SAME: Spatial-Aware Multimodal Egocentric Human Pose Estimation

Yurong Fu[*1,2], **Peng Dai**[*1], **Yu Zhang**[1], **Feng Yiqiang**[1], **Yang Zhang**[†1], **Haoqian Wang**[†2],

[1]PICO
[2]Tsinghua University

Abstract

Egocentric human pose estimation (HPE) plays a crucial role in immersive applications such as virtual and augmented reality. However, existing methods relying on either visual or sparse inertial data alone often suffer from occlusion or ill-posed problems. In this work, we propose SAME, a novel spatial-aware multi-modal fusion framework combining the complementary signals from the stereo images and sparse IMUs for accurate and robust egocentric HPE. It adopts a two-stage network based on a dual coordinate frame to mitigate the coordinate inconsistencies among the stereo cameras and the IMUs. In the first stage, the IMU signals are transformed into the local frame and iteratively fused with the stereo images for estimating 3D poses in the local frame. In the second stage, the local poses are transformed into the global frame with the 6DOF head poses provided by the head-mounted display's (HMD) SLAM algorithm and then temporally aggregated via a temporal Transformer network. Meanwhile, to achieve geometric and semantic alignment among multi-modal features, we present a depth-guided spatial-aware deformable stereo attention network and a modality-aware Transformer decoder for cross-view and cross-modal feature fusion. Extensive experiments demonstrate that our approach achieves state-of-the-art performance on the public EMHI multi-modal egocentric pose estimation benchmark.

Introduction

Egocentric HPE aims to estimate 3D human poses from wearable sensors such as ego-facing head-mounted cameras, sparse inertial measurement units (IMUs), etc (Akada et al. 2025; Yi et al. 2025). It has gained great attention in recent years for its numerous applications (Li, Liu, and Wu 2023; Du et al. 2023), including virtual reality (VR), augmented reality (AR), etc. However, egocentric HPE faces great challenges due to severe self-occlusions and frequently varying viewpoints in egocentric settings (Grauman et al. 2022).

These methods inevitably suffer from inherent limitations of single-modal data: specifically, visual-based approaches struggle with frequent lower body occlusion and body parts falling outside the FOV of egocentric images, while IMU-based methods avoid occlusion but face drifting over time and ill-posed problems due to sparse observations. Recent works (Camiletto et al. 2025; Lee et al. 2025; Fan et al. 2025) thus shift to multi-modal fusion, but effective cross-sensor signal fusion remains challenging. Most existing methods adopt a late-fusion strategy, which involves training multiple models for different modalities separately and then integrating the output results of multiple models—this is suboptimal as it does not allow for training all data simultaneously. Besides, MEPoser (Fan et al. 2025) presents a baseline multi-modal fusion method which simply concatenates the IMU and visual features for regression and ignores the differing characteristics of each modality.

To address the above limitations, we propose a Spatial-Aware Multimodal Egocentric (SAME) HPE method, as shown in Fig. 1. First, a two-stage network based on a dual coordinate frame is proposed to fully utilize the characteristics of the two input modalities. Specifically, in the first stage, the IMU signals are transformed into the left camera coordinate system, defined as the local frame in this paper, and integrated with visual signals to estimate human poses in the local frame. The pose estimation in the local frame follows a coarse-to-fine manner. In the second stage, to explicitly leverage the 6DOF head poses provided by the HMD's SLAM system and improve the physical plausibility via a stable reference aligned with the ground, we transform the local poses into the global frame, and utilize a temporal Transformer network to fuse data over a time window. Notably, the whole network can be trained end-to-end.

Second, different from the late-fusion strategy in previous methods, we present a mid-level fusion method where the egocentric stereo image features and the IMU features are iteratively fused after feature extraction. One of the biggest challenges in mid-level fusion arises from building *geometric* and *semantic* correspondence among multi-modal features. To achieve *geometric* alignment among all input data, we propose a new depth-guided spatial-aware deformable stereo attention (DSA) network in the multi-view fusion module to encode 3D positional embeddings (PEs) into image features. It is of great value for bridging the geometric correspondence between stereo images and 3D space. Meanwhile, we adopt a modality-aware Transformer decoder for *semantic*-aligned multi-modal fusion. The core idea of this network is twofold. First, the unique characteristics of each modality are preserved by using different key

[*]These authors contributed equally.
[†]Corresponding Authors

and value projection layers. Second, the modality collaboration is achieved via a shared query projection layer.

In summary, our main contributions are as follows:

- A two-stage mid-level multi-modal fusion framework that fuses stereo images with sparse IMU signals for egocentric human pose estimation.
- A depth-guided spatial-aware DSA network and a modality-aware Transformer decoder improve the cross-view and cross-modal feature fusion performance.
- Extensive experiments demonstrate the effectiveness and superiority of our proposed method.

Related Works

Egocentric HPE from Single-Modal Input

Early works in egocentric HPE mainly focus on using a single fisheye camera. For instance, Mo^{2}Cap^2 (Xu et al. 2019) first utilizes CNN-based 2D modules to estimate the 2D heatmaps and the distance between the camera and each joint, then recovers 3D joints by back-projecting 2D detections using the fisheye camera model. xR-EgoPose (Tome et al. 2019) adopted a similar approach with a dual-branch autoencoder, while Selfpose (Tome et al. 2020) added rotation prediction and joint rotation loss. However, these monocular methods struggle with depth ambiguity.

