *CVPR*, pages 2250–2262, 2024. 1, 2

[32] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Transpose: real-time 3d human translation and pose estimation with six inertial sensors. *ACM Trans. Graph.*, 40(4):86:1–86:13, 2021. 1, 2

[33] 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 *ICCV*, pages 8784–8795, 2023. 1, 2

[34] Weihao Yu and Xinchao Wang. Mambaout: Do we really need mamba for vision? In *CVPR*, pages 4484–4496, 2025. 2, 3, 4

[35] Xinyi Zhang, Qiqi Bao, Qinpeng Cui, Wenming Yang, and Qingmin Liao. Pose magic: Efficient and temporally consistent human pose with a hybrid mamba-gcn network. In *AAAI*, pages 10248–10256, 2025. 3

[36] Zhao and Jingdong. A review of wearable IMU (inertial-measurement-unit)-based pose estimation and drift reduction technologies. *Journal of Physics: Conference Series*, 1087: 042003, 2018. 2

[37] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N. Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *CVPR*, pages 3425–3435, 2019. 1, 2

[38] Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In *CVPR*, pages 8877–8886, 2023. 1, 2

[39] Ce Zheng, Sijie Zhu, Matías Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *ICCV*, pages 11636–11645, 2021. 1, 2

[40] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *ICML*, 2024. 2

[41] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *ICCV*, pages 15039–15053, 2023. 1, 2