EgoPoseVR: Spatiotemporal Multi-Modal Reasoning for Egocentric Full-Body Pose in Virtual Reality

Haojie Cheng, Shaun Jing Heng Ong, Shaoyu Cai,
Aiden Tat Yang Koh, Fuxi Ouyang, and Eng Tat Khoo*

Fig. 1: Egocentric full-body pose estimation in VR enabled by an integrated setup that combines a VR headset with hand controllers and a headset-mounted downward-facing RGB-D camera. The configuration supports real-time and accurate tracking of both upper- and lower-body movements to facilitate natural avatar control and immersive interaction.

Abstract— Immersive virtual reality (VR) applications demand accurate, temporally coherent full-body pose tracking. Recent head-mounted camera-based approaches show promise in egocentric pose estimation, but encounter challenges when applied to VR head-mounted displays (HMDs), including temporal instability, inaccurate lower-body estimation, and the lack of real-time inference. To address these limitations, we present *EgoPoseVR*, an end-to-end framework for accurate egocentric full-body pose estimation in VR that integrates headset motion cues with egocentric RGB-D observations through a dual-modality fusion pipeline. A spatiotemporal encoder extracts frame- and joint-level representations, which are fused via cross-attention to fully exploit complementary motion cues across modalities. A kinematic optimization module then imposes constraints from HMD signals, enhancing the accuracy and stability of pose estimation. To facilitate training and evaluation, we introduce a large-scale synthetic dataset of over 1.8 million temporally aligned HMD and RGB-D frames across diverse VR scenarios. Experimental results show that *EgoPoseVR* outperforms state-of-the-art egocentric pose estimation models. A user study in real-world scenes further shows that *EgoPoseVR* achieved significantly higher subjective ratings in accuracy, stability, embodiment, and intention for future use compared to baseline methods. These results show that *EgoPoseVR* enables robust full-body pose tracking, offering a practical solution for accurate VR embodiment without requiring additional body-worn sensors or room-scale tracking systems.

Index Terms—Virtual reality, head-mounted displays, egocentric full-body pose estimation, multimodal spatiotemporal fusion.

1 Introduction

Virtual reality (VR) enables a wide range of immersive applications, including embodiment entertainment [9, 47, 50], physical training and rehabilitation [5, 32, 52], and social collaboration [24, 42, 56]. Many of these scenarios require accurate and spatiotemporally consistent full-body pose estimation to faithfully reflect the user's physical state and enable intuitive interaction. However, current consumer VR devices offer limited full-body tracking capabilities. Therefore, researchers have widely explored full-body pose estimation by integrating external systems into VR setups, such as optical marker-based motion capture systems [12, 41] and third-person view cameras [48, 70]. Although current optical marker-based systems can achieve high-accuracy pose tracking, they require large capture spaces and meticulous calibration [53]. While third-person vision offers complete whole-body observations with minimal occlusion, its reliance on external viewpoints makes it unsuitable for immersive first-person applications such as VR, which inherently depend on egocentric sensing [3].

Modern VR hardware, including head-mounted displays (HMDs) with hand-held controllers, provides stable and low-latency motion measurements for the user's head and hands [44]. However, inferring a full-body pose from only head and hand tracking is fundamentally ambiguous, as similar upper-body motions may correspond to multiple plausible full-body configurations [20, 68]. This limitation underscores the need to capture lower-body movement data in VR. One approach is to attach additional body-worn sensors on the lower limbs to provide extra motion cues [8, 17, 39, 55, 63, 64, 69]. Among these, inertial measurement units (IMUs) are the most widely used, as they are compact and robust to occlusion, but require precise calibration and suffer from drift and usability burdens [39]. Another promising alternative

- *Haojie Cheng, Shaun Jing Heng Ong, Shaoyu Cai, Aiden Tat Yang Koh, and Eng Tat Khoo are with National University of Singapore. E-mail: hjcheng@nus.edu.sg, jong1.05@nus.edu.sg, shaoyucai@nus.edu.sg, aiden@nus.edu.sg, etkhoo@nus.edu.sg.*
- *Fuxi Ouyang is with Singapore University of Technology and Design. E-mail: fuxi_ouyang@mymail.sutd.edu.sg.*
- * *Corresponding author.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

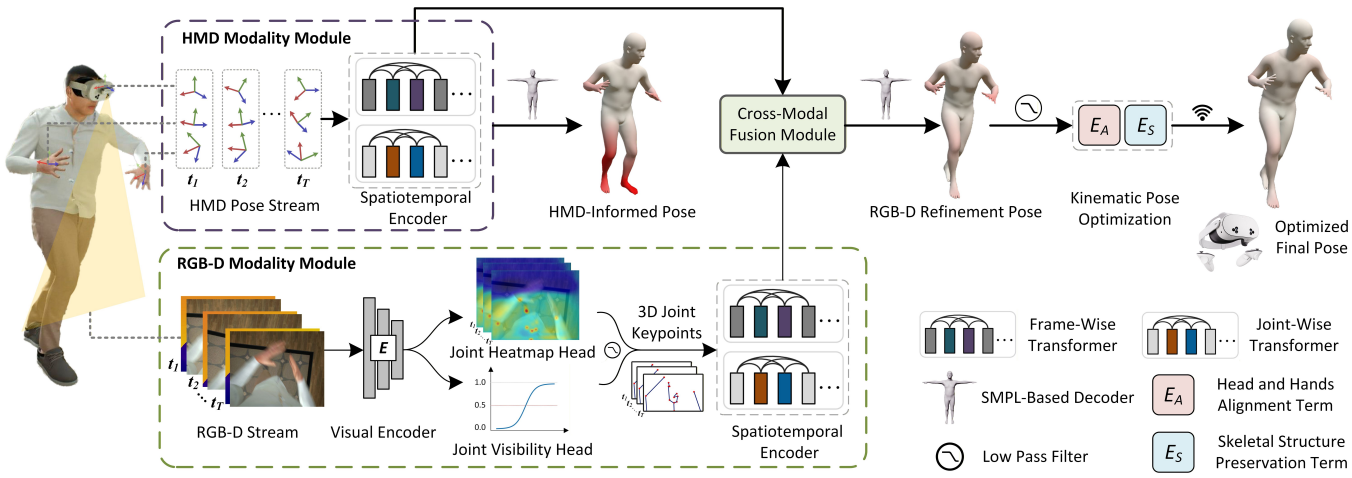


Fig. 2: Workflow of our EgoPoseVR framework for egocentric full-body pose estimation in VR. HMD motion and egocentric downward-facing RGB-D inputs are jointly encoded by the spatiotemporal module to predict full-body poses. A tailored kinematic optimization via energy functions ensures structurally consistent and immersive avatar rendering.

is to use an egocentric downward-facing camera [43], which captures first-person observations of the user's body. Recent studies in computer vision have demonstrated that this configuration substantially improves the accuracy of lower-body pose [7,21,22,25,34,37,51,61,62], demonstrating improved accuracy of lower-body pose estimation. However, the direct transfer of these methods to VR is impeded not only by camera access constraints [13], but also by fundamental challenges in modality integration.

Integrating HMD motion data with egocentric visual input in a unified framework poses significant challenges. HMD motion data are spatially sparse, structured, and low-dimensional, whereas egocentric vision data is high-dimensional and highly sensitive to even minor head movements, which often induce abrupt scene changes or motion blur and compromise temporal stability [3,21,58]. Moreover, existing egocentric vision methods are already computationally demanding [3], and extending them with additional VR motion signals from head-mounted displays and controllers further increases computational complexity and inference latency. Beyond these representational differences, progress is further constrained by the absence of multi-modal datasets that provide temporally synchronized headset motion and egocentric vision data in VR environments. These challenges are further amplified in VR, where users are highly sensitive to inaccuracies and latency [45], motivating the development of a principled framework that unifies motion and vision under practical immersive conditions.

To address these issues, we present *EgoPoseVR* as shown in Fig. 2, a dual-stream spatiotemporal framework for egocentric full-body pose estimation that combines HMD-based motion data with visual input from a downward-facing egocentric camera, specifically designed for VR scenarios. Each stream is encoded with spatiotemporal context, capturing both temporal dynamics and joint-level dependencies. Within the RGB-D modality, we design a visibility-aware temporal joint detection module that mitigates instability from motion blur and scene variations, thereby enhancing robustness to visual noise. To exploit complementary information across modalities, we further incorporate a cross-attention fusion module that refines HMD motion features with visual cues. Finally, a kinematic optimization module is introduced to improve pose accuracy by enforcing consistency with VR signals while preserving skeletal structure. To support both the training and inference, we construct a multi-modal dataset tailored for egocentric full-body pose estimation in VR. Evaluations on the synthetic dataset demonstrate that *EgoPoseVR* achieves more accurate and stable pose estimation than existing state-of-the-art methods, while maintaining real-time performance at 97 FPS. A real-world user study further validated these findings, showing significantly higher subjective ratings in accuracy, stability, embodiment, and future-use intention.

In summary, the main contributions of our work are as follows:

- We introduce *EgoPoseVR*, a novel egocentric pose estimation system for VR that combines HMD motion signals and egocentric visual observations through a headset-mounted RGB-D camera, enabling deployable full-body tracking without external infrastructure.
- We propose a dual-stream spatiotemporal architecture that encodes frame-wise dynamics and joint-level dependencies, integrates complementary cues through a cross-modal fusion module and refines predictions with a kinematic optimizer enforcing structural consistency with VR-tracked signals.
- We construct the first large-scale synthetic dataset for egocentric full-body pose estimation in VR, featuring temporally synchronized HMD motion and RGB-D data with 2D and 3D joint annotations, comprising 1.8 million frames across varied VR scenes and applicable to different HMD–camera setups.