To address the depth-ambiguity issue inherent to the single-camera setup, researchers explore the stereo setup as they provide enhanced depth information. For instance, EgoGlass (Zhao et al. 2021) predicts 2D heatmaps for each image first and then uses a 3D lifting module to generate 3D joint positions from the 2D heatmaps. UnrealEgo (Akada et al. 2022) adopts a similar pipeline but presents a new 2D module consisting of two weight-sharing encoders and one decoder to better utilize stereo information for heatmap estimation. UnrealEgo2 (Akada et al. 2024) expands on this by adding segmentation masks, the scene information, and temporal context to a Transformer-based framework. EgoTAP (Kang and Lee 2024) proposes a propagation network to leverage the kinematic structure of joints. EgoPoseFormer (Yang et al. 2024) presents a DSA operation to effectively fuse multi-view features, and employs a DETR-style (Carion et al. 2020) Transformer to predict poses in a coarse-to-fine manner. However, these methods omit encoding 3D PEs into image features, which would be helpful for bridging the geometric correspondence between stereo images and 3D space. In the field of multi-view 3D object detection with external cameras, a few methods (Liu et al. 2023; Xiong et al. 2023) have demonstrated the effectiveness of encoding 3D PEs into image features. However, to detect 3D objects in a large space, these methods usually use a large frustum shared by all views for 3D PE. Different from these methods, we propose a depth-guided 3D PE network to better adapt to the egocentric HPE scenarios. As shown in Fig. 2, it focuses only on a small shell range to iteratively finetune the 3D coordinates of each body joint.

Egocentric HPE from Multi-Modal Inputs

Common limitations such as severe occlusions still persist in the ego-view stereo setup, and prevent these methods from accurate HPE in challenging motions, such as dancing, fitness, yoga, etc. More recent works try to leverage complementary information provided by different sensor modalities to overcome these limitations. FRAME (Camiletto et al. 2025) proposes a two-stage network to combine the egocentric stereo images with SLAM poses. EgoAllo (Yi et al. 2024) uses SLAM head poses to condition a diffusion-based prior over body pose and height, and incorporates hand observations from egocentric videos during sampling. Besides the SLAM head poses and egocentric stereo images, REWIND (Lee et al. 2025) introduces identity-aware motion data to further enhance pose estimation quality. As in our paper, MEPoser (Fan et al. 2025) incorporates the egocentric stereo videos with sparse IMU signals including three 6DOFs from the HMD and hand controllers and two 3DOFs from the IMUs worn on the lower legs. However, only a baseline multi-modal fusion method which simply concatenates the IMU and visual features and regresses the body poses via a LSTM network is proposed in their paper. In contrast, our approach introduces a modality-aware Transformer network to achieve effective modality collaboration while preserving the unique characteristics of each modality.

Method

Overview

As illustrated in Fig. 1, our network aims at estimating human poses from a sequence of multi-modal egocentric observations over T frames:

$$f_\phi(\boldsymbol{J}_G^{1:T} \mid \boldsymbol{I}_{\text{Left}}^{1:T}, \boldsymbol{I}_{\text{Right}}^{1:T}, \boldsymbol{x}^{1:T}) \tag{1}$$

where ϕ denotes the network weights, $\boldsymbol{J}_G^{1:T}$ represents a sequence of N_J SMPL keypoints in the global frame, $\boldsymbol{I}_{v\in\{\text{Left, Right}\}}^t \in \mathbb{R}^{C\times H\times W}$ is an egocentric image captured from the viewpoint v at frame t, and $\boldsymbol{x}^t = [x_{\text{Head}}^t, ..., x_{\text{LeftLeg}}^t]$ is a set of sparse IMU measurements at frame t including three 6DOFs from the HMD and hand controllers and two 3DOFs from the IMUs worn on the lower legs.

As demonstrated in (Pan et al. 2023; Fan et al. 2025), it is non-trivial to directly regress human motions in the global frame, since there are significant modality differences and coordinate inconsistencies among the egocentric videos and IMU measurements. To address this issue, we decompose the task into two sequential stages:

$$f_\phi = f_{\phi_L}(\boldsymbol{J}_L^{1:T} \mid \boldsymbol{I}_{\text{Left}}^{1:T}, \boldsymbol{I}_{\text{Right}}^{1:T}, \boldsymbol{x}^{1:T}) * f_{\phi_G}(\boldsymbol{J}_G^{1:T} \mid \boldsymbol{J}_L^{1:T}, x_{\text{Head}}^{1:T}) \tag{2}$$

where f_{ϕ_L} denotes the network in the first stage that focuses on estimating poses $\boldsymbol{J}_L^{1:T}$ in the local frame which is defined as the left camera coordinate in this paper. In this stage, the IMU signals are firstly transformed into the unified local frame and then integrated with stereo videos to improve the pose estimation accuracy and robustness. To further improve the accuracy and physical plausibility, a temporal refinement network f_{ϕ_G} is proposed to attend over the whole time window in the second stage, yielding temporally smoother and more physically plausible motions $\boldsymbol{J}_G^{1:T}$ in the global frame.

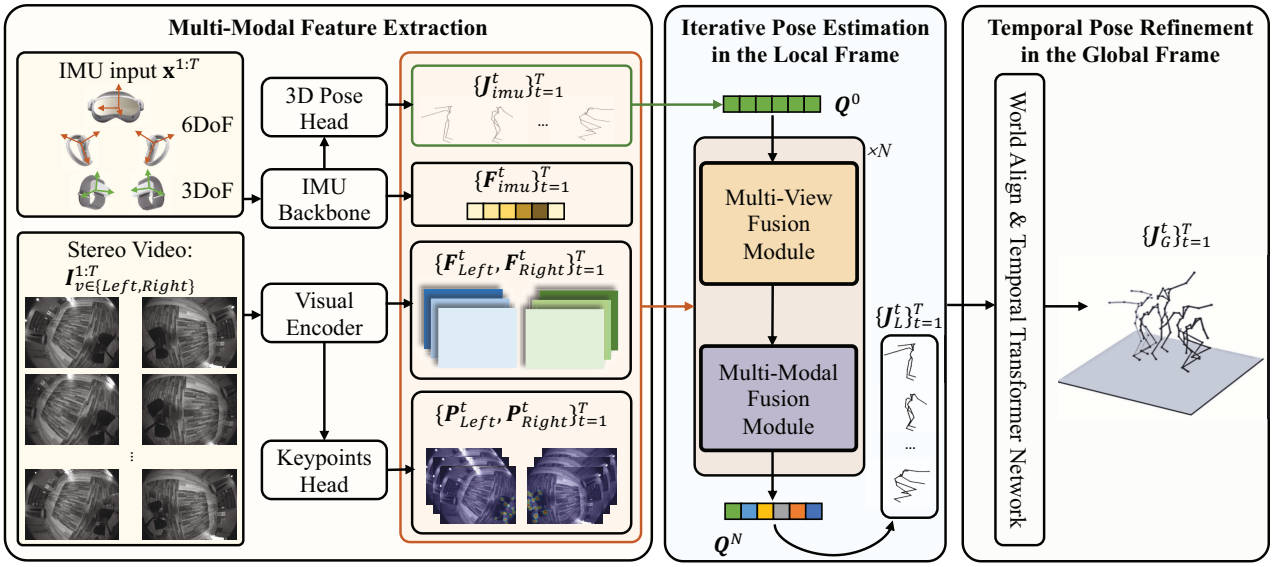


Figure 1: Overview of the proposed framework. Given a sequence of multi-modal egocentric observations, we first apply two dedicated backbones to extract image features $\{\boldsymbol{F}^t_{\text{Left}}, \boldsymbol{F}^t_{\text{Right}}\}^T_{t=1}$, 2D keypoints$\{\boldsymbol{P}^t_{\text{Left}}, \boldsymbol{P}^t_{\text{Right}}\}^T_{t=1}$, IMU features $\{\boldsymbol{F}^t_{\text{imu}}\}^T_{t=1}$ and initial 3D joint positions $\{\boldsymbol{J}^t_{\text{imu}}\}^T_{t=1}$. Next, for each timestamp t, we estimate the 3D poses in the local frame $\{\boldsymbol{J}^t_L\}^T_{t=1}$ via alternating between a multi-view fusion and a multi-modal fusion module. Lastly, we transform the local poses into the global frame and adopt a temporal Transformer network to generate the final 3D poses $\{\boldsymbol{J}^t_{\text{G}}\}^T_{t=1}$.

Multi-Modal Feature Extraction

To accommodate the heterogeneous nature of visual and IMU data, we employ two dedicated backbones for each input modality. First, in the IMU branch, a network similar to HMD-Poser (Dai et al. 2024) is adopted to estimate 3D joints $\{\boldsymbol{J}^t_{\text{imu}}\}^T_{t=1} \in \mathbb{R}^{N_J\times 3}$, and IMU features $\{\boldsymbol{F}^t_{\text{imu}}\}^T_{t=1} \in \mathbb{R}^{N_{\text{imu}}\times D}$, where D is the dimension of the IMU feature. Notably, different from (Dai et al. 2024) which operates in the global frame, our IMU network converts the raw IMU data into the local frame with the help of $x^{1:T}_{\text{Head}}$ and estimates 3D joints and features in the local frame. In this way, the coordinate inconsistencies among the egocentric videos and IMU measurements are alleviated. Second, the visual branch utilizes a pretrained ResNet (He et al. 2016) to extract stereo image features $\{\boldsymbol{F}^t_{\text{Left}}, \boldsymbol{F}^t_{\text{Right}}\}^T_{t=1} \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times D}$ and estimate 2D joint positions $\{\boldsymbol{P}^t_{\text{Left}}, \boldsymbol{P}^t_{\text{Right}}\}^T_{t=1} \in \mathbb{R}^{N_J\times 2}$. The estimated 2D joints will act as reference anchors in the subsequent attention-based multi-view fusion module.

Iterative Pose Estimation in the Local Frame

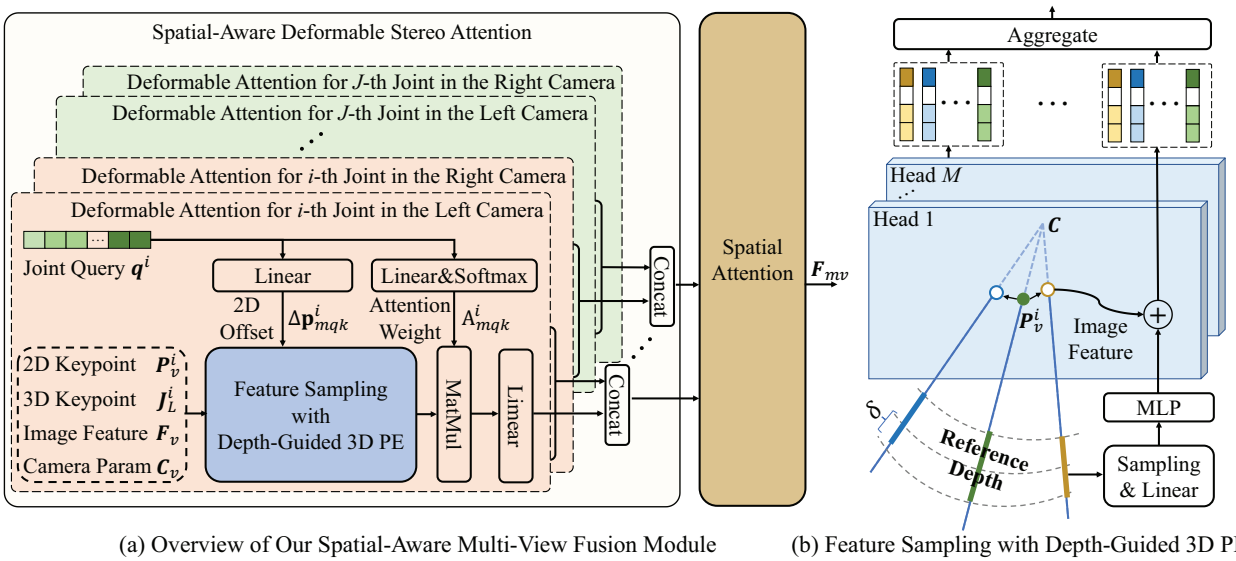
After collecting all stereo visual clues and IMU features as described above, we propose an iterative pose estimation network operating in a coarse-to-fine manner. Since all frames are processed identically in this network, we remove the subscript t in this subsection for the sake of simplicity. As shown in Fig. 1, the iterative pose estimation network starts from the 3D poses estimated from the IMU branch $\boldsymbol{J}^0_L = \boldsymbol{J}_{\text{imu}}$, and alternates between a multi-view fusion module for fusing multi-view information and a multi-modal fusion module for enabling cross-modal collaboration. After repeating N iterations, our model obtains the final 3D poses in the local frame $\boldsymbol{J}_L = \boldsymbol{J}^N_L$ with higher accuracy.