2 Related Works

2.1 Egocentric Pose Estimation in VR

In VR applications, a common and convenient configuration is to estimate full-body pose using only HMD input, without relying on additional sensors or visual observations. Physics-based approaches like QuestSim [60] and QuestEnvSim [27] generate physically plausible motion via reinforcement learning and simulation, but are difficult to integrate into real-time systems due to their reliance on non-differentiable simulators. In contrast, data-driven models (e.g., AvatarPoser [20], AvatarJLM [68] and EgoPoser [19]) regress poses directly using neural architectures with temporal modeling or inter-joint constraints, which contribute to improved performance of pose estimation. More recently, diffusion-based methods such as AGRoL [11] and EgoEgo [28] produce realistic motion sequences, though their iterative sampling and prediction of future frames limit real-time applicability. These HMD-only methods often struggle to predict lower-body pose accurately, due to the lack of global body context in spatially sparse HMD input.

To enhance full-body pose estimation beyond spatially sparse HMD inputs, recent work has explored integrating additional body-worn sensors (e.g., IMUs) [13]. To recover full-body motion from sparse signals, DIP [17] employs recurrent sequence modeling, while SIP [55] frames the task as an optimization over sparse-to-dense pose trajectories. Other methods incorporate biomechanical constraints and structural priors to improve physical plausibility, such as TransPose [64] and PIP [63]. Real-time applications have also been explored in HMD-Poser [8], which combines lightweight neural networks with online shape

Table 1: Comparison of representative datasets for egocentric full-body human pose estimation.

	Mo2Cap2 [61]	xR-EgoPose [51]	EgoGlass [66]	UnrealEgo [2]	ARES [28]	SynthEgo [7]	EgoPoseVR (Ours)
Camera Lens Type	Fisheye	Fisheye	Perspective	Fisheye	Perspective	Perspective	Perspective
Mono/Stereo	Mono	Mono	Stereo	Stereo	Mono	Stereo	Mono
Body Visibility	✓	✓	✓	✓	✗	✓	✓
Environment	In- & Outdoor	Mostly Indoor	Indoor	Indoor	In- & Outdoor	In- & Outdoor	In- & Outdoor
Joint Location	✓	✓	✓	✓	✓	✓	✓
Joint Rotation	✗	✗	✗	✗	✓	✓	✓
Depth Availability	✗	✗	✗	✗	✗	✗	✓
HMD Adaptability	✗	✗	✗	✗	✗	✓(Implicit)	✓(Explicit)
Temporal Continuity	✗	✗	✗	✗	✗	✗	✓
Dataset Size	$530k$	$383k$	$2 \times 170k$	$2 \times 450k$	$1.2M$	$2 \times 60k$	$1.8M$

adaptation for practical VR use. Beyond IMUs, alternative sensing strategies include wearable cameras attached to the torso or limbs, as in Nymeria [35] and EgoSim [16], which combine RGB camera with motion capture suits. However, these wearable sensors typically require bespoke hardware setup and are vulnerable to magnetic interference [23], limiting their scalability and practicality in VR contexts.

2.2 Motion Tracking from Egocentric Camera

A prominent direction in egocentric pose estimation refines lower-body predictions by employing a downward-facing head-mounted camera to capture visual observations [3]. Early methods such as Mo2Cap2 [61], xR-EgoPose [51] and EgoPW [57] demonstrate feasibility but are fundamentally limited by severe self-occlusions and the lack of explicit kinematic modeling. Recent approaches extend beyond monocular baselines by incorporating additional cues. Geometry-aware methods, such as Ego3DPose [21], exploit stereo correspondence and orientation cues to mitigate self-occlusions. Scene-aware methods, exemplified by Scene-Aware EgoPose [59], integrate depth and voxelized context to enforce plausibility. More recently, generative motion models, including REWIND [25], adopt diffusion-based formulations for real-time whole-body estimation. EgoPoseFormer [62] and Ego4View [1] propose coarse-to-fine refinement strategies to progressively lift egocentric observations to 3D poses. While improving robustness, such coarse-to-fine designs incur high inference latency and rely heavily on pelvis-to-camera constraints.

Current image-based egocentric pose estimation methods are primarily evaluated on synthetic datasets (Table 1), such datasets are widely adopted as they enable scalable data generation without site-specific constraints, support diverse and controllable variations in environments and clothing, and avoid the need for additional complex motion capture setups. While synthetic data may not fully capture all aspects of real-world variability, it provides a practical and controllable foundation for training and evaluation. Rather than addressing the sim-to-real gap solely through expanded data coverage, we focus on algorithmic designs that improve robustness under distribution shifts between synthetic and real-world scenarios. Beyond this limitation, current methods typically perform pose estimation on a per-frame basis, which leads to temporal inconsistency. Self-occlusion and motion blur from abrupt head movements can further distort the egocentric view and destabilize predictions. In addition, prior works have overlooked the practical constraint that commercial headsets (e.g., Meta Quest and Apple Vision Pro) do not grant the data access of downward-facing fisheye cameras, which restricts their applicability and prevents rigorous validation in real-world VR environments [34].

2.3 Spatiotemporal Modeling for Temporal Dynamics

Spatiotemporal modeling is already widely used in third-person 3D human pose estimation, as temporal context alleviates frame-wise ambiguities and enforces kinematic consistency across frames [10]. Transformer-based architectures such as PoseFormer [67], MixSTE [65], and MHFormer [29] leverage token-wise attention to capture long-range dependencies across RGB sequences, while diffusion-based models, like DiffPose [14] and S^2Fusion [49], further incorporate probabilistic temporal priors. Despite their effectiveness, these methods generally use dense visual frames and are often tailored for offline settings. Recent efficient designs such as UNSPAT [26] and HoT [30] attempt to reduce the computation of full-frame processing by aligning spatial positions or pruning. In contrast, several studies demonstrate that reliable temporal modeling can be achieved without densely sampled video image sequences. PhaseMP [46] and ReMP [18] encode long-range dynamics from sparse motion streams using phase representations or reusable motion priors, while KTPFormer [38] improves spatial-temporal stability by injecting kinematic priors.

Despite advances in leveraging spatiotemporal visual information, egocentric pose estimation remains hindered by dataset limitations. As shown in Table 1, popular benchmarks such as Mo2Cap2 [61], xR-EgoPose [51] and UnrealEgo [2] primarily provide isolated frames without temporally aligned image sequences, restricting the development of spatiotemporal models. Moreover, most datasets do not include explicit annotations of joint rotations (e.g., EgoGlass [66]), which are required to simulate dynamic six degrees of freedom (6DoF) motion signals from headsets and controllers. The absence of both temporal dynamics and rotation information further prevents consistent alignment between headset and controller signals and egocentric visual inputs. This gap motivates our creation of a spatiotemporally aligned egocentric pose dataset tailored for VR applications.

3 Methodology

3.1 Overview

We tackle the task of real-time full-body 3D pose estimation for VR, predicting both joint orientations and positions from synchronized egocentric signals. The input consists of temporal motion data from the headset and hand controllers, complemented by downward-facing RGB-D observations captured by a commercial camera mounted on the headset. In this setting, ensuring computational efficiency while enabling complementary use of modalities without cross-interference is a central challenge of our design.

To this end, we propose *EgoPoseVR*, a dual-stream spatiotemporal framework (Fig. 2). Each stream is processed by a designed spatiotemporal encoder built with Transformer layers that capture frame-wise temporal dynamics and joint-wise dependencies, yielding modality-specific representations. The motion stream operates on headset and controller trajectories, while the visual stream encodes temporally stable joints extracted from RGB-D input, providing compact visual-geometric cues. The two representations are fused in a shared latent space to predict a refined full-body pose, which is subsequently refined by a tailored kinematic optimizer to ensure consistency with VR tracking signals and skeletal structural constraints. The final output is a temporally stable and spatially accurate full-body pose estimation, well-suited for low-latency VR deployment.

3.2 Input Modalities and Pose Representation

The input signal $\mathbf{x}_t$ at the current frame t from the HMD consists of spatially sparse motion measurements. For each device $i \in \{H, C_L, C_R\}$ corresponding to the head, left controller and right controller respectively, we extract the global 3D position $\mathbf{p}_t^i \in \mathbb{R}^3$, 6D orientation $\boldsymbol{\theta}_t^i \in \mathbb{R}^6$, linear velocity $\mathbf{v}_t^i \in \mathbb{R}^3$, and 6D angular velocity $\boldsymbol{\omega}_t^i \in \mathbb{R}^6$. These components are concatenated into a global motion descriptor $\mathbf{g}_t^i = [\mathbf{p}_t^i, \boldsymbol{\theta}_t^i, \mathbf{v}_t^i, \boldsymbol{\omega}_t^i] \in \mathbb{R}^{18}$. To reduce sensitivity to global pose and facilitate body-centric reasoning, we explicitly compute the relative pose of each controller with respect to the headset. For each controller

$c \in \{C_L, C_R\}$, we define the head-relative position $\tilde{\mathbf{p}}_t^c \in \mathbb{R}^3$ and 6D rotation $\tilde{\boldsymbol{\theta}}_t^c \in \mathbb{R}^6$ by transforming the controller pose into the HMD coordinate frame. These features are concatenated into local motion descriptor $\mathbf{r}_t^c = [\tilde{\mathbf{p}}_t^c,\ \tilde{\boldsymbol{\theta}}_t^c] \in \mathbb{R}^9$. The HMD input vector $\mathbf{x}_t \in \mathbb{R}^{72}$ is constructed by concatenating all motion descriptors and can be written as:

$$\mathbf{x}_t = [\mathbf{g}_t^H,\ \mathbf{g}_t^{C_L},\ \mathbf{g}_t^{C_R},\ \mathbf{r}_t^{C_L},\ \mathbf{r}_t^{C_R}] \in \mathbb{R}^{72}. \tag{1}$$