Spatial-Aware Multi-View Fusion. For multi-view fusion, EgoPoseFormer (Yang et al. 2024) applies a DSA operation to process multi-view features and achieves state-of-the-art results on public egocentric datasets. However, it ignores the spatial-related information, which is also crucial for generating accurate 3D poses. In this paper, a spatial-aware multi-view fusion (SA-MVF) module is proposed, as shown in Fig. 2. It takes the previous iteration's output query features $\boldsymbol{Q}^{n-1}$ and 3D poses $\boldsymbol{J}^{n-1}_L$, the stereo image features $[\boldsymbol{F}_{\text{Left}}, \boldsymbol{F}_{\text{Right}}]$, the 2D joint positions $[\boldsymbol{P}_{\text{Left}}, \boldsymbol{P}_{\text{Right}}]$, and the camera parameters $[\boldsymbol{C}_{\text{Left}}, \boldsymbol{C}_{\text{Right}}]$ as inputs, and outputs spatial-aware multi-view fused image features $\boldsymbol{F}^n_{\text{mv}}$. The initial query features $\boldsymbol{Q}^0$ are computed by feeding $\boldsymbol{J}_{\text{imu}}$ and a scalar identifier into a multilayer perceptron (MLP) network. Structurally, the SA-MVF module consists of a spatial-aware DSA network for joint-level feature extraction and a spatial attention network for cross-joint feature fusion.

As shown in Fig. 2 (a), for i-th body joint in viewpoint v, given a query element with content feature $\boldsymbol{q}^i \in \boldsymbol{Q}^{n-1}$, a 2D reference point $\boldsymbol{p}^i_v$, and a 3D reference point $\boldsymbol{J}^i_L \in \boldsymbol{J}^{n-1}_L$, the spatial-aware deformable attention is calculated by

$$\begin{aligned}\boldsymbol{o}^i_v &= \text{SpatialAwareDeformAttn}(\boldsymbol{q}^i, \boldsymbol{p}^i_v, \boldsymbol{J}^i_L, \boldsymbol{F}_v, \boldsymbol{C}_v) \\ &= \sum_{m=1}^{M_a} \boldsymbol{W}_m[\sum_{k=1}^{K_s} A_{mqk}\cdot \boldsymbol{W}'_m(\boldsymbol{F}_v(\boldsymbol{p}^i_v + \Delta\boldsymbol{p}^i_{mqk}) \\ &\quad + f_{\text{PE}}(\boldsymbol{p}^i_v + \Delta\boldsymbol{p}^i_{mqk}, \boldsymbol{J}^i_L, \boldsymbol{C}_v))]\end{aligned} \tag{3}$$

where m indexes the attention head, k indexes the sampled key, M_a and K_s are the number of attention heads



(a) Overview of Our Spatial-Aware Multi-View Fusion Module (b) Feature Sampling with Depth-Guided 3D PE

Figure 2: (a) The architecture of the proposed spatial-aware multi-view fusion network; (b) A diagram illustrating the feature sampling process with depth-guided 3D positional embedding.

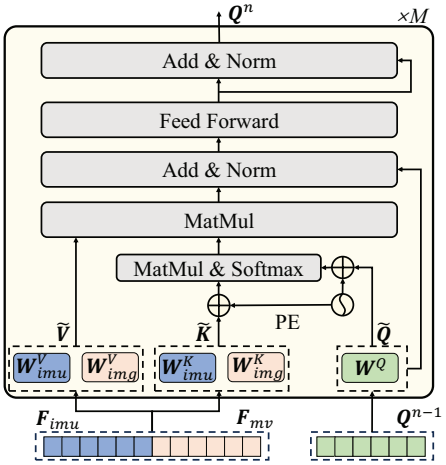


Figure 3: The modality-aware multi-modal fusion module.

and the total sampled key number respectively. $\boldsymbol{W}_m$ and $\boldsymbol{W}'_m$ are learnable weights as in standard multi-head attention network. $\Delta \boldsymbol{p}^i_{mqk}$ and A_{mqk} denote the sampling 2D offset and the attention weight of the k-th sampling point in the m-th attention head, respectively. Both $\Delta \boldsymbol{p}^i_{mqk}$ and A_{mqk} are obtained via linear projection over the query feature $\boldsymbol{q}^i$. As $\boldsymbol{p}^i_v + \Delta \boldsymbol{p}^i_{mqk}$ is fractional, a bilinear interpolation as in (Zhu et al. 2020) is applied to compute image feature $\boldsymbol{F}_v(\boldsymbol{p}^i_v + \Delta \boldsymbol{p}^i_{mqk})$. To help the model learn the geometric correspondence between images and the 3D space, we propose a new depth-guided 3D PE network to extract 3D PE for each sampled 2D point. Mathematically, the 3D PE of the sampled 2D point $\hat{\boldsymbol{p}}^i_v = \boldsymbol{p}^i_v + \Delta \boldsymbol{p}^i_{mqk}$ is calculated by

$$f_{\text{PE}}(\hat{\boldsymbol{p}}^i_v, \boldsymbol{J}^i_L, \boldsymbol{C}_v) = \text{MLP}(\Theta(\Pi^{-1}(\hat{\boldsymbol{p}}^i_v, \boldsymbol{C}_v), d(\boldsymbol{J}^i_L))) \quad (4)$$

where $\Pi^{-1}(\hat{\boldsymbol{p}}^i_v, \boldsymbol{C}_v)$ denotes the inverse projection function parameterized by the camera parameters $\boldsymbol{C}_v$. Θ is a sampling process along the un-projected ray direction. As shown in Fig. 2 (b), a fixed number of N_s 3D points are uniformly sampled in the depth range $[d(\boldsymbol{J}^i_L) - \delta, d(\boldsymbol{J}^i_L) + \delta]$, where $d(\cdot)$ is a function to extract the depth component of each 3D point and δ is the predefined sampling range. With the sampled 3D points at hand, an MLP network is adopted to mapping them into a single feature embedding. Note that the inverse projection function ensures that the sampled 3D points from all cameras are within the unified local frame. To this end, the 3D PEs from different cameras are aligned. The details of the inverse projection and the sampling process are provided in the supplementary material.

The multi-view features $\{\boldsymbol{o}^i_{\text{Left}}, \boldsymbol{o}^i_{\text{Right}}\}_{i=1}^{N_J}$ are concatenated and fed into a linear projection layer to generate spatial-aware stereo image features.

$$\boldsymbol{o}^i = \text{Linear}(\text{Concat}(\boldsymbol{o}^i_{\text{Left}}, \boldsymbol{o}^i_{\text{Right}})),\ i \in [1, ..., N_J] \quad (5)$$

Notably, for joints with low 2D detection confidence, we mask out their features to prevent unreliable information from being incorporated. Next, a self-attention operation is applied to extract spatial and human kinematic information among all joint features.

$$\boldsymbol{F}^n_{\text{mv}} = \text{SelfAttention}([\boldsymbol{o}^1, \boldsymbol{o}^2, ..., \boldsymbol{o}^{N_J}]) \quad (6)$$

Modality-Aware Multi-Modal Fusion. Prior multi-modal feature alignment methods (Lyu et al. 2024) typically align IMU features with visual features by projecting IMU features into the visual semantic space. However, as demonstrated in (Ye et al. 2024), this strategy can cause a mismatch in granularity, where visual features often contain fruitful semantic information compared to the discrete information within IMU features. To this end, a modality-aware multi-modal fusion (MA-MMF) module is proposed in this

paper, as shown in Fig. 3. It is composed of a stack of M identical layers. Each layer has two sub-layers. The first is a modality-specific self-attention mechanism, and the second is a fully connected feed-forward network. Meanwhile, a residual connection followed by layer normalization is employed around each of the two sub-layers.

Formally, in the m-th layer, given the spatial-aware multi-view features $\boldsymbol{F}_{\text{mv}}^{n} \in \mathbb{R}^{N_J \times D}$, the IMU features $\boldsymbol{F}_{\text{imu}} \in \mathbb{R}^{N_{\text{imu}} \times D}$ and the previous layer's output query features $\boldsymbol{Q}_{m-1}^{n} \in \mathbb{R}^{N_J \times D}$, we leverage separate linear projection layers for the key and value projection matrices while preserving query projection matrix shared as follows:

$$\tilde{\boldsymbol{K}} = [\boldsymbol{F}_{\text{imu}} \boldsymbol{W}_{\text{imu}}^{K};\ \boldsymbol{F}_{\text{mv}}^{n} \boldsymbol{W}_{\text{image}}^{K}] \tag{7}$$

$$\tilde{\boldsymbol{V}} = [\boldsymbol{F}_{\text{imu}} \boldsymbol{W}_{\text{imu}}^{V};\ \boldsymbol{F}_{\text{mv}}^{n} \boldsymbol{W}_{\text{image}}^{V}] \tag{8}$$

$$\tilde{\boldsymbol{Q}} = \boldsymbol{Q}_{m-1}^{n} \boldsymbol{W}^{Q} \tag{9}$$

$$\boldsymbol{Q}_{m}^{n} = \text{Softmax}\left(\frac{\tilde{\boldsymbol{Q}}\tilde{\boldsymbol{K}}^{\top}}{\sqrt{D}}\right)\tilde{\boldsymbol{V}} \tag{10}$$

where $\boldsymbol{W}_{\text{imu}}^{K}, \boldsymbol{W}_{\text{image}}^{K}, \boldsymbol{W}_{\text{imu}}^{V}, \boldsymbol{W}_{\text{image}}^{V}, \boldsymbol{W}^{Q} \in \mathbb{R}^{D \times D}$ are the learnable projection matrices. $\boldsymbol{Q}_{m}^{n}$ is the updated query feature. By preserving the unique characteristics of each modality through different key and value projection layers, we can avoid interference between the two modalities, particularly in relation to granularity mismatch. Meanwhile, using a shared query projection matrix can achieve modality collaboration. After M layers' modality-aware attention, we can obtain the multi-view and multi-modal fused features $\boldsymbol{Q}_{M}^{n}$ which will be adopted as the query features in the next iteration. Meanwhile, a lightweight MLP regression head is employed to predict the joint positions $\boldsymbol{J}_{L}^{n}$ from $\boldsymbol{Q}_{M}^{n}$.

Pose Refinement in the Global Frame

Having the estimated body joints $\{\boldsymbol{J}_{L}^{t}\}_{t=1}^{T}$ in the local frame, we first transform them to the global frame $\{\hat{\boldsymbol{J}}_{\text{G}}^{t}\}_{t=1}^{T}$ with the help of head motions $\{x_{\text{Head}}^{t}\}_{t=1}^{T}$ from the HMD's SLAM system. Then, we propose a temporal Transformer network (TTN) for pose refinement in the global frame. This design aims at taking the following advantages of optimizing poses in the global frame. First, a stable reference aligned with the ground can be established in the global frame, with which we can correct artifacts such as skating or ground penetration. Second, we can reduce unstable poses as the direction of gravity is known in the global frame. Specifically, the TTN is composed of a stack of 3 Transformer layers, and each having a feedforward dimension of 256 and 8 attention heads. It attends all the information over a time-window of $T = 32$. Finally, an MLP network is used to output the refined body joints in the global frame $\{\boldsymbol{J}_{\text{G}}^{t}\}_{t=1}^{T}$.