In parallel, a downward-facing RGB-D camera mounted on the HMD captures egocentric visual observations, denoted as $\mathbf{y}_t \in \mathbb{R}^{C\times H\times W}$, where C, H and W denote the number of channels, height and width, respectively. To leverage the temporal continuity of HMD motion and RGB-D image data, we construct a preset window of T consecutive frames ending at time t, i.e., $\mathbf{X} = [\mathbf{x}_{t-T+1}, \ldots, \mathbf{x}_t] \in \mathbb{R}^{T\times 72}$, $\mathbf{Y} = [\mathbf{y}_{t-T+1}, \ldots, \mathbf{y}_t] \in \mathbb{R}^{T\times C\times H\times W}$. Therefore, the final input for each frame t is defined as the combination of HMD motion and visual modalities:

$$\mathcal{I}_{t-T+1:t} = \{\mathbf{X},\ \mathbf{Y}\}. \tag{2}$$

Our model predicts full-body human pose using the SMPL parameterization [33]. Following prior work [20], we focus on the main skeletal structure and exclude the two palm joints, resulting in 22 output joints. For each joint, the model predicts its rotation relative to its parent joint in the kinematic chain. This yields an output of dimension $\mathbb{R}^{132}$ per frame, corresponding to 22 joints represented by a 6D rotation for SMPL-based avatar control.

3.3 HMD and RGB-D Feature Modeling

3.3.1 HMD Stream Encoder

To effectively capture temporal dynamics and spatial structure from egocentric motion data, we introduce a spatiotemporal encoder that operates on HMD-based measurements, as defined in Eq. (3). Given the temporal HMD descriptors $\mathbf{X}$ in Eq. (2), we first project each frame into a latent space via a linear embedding $E_l(\cdot)$, producing a sequence of feature vectors. To capture long-range temporal dependencies and mitigate the limitations of frame-wise or short-term modeling, we adopt a frame-wise transformer encoder $\mathcal{T}_f(\cdot)$. Its self-attention mechanism enables dynamic weighting of past motion cues across the entire temporal window, thereby facilitating robust trend extraction and improved temporal coherence in downstream pose estimation.

$$\mathbf{M}_t = \mathcal{T}_j\left(\mathbf{F}\left(\left[\mathcal{T}_f\left(E_l(\mathbf{X})\right)\right]_t\right) + \mathbf{e}_j\right). \tag{3}$$

To retain temporal context while focusing on the current t-th frame, we extract the final temporal token and project it into a structured pose representation via a multi-layer perceptron (MLP), denoted as $\mathbf{F}(\cdot)$. This step summarizes long-term dynamics into a frame-specific feature, ensuring that subsequent modules operate with temporally aware information. To encourage the model to learn joint-specific semantics rather than treating all joints uniformly, we add a learnable embedding $\mathbf{e}_j$ to each joint feature. This provides an identity cue that helps the network disambiguate joints with similar motion patterns and facilitates reasoning about cross-joint relationships. Finally, we further refine the joint features using a joint-wise transformer encoder $\mathcal{T}_j(\cdot)$. This allows the network to model kinematic constraints and generate anatomically consistent pose hypotheses. The resulting joint-wise feature representation $\mathbf{M}_t$ forms a tensor of shape $\mathbb{R}^{J\times S}$, where J is the number of joints and S is the per-joint feature dimension.

3.3.2 RGB-D Stream Encoder.

While the HMD stream captures head and hand trajectories and can coarsely infer lower-body motion, its predictions remain ambiguous. The same upper-body movement can correspond to multiple plausible lower-body poses. To resolve this ambiguity, we introduce a complementary vision stream that processes RGB-D inputs from a downward-facing camera.

To utilize the RGB-D stream effectively, a convolutional heatmap encoder, i.e., a ResNet backbone with a feature pyramid network (FPN) [31], first processes the RGB-D input to estimate 3D joint keypoint sequence $\mathbf{Z} \in \mathbb{R}^{T\times J\times 3}$. Compared to directly apply spatiotemporal modeling on dense image sequences $\mathbf{Y}$ in Eq. 2, 3D joint keypoints significantly reduce the computational complexity of temporal reasoning from image-level (i.e., $O(T \cdot C \cdot H \cdot W)$) to joint-level (i.e., $O(T \cdot J)$). In addition, the explicit 3D joint supervision stabilizes training and provides spatially grounded intermediate features.

Algorithm 1: Temporal refinement of 3D joint estimates

Input : $\mathbf{Y}$, $\mathbf{Z}$, T, J and ζ
Output : Refined 3D joint keypoint sequence $\tilde{\mathbf{Z}}$

if $\mathbf{Y}$ *is the first joint keypoint batch* **then**
  foreach (t', j) *where* $t' \in [t-T+1, t]$, $j \in [1, J]$ **do**
    $\tilde{\zeta}_{t',j} \leftarrow \mathcal{F}(\zeta_{t',j})$;
  end
  Joint-wise mask $\eta \leftarrow \mathrm{ReLU}(\tilde{\zeta} - 0.5)$, $\tilde{\mathbf{Z}} \leftarrow \mathbf{Z} \odot \eta$;
else
  Let N be the number of newly appended frames;
  New frames $\mathcal{N}$: $[t-N+1,\ t]$, Previous frames $\mathcal{P}$: $[t-T+1,\ t-N]$;
  if $N > 0$ **then**
    $\mathbf{Z}^+,\ \zeta^+ \leftarrow \mathbf{Z}_{\mathcal{N}},\ \zeta_{\mathcal{N}},\ \mathbf{Z}^-,\ \zeta^- \leftarrow \tilde{\mathbf{Z}}_{\mathcal{P}},\ \tilde{\zeta}_{\mathcal{P}}$;
    $\hat{\zeta} = \mathrm{Concat}(\zeta^-, \zeta^+)$;
    foreach (t, j) *where* $t' \in \mathcal{N}$, $j \in [1, J]$ **do**
      $\tilde{\zeta}_{t',j} \leftarrow \mathcal{F}(\hat{\zeta}_{t',j})$;
    end
    $\eta \leftarrow \mathrm{ReLU}(\tilde{\zeta}_{\mathcal{V}} - 0.5)$;
    $\tilde{\mathbf{Z}}^+ \leftarrow \mathbf{Z}^+ \odot \eta$, $\tilde{\mathbf{Z}} \leftarrow \mathrm{Concat}(\mathbf{Z}^-, \tilde{\mathbf{Z}}^+)$;
  end
end

To improve the robustness and stability of 3D joint estimation, we further introduce a new refinement strategy, as outlined in Alg. 1. Because each heatmap always produces a peak, joints that are outside the camera's view may be erroneously detected as visible, resulting in false positives. This misrepresentation can lead to inaccurate joint localization and unstable predictions in real-world scenes. To address this, we design a visibility-aware probability prediction module, which outputs an initial visibility probability sequence $\zeta \in \mathbb{R}^{T\times J}$ indicating the likelihood that each joint is visible in the current view. Since the visual encoder processes multiple frames jointly, a one Euro filter $\mathcal{F}$ is further added to smooth ζ. In addition, to accelerate inference, we design an incremental feature update scheme that extracts features only from newly appended frames and fuses them with cached features from previous frames. This design significantly reduces redundant computation while preserving temporal context.

Following the estimation of the refined 3D joint keypoint sequence $\tilde{\mathbf{Z}}$ calculated by Alg. 1, the same spatiotemporal encoder with the architecture as Eq. 3 is applied to frame-wise and joint-wise keypoint features $\mathbf{N}_t \in \mathbb{R}^{J\times S}$ from $\tilde{\mathbf{Z}}$. These features capture the temporal dynamics and inter-joint relationships of image-derived keypoints and are compatible with the HMD stream for subsequent cross-attention fusion. Notably, although both streams employ the identical spatiotemporal encoder design, their parameters are trained independently.

$$\mathbf{N}_t = \mathcal{T}_j\left(\mathbf{F}\left(\left[\mathcal{T}_f\left(E_l(\tilde{\mathbf{Z}})\right)\right]_t\right) + \mathbf{e}_j\right). \tag{4}$$

3.3.3 Cross-Modal Spatiotemporal Integration.

To integrate complementary information from the RGB-D stream with the HMD stream, we employ a standard multi-head cross-attention mechanism [54]. The HMD features $\mathbf{M}_t$ serve as queries, while RGB-D features $\mathbf{N}_t$ provide the keys and values, enabling HMD features to selectively attend to relevant visual cues. The resulting enriched representation $\tilde{\mathbf{M}}_t = \mathrm{CrossAttn}(\mathbf{M}_t, \mathbf{N}_t)$ denotes HMD features refined with visual guidance.

The fused features $\tilde{\mathbf{M}}_t$ are then fed into two parallel feed-forward networks, $\mathcal{T}_g(\cdot)$ and $\mathcal{T}_l(\cdot)$, to decode joint rotations represented in a continuous 6D rotation format. The first branch, $\mathcal{T}_g$, focuses on the root joint (i.e., the pelvis), whose global rotation determines the overall body orientation. Given its relatively stable and smooth trajectory, the pelvis benefits from a dedicated modeling pathway. The second branch,

$\mathcal{T}_l$, processes the local rotation of the remaining joints to capture finer articulation patterns. This architectural decoupling reduces interference across spatial scales and facilitates more precise attention allocation to relevant features.