Loss Function

We define the overall loss function $\mathcal{L}$ as a combination of the 3D joint loss in the local frame $\mathcal{L}_{L}^{3D}$, the 3D joint loss in the global frame $\mathcal{L}_{G}^{3D}$, and a ground-penetration loss $\mathcal{L}_{ground}$:

$$\mathcal{L} = \lambda_{L}\mathcal{L}_{L}^{3D} + \lambda_{G}\mathcal{L}_{G}^{3D} + \lambda_{ground}\mathcal{L}_{ground} \tag{11}$$

where λ_{L}, λ_{G}, and λ_{ground} are the weights for the respective loss terms. $\mathcal{L}_{L}^{3D}$ and $\mathcal{L}_{G}^{3D}$ are calculated as the mean of mean squared errors between the predicted and the ground-truth values. The ground-penetration loss $\mathcal{L}_{\text{ground}}$ is defined as:

$$\mathcal{L}_{\text{ground}} = \frac{1}{T}\sum_{t=1}^{T}\left(\tanh(\min(0, h_{\text{min}}^{t}))\right)^{2} \tag{12}$$

$$h_{\text{min}}^{t} = \min_{j \in \{1,\ldots,N_J\}}(y_{j}^{t} - y_{\text{ground}}) \tag{13}$$

where y_{ground} is the height of the ground plane in the global frame, y_{j}^{t} is the predicted height of j-joint at frame t. $\mathcal{L}_{\text{ground}}$ penalizes any joint that falls below the ground plane.

Experiments

Implementation Details. SAME is implemented in Pytorch and trained using the AdamW optimizer (Loshchilov and Hutter 2017) with a batch size of 4. We first train two dedicated backbones for each input modality. Due to limitations of GPU resources, these two pretrained backbones are frozen when training the full model, though we acknowledge that end-to-end fine-tuning of the entire model has the potential to yield further performance improvements. Then, we train our full model with an initial learning rate of 1×10^{-3} and a weight decay of 3×10^{-5} for 20 epochs. It takes about 20 hours to train our model on 8 NVIDIA A100 GPUs. The detailed network architectures and parameters are present in the supplementary materials.

Dataset. We conduct all comparison experiments and ablation studies on the public EMHI (Fan et al. 2025) dataset, which is a large-scale multi-modal egocentric HPE benchmark providing both egocentric vision and IMU signals captured by the real VR product suite. It consists of 885 sequences captured by 58 subjects performing 39 actions, totaling about 3.07M frames. We follow the official practice of using EMHI dataset by splitting the dataset into three parts: one for training (70%) and two separate sets, i.e., **Protocol1** (16%) and **Protocol2** (14%), for testing. Note that **Protocol1** contains the same set of actions as in the training set and is used to evaluate cross-subject generalization, while **Protocol2** consists of unseen actions not present in the training set and is designed to assess the model's generalization ability in out-of-distribution actions.

Metrics. As in prior works, we use a total of 7 metrics which can be divided into two categories. The first category measures the HPE accuracy and includes the mean per joint positional error (MPJPE), Procrustes-aligned MPJPE (PA-MPJPE), upper-body MPJPE (U-PE), lower-body MPJPE (L-PE), foot MPJPE (FootPE), and hand MPJPE (HandPE). The second category measures the physical plausibility of the generated poses. Following (Zheng et al. 2023), we use the Ground metric in this category. All metrics are reported in millimeters in this paper.

Comparisons

We compare our SAME with both the latest single-modal and multi-modal methods. For single-modal baselines, we

Dataset	Methods	Input	PA-MPJPE↓	MPJPE↓	U-PE↓	L-PE↓	FootPE↓	HandPE↓	Ground↓
Protocol1	HMD-Poser	IMU	27.8	57.6	48.4	70.9	78.9	52.3	19.6
	UnrealEgo	Stereo	38.6	55.2	39.9	77.3	100.2	85.2	22.1
	EgoPoseFormer	Stereo	29.9	50.7	32.9	76.5	96.6	51.0	31.9
	FRAME	IMU+Stereo	28.2	37.4	28.0	51.1	62.1	45.1	2.4
	MEPoser	IMU+Stereo	23.8	36.9	27.1	51.0	63.5	45.0	23.7
	Ours	IMU+Stereo	**20.7**	**33.4**	**22.8**	**48.8**	**57.5**	**23.5**	**2.0**
Protocol2	HMD-Poser	IMU	33.5	70.5	51.8	97.5	112.6	57.1	25.7
	UnrealEgo	Stereo	42.8	63.7	46.4	88.7	115.5	100.8	28.7
	EgoPoseFormer	Stereo	32.7	62.6	42.0	92.3	118.1	66.8	41.8
	FRAME	IMU+Stereo	44.3	60.5	55.8	67.4	78.6	75.5	6.0
	MEPoser	IMU+Stereo	29.4	47.6	32.1	70.2	92.2	53.8	23.7
	Ours	IMU+Stereo	**21.6**	**38.0**	**25.5**	**56.1**	**66.9**	**25.4**	**4.7**

Table 1: Comparison with state-of-the-art methods on the EMHI dataset. The best results are in **bold**.

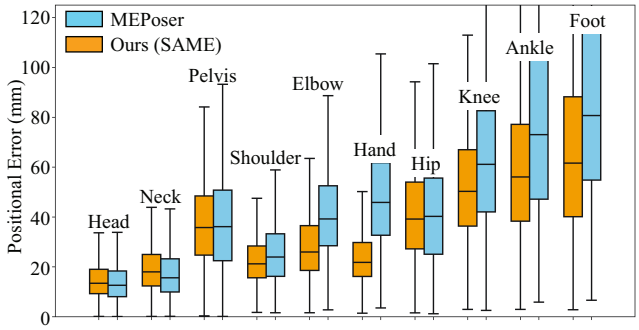


Figure 4: Per joint error analysis in Protocol2.

consider HMD-Poser (Dai et al. 2024) using three 6DOFs from HMD and two 3DOFs from IMUs, and two state-of-the-art methods for egocentric HPE from stereo images: UnrealEgo (Akada et al. 2022) and EgoPoseFormer (Yang et al. 2024). For multi-modal baselines, we choose two most recent methods: FRAME (Boscolo Camiletto et al. 2025) using both stereo images and the head 6D pose as input, and MEPoser (Fan et al. 2025) having the same input data as ours. To ensure a rigorous comparison with the existing methods, we directly adopt the evaluation results of HMD-Poser, UnrealEgo, and MEPoser provided in (Fan et al. 2025). As for EgoPoseFormer and FRAME, we retrain their models on EMHI dataset according to the original settings described in their papers.