The 6D joint rotation representations are subsequently propagated through a forward kinematics (FK) layer with the VR headset's world coordinates to derive the full-body 3D joint positions. To mitigate high-frequency jitter and improve temporal coherence across frames, we further apply a one Euro filter $\mathcal{F}$ to the output position trajectories, yielding smoother full-body joint position sequences $\mathbf{P}_t \in \mathbb{R}^{22\times 3}$ at the current t-th frame:

$$\mathbf{P}_t = \mathcal{F}\left(\mathrm{FK}\left(\mathcal{T}_g(\tilde{\mathbf{M}}_t), \mathcal{T}_l(\tilde{\mathbf{M}}_t)\right)\right). \tag{5}$$

3.4 Kinematic Pose Optimization via Energy Functions

Despite using headset and controller poses in the system input, the network may still yield inaccurate joint positions due to occlusions, sensor noise, and the ambiguity of inferring full-body motion from only head and hand tracking. Inverse kinematics (IK) can improve temporal continuity but remains under-constrained with head- and hand-only inputs, producing ambiguous or implausible lower-body poses without motion priors [60]. To address this, we propose an energy-based kinematic skeleton optimization that refines joint positions while preserving skeletal consistency.

(a) Original Pose (b) E_{A} Only (c) $E_{\mathrm{A}} + E_{\mathrm{S}}$

Fig. 3: Pose refinement results under different kinematic pose optimization settings. The semi-transparent mesh is the ground-truth pose, while the predicted mesh is color-coded by the pose error (redder indicates higher error).

3.4.1 Head and Hands Alignment Term

To ensure that the optimized skeleton remains consistent with external positional observations from the input HMD signals, we define a position alignment energy that enforces consistency between optimized joint positions and external observations from the HMD and controllers. Specifically, let O denote the set of observable joints (head, left hand, right hand) with ground-truth positions $\mathbf{q}_k \in \mathbb{R}^3$ provided by the VR system, where $k \in O$. Let $\tilde{\mathbf{p}}_k$ denote the initial network prediction, and $\mathbf{p}_k$ denotes the optimized result. The energy function for head and hands alignment is formulated as:

$$E_{\mathrm{A}} = \sum_{k\in O} \lambda_a \left\|\mathbf{p}_k - \mathbf{q}_k\right\|^2 + \sum_{k\notin O} \lambda_s \left\|\mathbf{p}_k - \tilde{\mathbf{p}}_k\right\|^2. \tag{6}$$

The first term encourages $\mathbf{p}_k$ to align with $\mathbf{q}_k$ for visible joints, ensuring consistency with real-world sensor input. The second term acts as a self-regularization constraint for unobserved joints ($k \notin O$), guiding their optimized positions to remain close to their initial predictions $\tilde{\mathbf{p}}_k$. This formulation ensures that the optimization adjusts the original predictions only when supported by external observations, while preserving plausible pose structure in the absence of ground-truth data. The weight factors λ_a and λ_s control the relative importance of alignment with observed joints and regularization of unobserved joints, respectively.

3.4.2 Skeletal Structure Preservation Term

To ensure local skeletal structure is preserved during optimization, we define a structure-preserving energy that penalizes deviations in both bone length and bone direction between neighbouring joints. Let $\tilde{\mathbf{J}}_{ij}$ and $\mathbf{J}_{ij}$ denote the initial and optimized joint direction vector prediction from joint j to i, respectively. The constraint energy function for local structure preserving is defined as:

$$E_{\mathrm{S}} = \sum_i \sum_{j\in\mathcal{M}(i)} \left[\lambda_l \left\|\ell(\mathbf{J}_{ij}) - \ell(\tilde{\mathbf{J}}_{ij})\right\|^2 + \lambda_d \left\|\mathbf{J}_{ij} - \tilde{\mathbf{J}}_{ij}\right\|^2\right], \tag{7}$$

where $\mathcal{M}(i)$ denotes the set of joints that are directly connected to joint i in the skeletal hierarchy, typically defined by the SMPL kinematic tree, $\ell(\cdot)$ represents the distance calculation of adjacent joints. The first term enforces bone length consistency, while the second term encourages directional consistency of neighboring joints. The weight factors λ_l and λ_d control the relative importance of bone length and direction preservation, respectively.

4 Synthetic Dataset Generation for VR Avatar

To address the lack of datasets suitable for immersive VR applications that require spatiotemporally aligned pose data, we introduce *EgoPoseVR*, a synthetic dataset featuring egocentric RGB-D imagery, HMD motion trajectories, and full-body 2D/3D pose annotations, as listed in Table 1. EgoPoseVR comprises 18,235 motion sequences and 1.8 million frames rendered at 320×256 resolution and 60 FPS, establishing a large-scale benchmark for training and evaluating VR avatar pose estimation systems.

The dataset is constructed with five key design objectives. First, to ensure motion diversity, we curate 2,350 sequences from AMASS [36], guided by category-level annotations from BABEL [40], covering a broad range of daily and task-oriented movements. Second, to promote scene and illumination generalization [6], we include 8 large-scale indoor 3D environments together with 100 HDR panoramic maps, providing both indoor- and outdoor-style lighting conditions. Third, to enhance appearance variability, the dataset incorporates 100 clothing sets (50 male and 50 female) and 20 footwear assets, increasing visual realism in egocentric views. Fourth, to ensure cross-headset compatibility, we provide both pelvis-centered and camera-centered joint coordinates. The latter are obtained by applying forward kinematics from the pelvis to the head, followed by a pre-calculated relative pose between the head and egocentric camera, enabling accurate egocentric projections across arbitrary headset configurations. Finally, to support efficient and temporally consistent egocentric rendering, all RGB-D sequences are rendered at 320×256 resolution with alignment to HMD and hand controller trajectories. To account for partial observability in egocentric views, EgoPoseVR additionally provides per-joint visibility labels for each frame, encoded as binary indicators of whether a joint lies within the camera's field of view. Further implementation details of EgoPoseVR dataset are provided in the supplementary material for reproducibility.

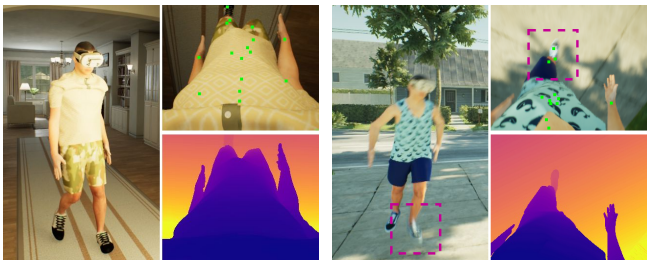


Fig. 4: Example images from our dataset showing third-person and egocentric RGB-D views. Green dots indicate 2D joint projections and pink dashed boxes mark the motion blur regions.

5 Cross-Device Data Transmission Pipeline

To support real-time full-body pose estimation across separate sensing and rendering modules, we design a lightweight cross-device architecture based on a persistent WebSocket communication framework. The system consists of three components: an input client, a rendering client

and an inference server. The input client integrates a VR HMD and an RGB-D camera, physically connected via USB Type-C. RGB-D images and HMD pose data are synchronized and transmitted in real time to the inference server via a persistent WebSocket connection. The inference server is responsible for running our EgoMotionVR model, and can host multiple models simultaneously, enabling flexible switching without altering the client architecture.

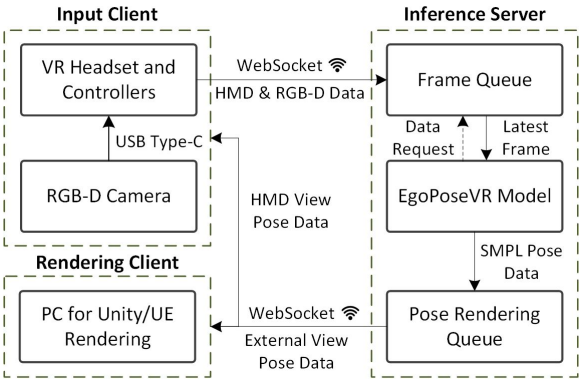


Fig. 5: System architecture illustrating cross-device data transmission for real-time SMPL pose estimation and visualization.

Upon receiving data, the inference server temporarily stores the latest RGB-D frame in a buffer queue. Our EgoPoseVR model continuously retrieves the most recent frame on demand, estimates the full-body pose in SMPL format, and performs internal pose refinement to reduce noise and improve temporal stability. The refined SMPL pose is returned via WebSocket to the input client for in-headset visualization and to the rendering client for third-person monitoring, supporting synchronized third-person perspective feedback via Unity or UE render. The entire system operates asynchronously and supports bidirectional communication between the clients and server, enabling low-latency streaming of sensory data and immediate return of model predictions.

6 Evaluations

6.1 Experiments

We implemented our system and trained all models using an NVIDIA GeForce RTX 3090 GPU. The training was conducted on our proposed EgoPoseVR dataset synthesized from human motions in the AMASS dataset [36]. Each dataset was split into 90% for training and 10% for testing. The network was optimized using AdamW with the learning rate of $5.0e-4$. The batch size was set to 32. The overall training objective comprises 2D joint estimation loss and 3D pose estimation loss that enforce spatial accuracy and kinematic plausibility. For a complete formulation and loss weighting strategy, please refer to the supplementary material.

In the absence of prior approaches that jointly integrate HMD motion and egocentric visual cues, we compared our EgoPoseVR system with the two state-of-the-art single-modality approaches, i.e., the HMD-based method EgoPoser [19] and the egocentric vision-based method EgoPoseFormer [62], as well as their complementary configurations. To comprehensively evaluate performance, we adopt four standard metrics: MPJPE (Mean Per-Joint Position Error, in cm), PA-MPJPE (Procrustes Aligned MPJPE, in cm), MPJRE (Mean Per-Joint Rotation Error, in degrees), and FPS (Frames Per Second).

6.2 Comparison with Existing Methods

To ensure a fair and rigorous comparison with existing approaches, we first adopt EgoPoser [19] to reveal the limitations of using HMD motion alone. To incorporate visual cues, we integrate EgoPoseFormer [62] as a refinement module. Notably, EgoPoseFormer alone is excluded as a standalone baseline because it predicts only joint positions rather than rotations, making it incompatible with SMPL-based motion reconstruction, and it depends on ground-truth pelvis-to-camera transformations during training, which are unavailable in VR inference. To overcome these limitations, we provide EgoPoseFormer with pelvis poses predicted by EgoPoser, enabling a fair and applicable integration.

To rigorously assess the effectiveness of each proposed component, we employ a staged evaluation strategy based on controlled module replacement. Starting from *EgoPoser + EgoPoseFormer* baseline, we first replace the motion stream with our HMD encoder while retaining the original visual module, resulting in *Proposed HMD + EgoPoseFormer*. This configuration isolates the contribution of our motion representation under an otherwise identical visual refinement pipeline. We then substitute the visual stream with our RGB-based encoder, forming *Proposed HMD + Proposed RGB (M)*, where (M) denotes the multi-frame input. This intermediate variant highlights the effectiveness of our visual module and ensures a fair comparison with the *EgoPoser + EgoPoseFormer* baseline. Finally, we evaluate our full system (i.e., *EgoPoseVR (HMD + RGB-D (M) + KPO)*), which integrates all proposed modules and achieves superior accuracy in VR-based full-body pose estimation.

6.2.1 Quantitative Evaluation

To objectively assess the effectiveness of different methods, we present quantitative comparisons in Table 2. Metrics are reported separately for the upper body (-U) and lower body (-L) to better isolate the impact of each module. Additional per-joint metrics are provided in the supplementary material.

As shown in Row B, incorporating visual cues into EgoPoser significantly improves the accuracy of lower-body joint estimation, which primarily depends on visual inference in egocentric settings due to the absence of direct sensing. In Row C, replacing EgoPoser with our proposed HMD module leads to further performance gains across all metrics, with a particularly notable reduction in MPJRE-U. This improvement stems from our dedicated joint embedding design, which more effectively encodes sensor input. Such enhancement is especially valuable for SMPL-based avatar control, where accurate joint rotation is essential for realistic motion synthesis.

To ensure a fair comparison with the baseline in Row B, Row D retains the same input modalities but substitutes EgoPoseFormer with our proposed visual module. Although the accuracy improvement over Row C is relatively modest, the system achieves a substantial increase in FPS, demonstrating a more favorable trade-off between precision and runtime speed. Finally, Row E presents the complete configuration of EgoPoseVR. By incorporating depth input and the KPO module, the system effectively mitigates challenges such as motion blur and occlusion while enforcing kinematic consistency in the upper body. Despite the added computational cost, this full setup delivers the highest accuracy, particularly in upper-body estimation (e.g., MPJPE-U), while maintaining real-time capability, making it well-suited for interactive VR scenarios.

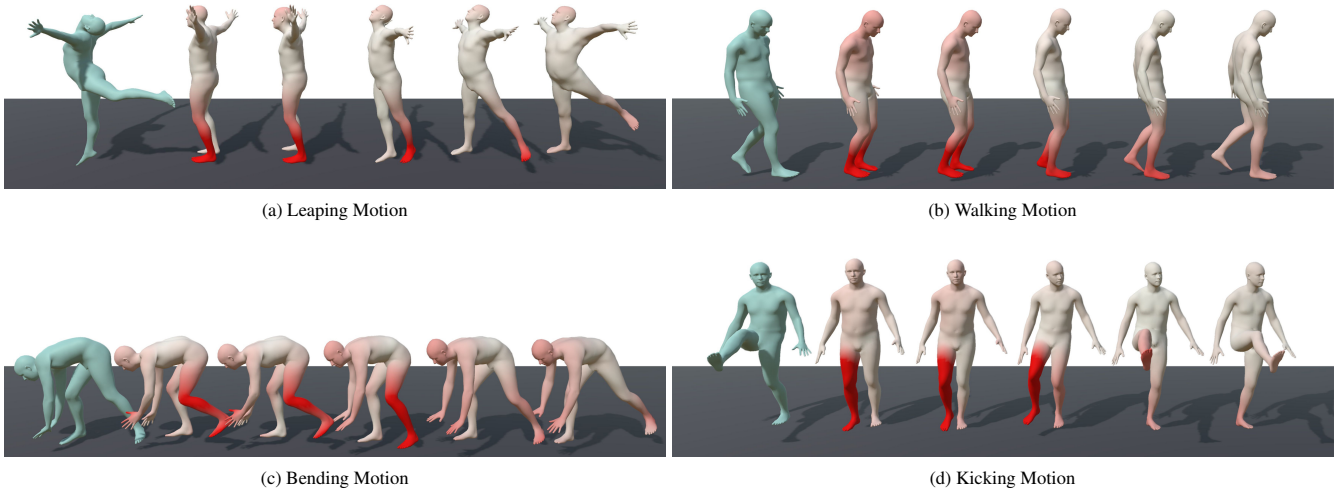
6.2.2 Qualitative Evaluation

The methods corresponding to Rows A–E in Table 2 are visualized under scenarios where the lower body exhibits noticeable motion, as shown in Fig. 6. EgoPoser infers lower-body motion primarily from contralateral arm–leg coordination. However, when arm swings are not evident (e.g., Fig. 6(b)), its predictions become unreliable. The approaches in Rows B–C, which build on EgoPoseFormer, depend heavily on the pelvis position within the image view. Consequently, when the user makes large head movements upward (Fig. 6(a)), downward (Figs. 6(b) and 6(c)) or to the side (Fig. 6(d)), the predicted lower-body poses are unreliable.

In contrast, the proposed RGB module (Row D) leverages visible joints in the image to enhance the HMD-based estimation of lower-body joints, without depending on pelvis localization, thereby accommodating larger viewpoint variations. Incorporating depth information and KPO further improves performance by enhancing 3D spatial understanding of the lower body and enforcing upper-body constraints based on HMD sensor input, resulting in more accurate and robust full-body pose estimation.

Table 2: Quantitative results of module-level comparison with state-of-the-art methods on our synthetic dataset. Row colors indicate performance improvement from the previous row: Yellow (<10%), Light Green (10–20%), Dark Green (>20%).

	Method	MPJPE-U ↓	MPJPE-L ↓	PA-MPJPE-U ↓	PA-MPJPE-L ↓	MPJRE-U ↓	MPJRE-L ↓	FPS ↑
A	EgoPoser [19]	5.61 ± 3.76	8.45 ± 5.99	4.13 ± 2.56	7.36 ± 5.08	15.48 ± 6.24	10.72 ± 4.89	216 ± 16
B	EgoPoser [19] + EgoPoseFormer [62]	4.92 ± 3.35	6.44 ± 4.92	3.66 ± 2.41	5.34 ± 3.94	15.39 ± 6.21	9.97 ± 4.63	29 ± 1
C	Proposed HMD + EgoPoseFormer [62]	3.60 ± 2.80	5.82 ± 4.88	2.68 ± 1.81	4.83 ± 3.80	8.59 ± 4.06	7.73 ± 4.24	29 ± 2
D	Proposed HMD + Proposed RGB (M)	3.57 ± 2.71	5.23 ± 4.86	2.55 ± 1.80	4.20 ± 3.73	8.33 ± 4.04	7.29 ± 4.28	155 ± 6
E	EgoPoseVR (HMD + RGB-D (M) + KPO)	1.66 ± 1.05	4.75 ± 4.35	2.05 ± 1.53	3.63 ± 3.31	8.21 ± 3.96	6.74 ± 3.94	97 ± 6



(a) Leaping Motion

(b) Walking Motion

(c) Bending Motion

(d) Kicking Motion

Fig. 6: Qualitative results of module-level comparison with state-of-the-art methods. The leftmost blue avatar represents the ground truth, while the remaining five avatars from left to right correspond respectively to the methods listed in Rows A–E of Table 2. Pose errors are color-coded relative to the ground truth, with deeper red indicating greater error.

6.3 Ablation Study

We conduct an ablation study on both input modalities and model components to analyze their individual contributions (Table 3).

For input modalities, starting from an HMD-only baseline, we first add single-frame RGB input, which contributes per-frame visual features without temporal context. This yields a marked improvement in positional accuracy, particularly for lower-body joints where HMD signals alone provide weak kinematic constraints. We then replace single-frame input with a temporally stacked RGB sequence, allowing the network to exploit short-term motion continuity. This further reduces both positional noise and orientation instability, especially under abrupt head movement. The inclusion of depth in the multi-frame configuration brings additional gains across all metrics by providing geometric cues that disambiguate joint positions under occlusion and motion blur.

Table 3: Ablation study of input modalities and model components. (S): single-frame RGB input; (M): multi-frame RGB input. Row colors indicate accuracy improvement from the previous row: Yellow (<10%), Light Green (10–20%), Dark Green (>20%).