Table 1 provides the detailed quantitative results highlighting that our method outperforms previous single-modal and multi-modal methods across all evaluation metrics in both **Protocol1** and **Protocol2**. In particular, our method surpasses all existing methods by a large margin in **Protocol2**, where the actions do not overlap with the training set. It demonstrates the superiority of our method not only in pose estimation accuracy but also in robustness and generalizability when applying to real-world VR and AR applications with unseen actions. To further understand how SAME achieves state-of-the-art accuracy, we visualize the detailed positional error improvement w.r.t. different body joints in Fig 4. It reveals that the performance improvement is particularly significant in the limbs including elbow, hand, knee, ankle and foot. Taking hand as an example, our method reduces the HandPE from 45.0mm to 23.5mm in **Protocol1**, and 53.8mm to 25.4mm in **Protocol2**, achieving over 45% improvement in both testing sets. These improvements are largely attributed to our spatial-aware multi-view fusion and modality-aware multi-modal fusion modules, which effectively incorporates the sparse IMU observations of hands and lower legs with the stereo images, leading to a significant MPJPE reduction in these joints.

SA-MVF	MA-MMF	TTN	PA-MPJPE↓	MPJPE↓	Ground↓
✗	✗	✗	23.9	42.5	25.8
✗	✓	✓	23.1	40.7	6.0
✓	✗	✓	22.5	38.8	5.9
✓	✓	✗	22.3	39.0	24.5
✓	✓	✓	**21.6**	**38.0**	**4.7**

Table 2: Ablation study for key components in our model.

IMU	I_{Left}	I_{Right}	PA-MPJPE↓	MPJPE↓	Ground↓
✓	✗	✗	32.1	52.1	6.4
✗	✓	✓	36.0	62.4	13.0
✓	✓	✗	25.3	41.0	6.2
✓	✓	✓	**21.6**	**38.0**	**4.7**

Table 3: Ablation study for the input modalities.

In Fig 5, we present a few visualization results for qualitative comparisons. Obviously, our model could achieve significantly more accurate results. And the most significant improvement happens in the limbs which is consistent with the results in Fig 4 and Table 1.

We also provide more visualization results and quantitative results including the computation efficiency in our supplementary materials. Please kindly refer to our supplementary materials for more details.

Ablation Study

We ablate our method in various settings to evaluate the effects of the key components, the input modalities, etc. All

Figure 5: Qualitative comparison on a few challenging inputs. For better illustration, the predicted (red) and ground-truth (green) 3D poses are re-projected onto external reference views which are not used for pose estimation. Note that the input IMU data is not presented in this figure for simplicity. The qualitative results confirm that our SAME obtains more accurate 3D poses than existing single-modal and multi-modal methods. Best viewed in color.

ablation studies are conducted on **Protocol2**.

Impact of Proposed Key Components. As described previously, there are three key components in our model: SA-MVF, MA-MMF, and TTN for pose refinement in the global frame. Table 2 summarizes our ablation study by removing each model component separately to isolate their contribution and removing all components to identify the overall value of our model. First, when the depth-guided 3D PE is removed and only the DSA network is used for multi-view fusion (w/o SA-MVF), there is a significant performance drop in both MPJPE and PA-MPJPE metrics, suggesting our SA-MVF can help the model to encode accurate spatial information into image features, hence contributing to a significant improvement in HPE accuracy. Second, when replacing the MA-MMF with a vanilla Transformer decoder (w/o MA-MMF) or removing the TTN module, the MPJPE and PA-MPJPE metrics have deteriorated, indicating MA-MMF and TTN can contribute to additional accuracy improvement. Third, TTN plays an important role in reducing the Ground metric, validating the temporal pose refinement in the global frame is the key for improving the physical plausibility. Finally, the model removing all components obtains the worst results as expected. However, it is still better than MEPoser in both MPJPE and PA-MPJPE metrics, hinting that the simple multi-modal fusion method in MEPoser is less effective in obtaining accurate egocentric HPE.

$\mathcal{L}_{\text{ground}}$	PA-MPJPE↓	MPJPE↓	Ground↓
✗	22.4	38.5	8.1
✓	**21.6**	**38.0**	**4.7**

Table 4: Ablation study for the ground-penetration loss.

Impact of the Input Modalities. To investigate the effectiveness of the input signals, Table 3 reports the experimental results of using the following input settings: (1) IMU-only; (2) Stereo images; (3) IMU + the left camera view; (4) IMU + stereo images. Note that the missing data is replaced with zeros to keep our model unchanged. It can be concluded from the results that: (1) Both the IMU data and the stereo images play an important role in egocentric HPE. (2) Our model using only IMU and the left camera can achieve even better performance than MEPoser, validating the superiority of our method. (3) Our model can effectively incorporate the complementary signals from multi-modal data to achieve state-of-the-art accuracy in egocentric HPE.

Impact of the Ground-Penetration Loss. Table 4 shows the comparison results with and w/o the ground-penetration loss. As expected, introducing the ground-penetration loss in model training can contribute to additional improvement in both accuracy and physical plausibility.

For more ablation experiments on the effects of the model size and the choice of hyperparameters, please refer to our supplementary materials.