Method	MPJPE ↓	PA-MPJPE ↓	MPJRE ↓	FPS ↑
HMD Only	6.26 ± 4.03	4.85 ± 3.05	9.26 ± 3.85	212 ± 6
HMD + RGB (S)	4.42 ± 3.47	3.37 ± 2.57	8.12 ± 3.82	158 ± 8
HMD + RGB (M)	4.25 ± 3.28	3.22 ± 2.47	7.90 ± 3.68	155 ± 6
HMD + RGB-D (M)	3.95 ± 3.05	2.96 ± 2.21	7.61 ± 3.51	154 ± 7
+ KPO (Full System)	**2.92 ± 2.16**	**2.70 ± 2.17**	**7.61 ± 3.51**	**97 ± 6**
– Cross-Modal Fusion	4.08 ± 3.03	3.10 ± 2.19	7.65 ± 3.62	156 ± 5
– Spatiotemporal Encoder	6.42 ± 4.19	5.35 ± 3.28	14.29 ± 5.07	189 ± 6

For model components, the proposed KPO module in Sec. 3.4 is further applied on top of *HMD + RGB-D (M)*, and particularly improves full-body joint position consistency while preserving predicted joint orientations. Although the refinement introduces computational cost, the final system still achieves 97 FPS, ensuring its practicality for real-time VR interactions. Additional ablation study results are provided in the supplementary material. To assess the necessity of the spatiotemporal encoder and the cross-modal fusion module, we remove each component from the full system. While this expectedly improves inference speed, it leads to reduced full-body joint pose accuracy. Replacing cross-modal fusion with simple feature concatenation results in only a modest performance drop, as modality-specific features are already effectively encoded by the spatiotemporal encoder before fusion.

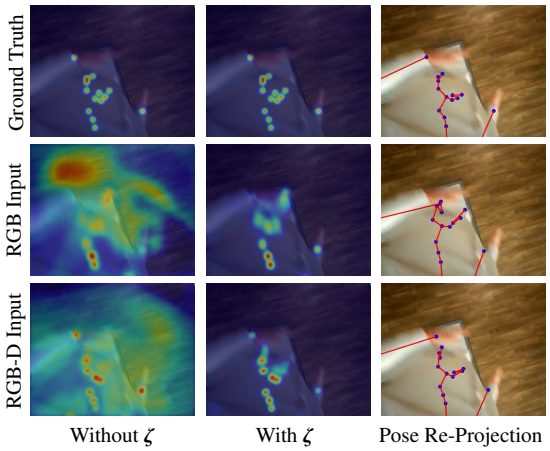
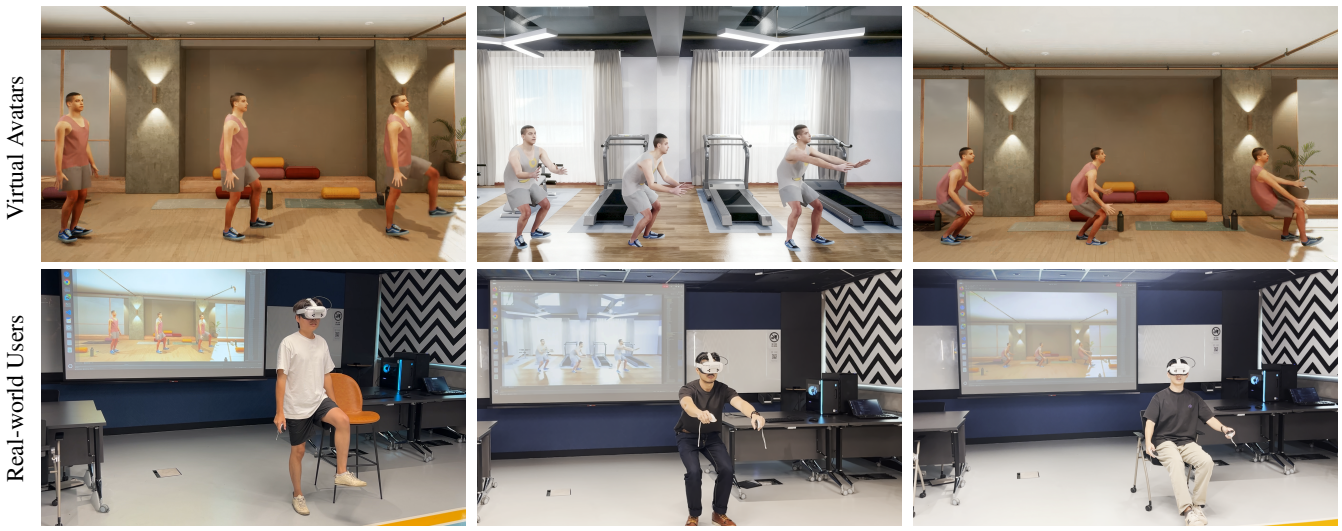


Fig. 7: Effect of visual input modality and the proposed visibility probability factor ζ on heatmap quality and the resulting 3D pose estimation, evaluated via 2D re-projections of the predicted joints.

6.4 Heatmap-Driven Joint Accuracy Analysis

To examine how different factors affect 2D heatmap quality, we compare input configurations across two dimensions: image modality and the integration of the proposed visibility probability factor ζ in



(a) High-Knee Motion (b) Squat Motion (c) Sitting Motion

Fig. 8: Comparison with existing methods in the real-world scenes. From left to right, the avatars are generated by EgoPoser [19], EgoPoser [19] + EgoPoseFormer [62], and the proposed EgoPoseVR, respectively. The foreground users illustrate the corresponding real-world motions, while the background avatars demonstrate their synchronized virtual counterparts in real time.

Sec. 3.3.2. For quantitative validation, the predicted 3D joints are re-projected onto the original 2D image plane via known camera intrinsics, allowing a direct assessment of spatial alignment with the input observations.

The leftmost column of Fig. 7 shows that without incorporating ζ, the heatmaps corresponding to out-of-view joints introduce substantial noise and artifacts. This occurs because fixed-threshold filtering cannot adaptively suppress heatmap responses from invisible joints. To address this limitation, we introduce a joint visibility head that predicts the probability of each joint being within the visible region. As shown in the middle column, integrating ζ effectively suppresses noisy heatmap signals and sharpens the heatmap distributions, thereby focusing on the valid joint regions. By comparing the RGB and RGB-D rows in Fig. 7, it is evident that depth cues enable the network to better capture the 3D spatial structure of the scene, resulting in more accurate joint localization. This refinement ensures temporally stable 2D joint sequences, which serve as reliable input to the spatiotemporal encoder and ultimately lead to reliable 3D pose estimation.

6.5 User Study in Real-life Environments

To evaluate the effectiveness of different approaches in the real-world scenes, we conducted a user study comparing EgoPoser [19] (EP), EgoPoser [19] + EgoPoseFormer [62] (EP+EF), and our proposed EgoPoseVR, i.e., Row A, B and E in Table 2. The study protocol was reviewed and approved by the Institutional Review Board (IRB) of our university. All compared models were trained using the proposed dataset. Building on the cross-device data transmission design described in Sec. 5, to ensure a fair and consistent comparison, virtual avatars generated by the different methods are simultaneously rendered side by side in UE render and driven by identical user input from the same VR device.

A total of 20 participants were recruited from the university staff and students (mean age = 24.9 years, standard deviation = 3.73), consisting of 11 males and 9 females. After reading and signing the consent form, participants first completed a demographic questionnaire to record their basic information. Participants were informed of the general purpose during the consent process but were blind to the specific study hypotheses. Participants exhibited heterogeneous prior VR experience, with 14 participants reporting prior VR exposure and 6 reporting none. In addition, 8 participants also reported prior experience with pose tracking or related development tasks. Following a short demonstration of our system, they were asked to wear the headset and freely perform basic whole-body movements (e.g., walking, sitting, leg raising, and bending) within 5 mins. After the session, participants filled out the 5-point Likert-scale questionnaire shown in Table 4. Based on standardized embodiment questionnaires [4, 15], our questionnaire was designed to assess four dimensions of the system: *tracking accuracy* (Q1–Q2), *responsiveness and stability* (Q3–Q4), *sense of embodiment* (Q5), and *intention for future use* (Q6). We adapted items most relevant to the system's functional evaluation rather than adopting a full embodiment inventory. During the experiment, participants were not informed of the arrangement of avatars corresponding to different methods in the UE environment, ensuring unbiased evaluation.

Table 4: Post-session questionnaire used for the subjective evaluation of different methods.

Question	Description
Q1	The avatar's upper body movements accurately reflected my real-world movements.
Q2	The avatar's lower body movements accurately reflected my real-world movements.
Q3	The avatar's movements were synchronized (real-time) with my actions in real time.
Q4	The avatar's motion appeared stable and free of noticeable jittering.
Q5	I felt that the avatar's body was a natural extension of my own body.
Q6	I would prefer to use this pose estimation method and its associated device in future VR applications.

The questionnaire ratings were collected on a 5-point Likert scale, yielding ordinal data that cannot be assumed to follow a normal distribution. Accordingly, we employed non-parametric analyses, applying Friedman tests to assess overall differences across methods and Wilcoxon signed-rank tests with the Holm–Bonferroni adjustment for pairwise comparisons. The Friedman test results revealed significant differences for all questionnaire items ($p < 0.001$). Post hoc tests further showed that EgoPoseVR yielded significantly higher rating than both EP and EP+EF ($p < 0.01$), as shown in Fig. 9. These findings demonstrate that participants consistently perceived evident differences among the three models across all evaluation aspects. Detailed statistical results are provided in the supplementary material.

The Friedman test indicated a significant effect for Q1 (*upper-body tracking accuracy*) ($\chi^2(2) = 21.794$, $p < 0.0001$, Kendall's W = 0.545). When comparing EP+EF and EP, EP+EF employed the same

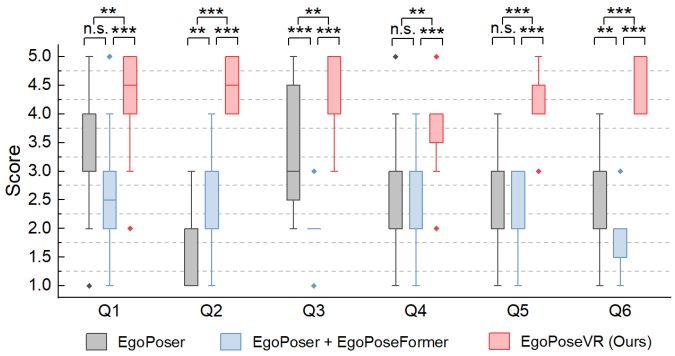


Fig. 9: Statistical analysis of the user study. The box represents the interquartile range (IQR), the central line indicates the median. Outliers are shown as individual points. (n.s.): no statistical significance, (*): $p \leq 0.05$, (**): $p \leq 0.01$, (***): $p \leq 0.001$.

HMD-based upper-body estimation strategy as EP, with no significant difference ($p = 0.051$). However, the additional latency introduced by EP+EF reduced perceived accuracy relative to EP. For *lower-body tracking accuracy* (Q2), the Friedman test also showed a significant effect for Q2 ($\chi^2(2) = 35.368$, $p < 0.0001$, Kendall's W = 0.884). EP performed worst as it relied solely on HMD motion, while EP+EF offered moderate improvements but still struggled with large joint-angle poses such as single-leg squats and knee lifts (Fig. 8). In contrast, EgoPoseVR enhanced by the dedicated HMD and KPO modules, yielded superior upper-body accuracy and significantly outperformed both baselines in lower-body motions (all $p < 0.001$), owing to the spatiotemporal encoder and cross-modal module.

For *responsiveness and stability* (Q3–Q4), significant condition effects were observed (Q3: $\chi^2(2) = 31.886$, $p < 0.001$, Kendall's $W = 0.797$; Q4: $\chi^2(2) = 25.323$, $p < 0.001$, Kendall's $W = 0.633$). EP+EF's computational overhead hindered timely avatar responses, resulting in markedly lower responsiveness scores. In terms of stability, EP exhibited jitter from its heuristic propagation of lower-body poses from upper-body motion, while EP+EF displayed frame-to-frame inconsistency amplified by latency, with no statistically significant difference between the two conditions ($p = 0.749$). The proposed method alleviated both issues (all $p < 0.001$), though when the arms occluded the lower body or when parts moved outside the field of view, reconstructed motion was prone to becoming less smooth.

For *sense of embodiment* (Q5) and *intention for future use* (Q6), significant condition effects were observed (Q5: $\chi^2(2) = 31.521$, $p < 0.001$, Kendall's $W = 0.788$; Q6: $\chi^2(2) = 34.274$, $p < 0.001$, Kendall's $W = 0.857$). Post hoc tests showed that participants rated the proposed method significantly higher than EP and EP+EF (all $p < 0.001$). In term of Q5, the lack of a significant difference between EP and EP+EF ($p = 0.090$) indicates that partial or unstable lower-body tracking is insufficient to substantially improve embodiment. Stronger embodiment arose from superior tracking accuracy and temporal stability, which also translated into a greater willingness to adopt the system in future VR applications, showing that subjective acceptance was consistent with improvements in the preceding technical questions.

7 Limitations and Future Work

Our method has several limitations that open avenues for future exploration. First, the effectiveness of the RGB-D refinement module relies on the visibility of the user body within the camera's view. In cases where the user body is outside the field of view or severely occluded (Fig. 10), the benefit of cross-modal fusion diminishes. Future work may consider incorporating temporal priors or generative models to better handle such visually sparse conditions. Second, although the avatar motions in our dataset are driven by real-life motion capture data, the corresponding visual observations are synthesized, which inevitably introduces a sim-to-real gap. To mitigate this issue, we have incorporated visibility-aware modeling (Sec. 3.3.2) and kinematic alignment with HMD sensor measurements (Sec. 3.4) to reduce sensitivity to complex motions and environment scenes. Nevertheless, future work will focus on collecting real-world, temporally synchronized HMD motion and RGB-D data across diverse physical environments to further improve robustness under realistic deployment conditions. Third, while SMPL enables efficient prediction and animation, it does not incorporate biomechanical constraints. As a result, the generated poses may lack physical plausibility or high-frequency muscular dynamics. Extending our framework to incorporate musculoskeletal or physics-based priors could lead to more anatomically realistic pose reconstruction.

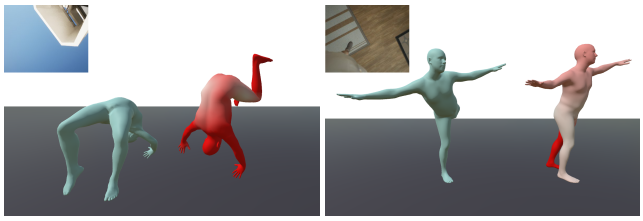


Fig. 10: Failure cases under partial or complete observability loss in egocentric views. The blue avatar denotes the ground truth, and the right avatar shows the predicted pose with color-coded errors.

8 Conclusion

We present *EgoPoseVR*, a unified framework for egocentric full-body pose estimation in VR, leveraging spatially sparse motion trajectories from HMD and spatial observations from a headset-mounted downward-facing RGB-D camera. Our method introduces a spatiotemporal fusion strategy that refines pose estimates derived from HMD motion input using visibility-aware RGB-D features, with a particular focus on improving lower-body accuracy under occlusion and limited viewpoints. To support this task, we construct a large-scale, VR-specific dataset featuring temporally synchronized RGB-D and HMD motion data, enabling effective learning under realistic egocentric conditions. Extensive evaluations across synthetic and real-world VR scenarios demonstrate that *EgoPoseVR* achieves higher accuracy in both joint positions and rotations than state-of-the-art egocentric pose estimation baselines, while maintaining real-time performance. It offers portable and infrastructure-free full-body tracking without the need for additional body-worn sensors. These findings underscore its promise for facilitating natural avatar embodiment and seamless interaction in rehabilitation training and other immersive VR applications.

References

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

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

[3] M. M. Azam and K. Desai. A survey on 3d egocentric human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1643–1654, 2024. 1, 2, 3

[4] A. Borrego, J. Latorre, M. Alcañiz, and R. Llorens. Embodiment and presence in virtual reality after stroke. a comparative study with healthy subjects. *Frontiers in neurology*, 10:1061, 2019. 8

[5] H. Cheng, C. Xu, X. Chen, Z. Chen, J. Wang, and L. Zhao. Realistic volume rendering with environment-synced illumination in mixed reality. In *2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 423–428. IEEE, 2023. 1

[6] H. Cheng, C. Xu, J. Wang, Z. Chen, and L. Zhao. Fast and accurate illumination estimation using ldr panoramic images for realistic rendering. *IEEE Transactions on Visualization and Computer Graphics*, 29(12):5235–5249, 2022. 5

[7] H. Cuevas-Velasquez, C. Hewitt, S. Aliakbarian, and T. Baltrušaitis. Simpleego: Predicting probabilistic body pose from egocentric cameras. In *2024 International Conference on 3D Vision (3DV)*, pp. 1446–1455. IEEE, 2024. 2, 3

[8] P. Dai, Y. Zhang, T. Liu, Z. Fan, T. Du, Z. Su, X. Zheng, and Z. Li. 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*, pp. 874–884, 2024. 1, 2

[9] T. D. Do, C. I. Protko, and R. P. McMahan. Stepping into the right shoes: The effects of user-matched avatar ethnicity and gender on sense of embodiment in virtual reality. *IEEE transactions on visualization and computer graphics*, 2024. 1

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[11] Y. Du, R. Kips, A. Pumarola, S. Starke, A. Thabet, and A. Sanakoyeu. 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*, pp. 481–490, 2023. 2

[12] S. Eom, D. Sykes, S. Rahimpour, and M. Gorlatova. Neurolens: Augmented reality-based contextual guidance through surgical tool tracking in neurosurgery. In *2022 IEEE International symposium on mixed and augmented reality (ISMAR)*, pp. 355–364. IEEE, 2022. 1

[13] Z. Fan, P. Dai, Z. Su, X. Gao, Z. Lv, J. Zhang, T. Du, G. Wang, and Y. Zhang. Emhi: A multimodal egocentric human motion dataset with hmd and body-worn imus. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 2879–2887, 2025. 2

[14] R. Feng, Y. Gao, T. H. E. Tse, X. Ma, and H. J. Chang. Diffpose: Spatiotemporal diffusion model for video-based human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14861–14872, 2023. 3

[15] M. Gonzalez-Franco and T. C. Peck. Avatar embodiment. towards a standardized questionnaire. *Frontiers in Robotics and AI*, 5:74, 2018. 8

[16] D. Hollidt, P. Streli, J. Jiang, Y. Haghighi, C. Qian, X. Liu, and C. Holz. Egosim: An egocentric multi-view simulator and real dataset for body-worn cameras during motion and activity. *Advances in Neural Information Processing Systems*, 37:106607–106627, 2024. 3

[17] Y. Huang, M. Kaufmann, E. Aksan, M. J. Black, O. Hilliges, and G. Pons-Moll. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics (TOG)*, 37(6):1–15, 2018. 1, 2

[18] H. Jang and Y. M. Kim. Remp: Reusable motion prior for multi-domain 3d human pose estimation and motion inbetweening. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2474–2483. IEEE, 2025. 3

[19] J. Jiang, P. Streli, M. Meier, and C. Holz. Egoposer: Robust real-time egocentric pose estimation from sparse and intermittent observations everywhere. In *European Conference on Computer Vision*, pp. 277–294. Springer, 2024. 2, 6, 7, 8