Conclusion

We present a spatial-aware multimodal egocentric HPE method combining the complementary signals from the stereo images and sparse IMUs. To achieve geometric and semantic alignment among multi-modal features, we present a depth-guided spatial-aware deformable stereo attention network and a modality-aware Transformer decoder for cross-view and cross-modal feature fusion. Experiments validate that our approach achieves superior results with respect to state-of-the-art methods on the public EMHI dataset. Comprehensive ablation studies demonstrate the effectiveness of each proposed component.

Limitations. As a data-driven method, our approach is highly dependent on large-scale training data. It is really hard to collect in-the-wild dataset with high-quality GT labels. Investigating effective semi-supervised and unsupervised training methods for generalization to real-world scenarios would be an interesting direction for future work.

Acknowledgements

This work is supported by the Shenzhen Science and Technology Project under Grant KJZD20240903103210014.

References

Akada, H.; Wang, J.; Golyanik, V.; and Theobalt, C. 2024. 3d human pose perception from egocentric stereo videos. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 767–776.

Akada, H.; Wang, J.; Golyanik, V.; and Theobalt, C. 2025. Bring your rear cameras for egocentric 3d human pose estimation. *arXiv preprint arXiv:2503.11652*.

Akada, H.; Wang, J.; Shimada, S.; Takahashi, M.; Theobalt, C.; and Golyanik, V. 2022. Unrealego: A new dataset for robust egocentric 3d human motion capture. In *European Conference on Computer Vision*, 1–17. Springer.

Boscolo Camiletto, A.; Wang, J.; Alvarado, E.; Dabral, R.; Beeler, T.; Habermann, M.; and Theobalt, C. 2025. FRAME: Floor-aligned Representation for Avatar Motion from Egocentric Video. *arXiv preprint arXiv:2503.23094*.

Camiletto, A. B.; Wang, J.; Alvarado, E.; Dabral, R.; Beeler, T.; Habermann, M.; and Theobalt, C. 2025. FRAME: Floor-aligned Representation for Avatar Motion from Egocentric Video. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 17497–17507.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Dai, P.; Zhang, Y.; Liu, T.; Fan, Z.; Du, T.; Su, Z.; Zheng, X.; and Li, Z. 2024. 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.

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.

Fan, Z.; Dai, P.; Su, Z.; Gao, X.; Lv, Z.; Zhang, J.; Du, T.; Wang, G.; and Zhang, Y. 2025. Emhi: A multimodal egocentric human motion dataset with hmd and body-worn imus. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2879–2887.

Grauman, K.; Westbury, A.; Byrne, E.; Chavis, Z.; Furnari, A.; Girdhar, R.; Hamburger, J.; Jiang, H.; Liu, M.; Liu, X.; et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18995–19012.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Kang, T.; and Lee, Y. 2024. Attention-propagation network for egocentric heatmap to 3d pose lifting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 842–851.

Lee, J.; Xu, W.; Richard, A.; Wei, S.-E.; Saito, S.; Bai, S.; Wang, T.-L.; Sung, M.; Kim, T.-K.; and Saragih, J. 2025. REWIND: Real-Time Egocentric Whole-Body Motion Diffusion with Exemplar-Based Identity Conditioning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 7095–7104.

Li, J.; Liu, K.; and Wu, J. 2023. Ego-body pose estimation via ego-head pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17142–17151.

Liu, Y.; Yan, J.; Jia, F.; Li, S.; Gao, A.; Wang, T.; and Zhang, X. 2023. Petrv2: A unified framework for 3d perception from multi-camera images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3262–3272.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lyu, Y.; Zheng, X.; Zhou, J.; and Wang, L. 2024. Unibind: Llm-augmented unified and balanced representation space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26752–26762.

Pan, S.; Ma, Q.; Yi, X.; Hu, W.; Wang, X.; Zhou, X.; Li, J.; and Xu, F. 2023. Fusing monocular images and sparse imu signals for real-time human motion capture. In *SIGGRAPH Asia 2023 Conference Papers*, 1–11.

Tome, D.; Alldieck, T.; Peluse, P.; Pons-Moll, G.; Agapito, L.; Badino, H.; and De la Torre, F. 2020. Selfpose: 3d egocentric pose estimation from a headset mounted camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6): 6794–6806.

Tome, D.; Peluse, P.; Agapito, L.; and Badino, H. 2019. xr-egopose: Egocentric 3d human pose from an hmd camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7728–7738.

Xiong, K.; Gong, S.; Ye, X.; Tan, X.; Wan, J.; Ding, E.; Wang, J.; and Bai, X. 2023. Cape: Camera view position embedding for multi-view 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 21570–21579.

Xu, W.; Chatterjee, A.; Zollhoefer, M.; Rhodin, H.; Fua, P.; Seidel, H.-P.; and Theobalt, C. 2019. Mo 2 cap 2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera. *IEEE transactions on visualization and computer graphics*, 25(5): 2093–2101.

Yang, C.; Tkach, A.; Hampali, S.; Zhang, L.; Crowley, E. J.; and Keskin, C. 2024. EgoPoseFormer: A Simple Baseline for Stereo Egocentric 3D Human Pose Estimation. In *European Conference on Computer Vision*, 401–417. Springer.

Ye, Q.; Xu, H.; Ye, J.; Yan, M.; Hu, A.; Liu, H.; Qian, Q.; Zhang, J.; and Huang, F. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 13040–13051.

Yi, B.; Ye, V.; Zheng, M.; Li, Y.; Müller, L.; Pavlakos, G.; Ma, Y.; Malik, J.; and Kanazawa, A. 2024. Estimating body and hand motion in an ego-sensed world. *arXiv preprint arXiv:2410.03665*.

Yi, B.; Ye, V.; Zheng, M.; Li, Y.; Müller, L.; Pavlakos, G.; Ma, Y.; Malik, J.; and Kanazawa, A. 2025. Estimating body and hand motion in an ego-sensed world. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 7072–7084.

Zhao, D.; Wei, Z.; Mahmud, J.; and Frahm, J.-M. 2021. Egoglass: Egocentric-view human pose estimation from an eyeglass frame. In *2021 International Conference on 3D Vision (3DV)*, 32–41. IEEE.

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.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.