[20] J. Jiang, P. Streli, H. Qiu, A. Fender, L. Laich, P. Snape, and C. Holz. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *European conference on computer vision*, pp. 443–460. Springer, 2022. 1, 2, 4

[21] T. Kang, K. Lee, J. Zhang, and Y. Lee. Ego3dpose: Capturing 3d cues from binocular egocentric views. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–10, 2023. 2, 3

[22] T. Kang and Y. Lee. Attention-propagation network for egocentric heatmap to 3d pose lifting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 842–851, 2024. 2

[23] M. Kaufmann, Y. Zhao, C. Tang, L. Tao, C. Twigg, J. Song, R. Wang, and O. Hilliges. Em-pose: 3d human pose estimation from sparse electromagnetic trackers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11510–11520, 2021. 3

[24] A. Lammert, G. Rendle, F. Immohr, A. Neidhardt, K. Brandenburg, A. Raake, and B. Froehlich. Immersive study analyzer: Collaborative immersive analysis of recorded social vr studies. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 1

[25] J. Lee, W. Xu, A. Richard, S.-E. Wei, S. Saito, S. Bai, T.-L. Wang, M. Sung, J. Saragih, et al. Rewind: Real-time egocentric whole-body motion diffusion with exemplar-based identity conditioning. *arXiv preprint arXiv:2504.04956*, 2025. 2, 3

[26] M. Lee, H. Lee, B. Kim, and S. Kim. Unspat: Uncertainty-guided spatiotemporal transformer for 3d human pose and shape estimation on videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3004–3013, 2024. 3

[27] S. Lee, S. Starke, Y. Ye, J. Won, and A. Winkler. Questenvsim: Environment-aware simulated motion tracking from sparse sensors. In *ACM SIGGRAPH 2023 Conference Proceedings*, pp. 1–9, 2023. 2

[28] J. Li, K. Liu, and J. Wu. Ego-body pose estimation via ego-head pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17142–17151, 2023. 2, 3

[29] W. Li, H. Liu, H. Tang, P. Wang, and L. Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13147–13156, 2022. 3

[30] W. Li, M. Liu, H. Liu, P. Wang, J. Cai, and N. Sebe. Hourglass tokenizer for efficient transformer-based 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 604–613, 2024. 3

[31] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017. 4

[32] C. Liu, F. F.-Y. Tan, S. Zhao, A. Kanneganti, G. A. Tushar, and E. T. Khoo. Facilitating virtual reality integration in medical education: A case study of acceptability and learning impact in childbirth delivery training. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2024. 1

[33] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 851–866. 2023. 4

[34] Z. Luo, J. Cao, R. Khirodkar, A. Winkler, K. Kitani, and W. Xu. Real-time simulated avatar from head-mounted sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 571–581, 2024. 2, 3

[35] L. Ma, Y. Ye, F. Hong, V. Guzov, Y. Jiang, R. Postyeni, L. Pesqueira, A. Gamino, V. Baiyya, H. J. Kim, et al. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *European Conference on Computer Vision*, pp. 445–465. Springer, 2024. 3

[36] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5442–5451, 2019. 5, 6

[37] C. Millerdurai, H. Akada, J. Wang, D. Luvizon, C. Theobalt, and V. Golyanik. Eventego3d: 3d human motion capture from egocentric event streams. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1186–1195, 2024. 2

[38] J. Peng, Y. Zhou, and P. Mok. Ktpformer: Kinematics and trajectory prior knowledge-enhanced transformer for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1123–1132, 2024. 3

[39] J. L. Ponton, H. Yun, A. Aristidou, C. Andujar, and N. Pelechano. Sparseposer: Real-time full-body motion reconstruction from sparse data. *ACM Transactions on Graphics*, 43(1):1–14, 2023. 1

[40] A. R. Punnakkal, A. Chandrasekaran, N. Athanasiou, A. Quiros-Ramirez, and M. J. Black. Babel: Bodies, action and behavior with english labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 722–731, 2021. 5

[41] P. Rahimian and J. K. Kearney. Optimal camera placement for motion capture systems. *IEEE transactions on visualization and computer graphics*, 23(3):1209–1221, 2016. 1

[42] T. Rhee, S. Thompson, D. Medeiros, R. Dos Anjos, and A. Chalmers. Augmented virtual teleportation for high-fidelity telecollaboration. *IEEE transactions on visualization and computer graphics*, 26(5):1923–1933, 2020. 1

[43] H. Rhodin, C. Richardt, D. Casas, E. Insafutdinov, M. Shafiei, H.-P. Seidel, B. Schiele, and C. Theobalt. Egocap: egocentric marker-less motion capture with two fisheye cameras. *ACM Transactions on Graphics (TOG)*, 35(6):1–11, 2016. 2

[44] A. Rojo, J. Cortina, C. Sánchez, E. Urendes, R. García-Carmona, and R. Raya. Accuracy study of the oculus touch v2 versus inertial sensor for a single-axis rotation simulating the elbow's range of motion. *Virtual Reality*, 26(4):1651–1662, 2022. 1

[45] D. Roth and M. E. Latoschik. Construction of the virtual embodiment questionnaire (veq). *IEEE Transactions on Visualization and Computer Graphics*, 26(12):3546–3556, 2020. 2

[46] M. Shi, S. Starke, Y. Ye, T. Komura, and J. Won. Phasemp: Robust 3d pose estimation via phase-conditioned human motion prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14725–

14737, 2023. 3

[47] S. Starke, P. Starke, N. He, T. Komura, and Y. Ye. Categorical codebook matching for embodied character controllers. *ACM Transactions on Graphics (TOG)*, 43(4):1–14, 2024. 1

[48] Z. Sun, Y. Liang, Z. Ma, T. Zhang, L. Bao, G. Li, and S. He. Repose: 3d human pose estimation via spatio-temporal depth relational consistency. In *European Conference on Computer Vision*, pp. 309–325. Springer, 2024. 1

[49] J. Tang, J. Wang, K. Ji, L. Xu, J. Yu, and Y. Shi. A unified diffusion framework for scene-aware human motion estimation from sparse signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21251–21262, 2024. 3

[50] Y. Tao, C. Y. Wang, A. D. Wilson, E. Ofek, and M. Gonzalez-Franco. Embodying physics-aware avatars in virtual reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2023. 1

[51] D. Tome, P. Peluse, L. Agapito, and H. Badino. xr-egopose: Egocentric 3d human pose from an hmd camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7728–7738, 2019. 2, 3

[52] J. C. Uhl, H. Schrom-Feiertag, G. Regal, K. Gallhuber, and M. Tscheligi. Tangible immersive trauma simulation: is mixed reality the next level of medical skills training? In *Proceedings of the 2023 chi conference on human factors in computing systems*, pp. 1–17, 2023. 1

[53] E. Van der Kruk and M. M. Reijne. Accuracy of human motion capture systems for sport applications; state-of-the-art review. *European journal of sport science*, 18(6):806–819, 2018. 1

[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[55] T. Von Marcard, B. Rosenhahn, M. J. Black, and G. Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer graphics forum*, vol. 36, pp. 349–360. Wiley Online Library, 2017. 1, 2

[56] C. Wang, S. Zheng, L. Zhong, C. Yu, C. Liang, Y. Wang, Y. Gao, T. L. Lam, and Y. Shi. Pepperpose: Full-body pose estimation with a companion robot. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2024. 1

[57] J. Wang, L. Liu, W. Xu, K. Sarkar, D. Luvizon, and C. Theobalt. Estimating egocentric 3d human pose in the wild with external weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13157–13166, 2022. 3

[58] J. Wang, L. Liu, W. Xu, K. Sarkar, and C. Theobalt. Estimating egocentric 3d human pose in global space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11500–11509, 2021. 2

[59] J. Wang, D. Luvizon, W. Xu, L. Liu, K. Sarkar, and C. Theobalt. Scene-aware egocentric 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13031–13040, 2023. 3

[60] A. Winkler, J. Won, and Y. Ye. Questsim: Human motion tracking from sparse sensors with simulated avatars. In *SIGGRAPH Asia 2022 Conference Papers*, pp. 1–8, 2022. 2, 5

[61] W. Xu, A. Chatterjee, M. Zollhoefer, H. Rhodin, P. Fua, H.-P. Seidel, and C. Theobalt. 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, 2019. 2, 3

[62] C. Yang, A. Tkach, S. Hampali, L. Zhang, E. J. Crowley, and C. Keskin. Egoposeformer: A simple baseline for stereo egocentric 3d human pose estimation. In *European Conference on Computer Vision*, pp. 401–417. Springer, 2024. 2, 3, 6, 7, 8

[63] X. Yi, Y. Zhou, M. Habermann, S. Shimada, V. Golyanik, C. Theobalt, and F. Xu. Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13167–13178, 2022. 1, 2

[64] X. Yi, Y. Zhou, and F. Xu. Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions On Graphics (TOG)*, 40(4):1–13, 2021. 1, 2

[65] J. Zhang, Z. Tu, J. Yang, Y. Chen, and J. Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13232–13242, 2022. 3

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

[67] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, and Z. Ding. 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11656–11665, 2021. 3

[68] X. Zheng, Z. Su, C. Wen, Z. Xue, and X. Jin. Realistic full-body tracking from sparse observations via joint-level modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14678–14688, 2023. 1, 2

[69] Z. Zheng, T. Yu, H. Li, K. Guo, Q. Dai, L. Fang, and Y. Liu. Hybridfusion: Real-time performance capture using a single depth sensor and sparse imus. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 384–400, 2018. 1

[70] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15085–15099, 2023. 1