# Real-Time Full-Body Avatar Control in VR Using Glasses-Mounted Stereo Cameras

Seongmin Baek
Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea
baeksm@etri.re.kr

Hyunjun Lim
Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea
tp02134@etri.re.kr

Hyejin Hur
ETRI
Daejeon, Republic of Korea
Korea University
Seoul, Republic of Korea
hyejin.hur@etri.re.kr

Youn-Hee Gil
Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea
yhgil@etri.re.kr

## Abstract

We present an egocentric stereo framework for real-time full-body avatar control in VR. The proposed method combines temporal 2D refinement, geometry-based 3D initialization, and causal 3D pose refinement. Trained jointly on UnrealEgo and UnrealEgo-RW, the pose estimation pipeline runs at over 30 fps on an i9-13900K/RTX 4090. Pose estimation and avatar retargeting are qualitatively demonstrated on one additional glasses-mounted stereo rig not used during training.

## CCS Concepts

• **Human-centered computing** → **Virtual reality**; • **Computing methodologies** → **Motion capture**; *Reconstruction.*

## Keywords

Virtual Reality, Full-Body Avatar Control, Egocentric Stereo Vision, 3D Human Pose Estimation, Real-Time Tracking

**ACM Reference Format:**
Seongmin Baek, Hyunjun Lim, Hyejin Hur, and Youn-Hee Gil. 2026. Real-Time Full-Body Avatar Control in VR Using Glasses-Mounted Stereo Cameras. In *32nd ACM Symposium on Virtual Reality Software and Technology (VRST '26), November 16–18, 2026, Sendai, Japan.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3822517.3849922

## 1 Introduction

Full-body avatar control enhances embodiment and interaction in VR, but conventional systems often require external cameras or multiple wearable trackers. AvatarPoser [3] estimates full-body pose from head and hand tracking, while lower-body motion must be inferred without direct observations. UnrealEgo [2] introduced a large-scale synthetic egocentric stereo dataset for 3D human motion capture, and subsequent work extended stereo pose estimation using temporal information and real-world data [1]. More recently, REWIND [5] demonstrated real-time whole-body motion estimation from head-mounted stereo images and camera poses.

Practical egocentric stereo tracking remains challenging. Self-occlusion and fisheye distortion degrade frame-wise 2D pose estimation, while small left–right keypoint mismatches can cause large triangulation errors. Differences between training datasets and real capture conditions further complicate deployment. We present an integrated framework combining temporal stereo 2D refinement, constrained geometry-based 3D initialization, and causal rotation-space refinement for real-time full-body pose estimation. We evaluate accuracy on held-out benchmark data and qualitatively demonstrate pose estimation and avatar retargeting on an additional glasses-mounted stereo rig not used during training.

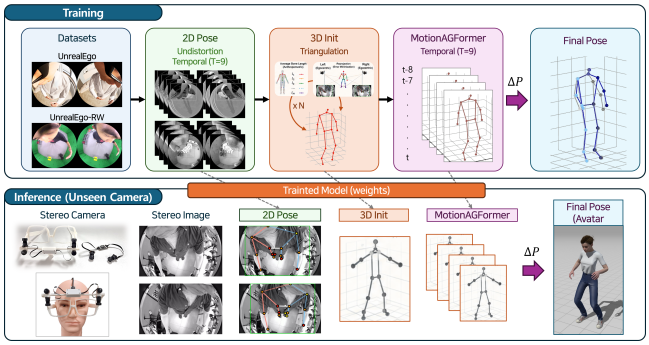



**Figure 1: Proposed pipeline and glasses-mounted stereo rig.**

## 2 Method

Figure 1 shows the proposed pipeline and our 67 g glasses-mounted stereo rig, consisting of two synchronized downward-facing grayscale fisheye cameras with an approximately 12 cm baseline, providing 2560 × 720 side-by-side stereo frames at up to 60 fps.

### 2.1 Temporal Stereo 2D Refinement

RTMDet detects the body, and RTMPose [4] estimates 16 body joints and their confidence scores from each stereo view. *GraphMLP2DRefine* processes a nine-frame causal stereo sequence ($T = 9$) to capture short-term motion. The last three frames ($t-2$, $t-1$, and $t$) are jointly supervised using position, velocity, and acceleration losses. The position term uses a Laplace negative log-likelihood to jointly learn keypoint uncertainty, which is combined with 2D confidence for subsequent stereo reconstruction. Only the current-frame output is used at inference.

### 2.2 Geometry-Based 3D Initialization

The refined stereo keypoints are undistorted and triangulated using the target-camera intrinsics and extrinsics to obtain an initial 3D pose.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
*VRST '26, Sendai, Japan*
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2811-2/2026/11
https://doi.org/10.1145/3822517.3849922

Since direct ray-based triangulation is sensitive to small 2D errors, the initial joints are iteratively corrected using a skeletal prior and reprojection consistency. Joints are adjusted toward average bone lengths computed from the training data and reprojected onto both camera views, with confidence weighting to reduce the influence of unreliable keypoints. This process is repeated three times ($K = 3$) to obtain the constrained initial pose.

## 2.3 Causal 3D Pose Refinement

We encode 14 bone rotations in 6D [7], excluding the fixed neck-to-head segment. Initial rotations align rest and observed directions by shortest-arc rotation in the current parent frame, without added twist. Forward kinematics (FK) preserves the initial neck position and per-subject mean bone lengths computed from training GT. We employ *MotionAGFormerLite*, a reduced variant of MotionAGFormer [6], which processes a nine-frame causal sequence ($T = 9$) using a spatial GCN and causal temporal attention. The network predicts residual rotation corrections for the last three frames, which are composed with the initial rotations and converted to 3D joint positions through FK.

Training uses rotation, angular-velocity, and angular-acceleration losses, together with an FK-based 3D position loss. Structured corruption simulates estimation failures by perturbing 2D positions and confidence scores, as well as triangulated 3D positions and depths before rotation conversion. Only the current-frame output is used at inference.

# 3 Results

We follow the official UnrealEgo and UnrealEgo-RW splits, combining their training splits for joint training and their validation splits for validation, and reporting test results separately. Nine-frame causal windows stay within individual sequences and splits; unavailable history is padded with the sequence's first frame.

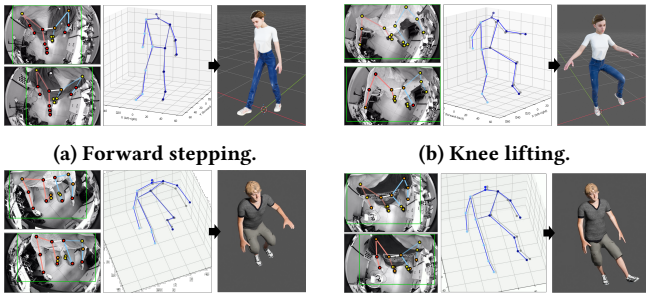
Table 1 reports device-relative MPJPE without root centering, scale normalization, or Procrustes alignment. Constrained initialization substantially improves upon direct triangulation; final refinement further reduces MPJPE from 10.26 to 8.79 cm on UnrealEgo and from 10.90 to 9.25 cm on UnrealEgo-RW. For context, prior studies report 7.91 cm on UnrealEgo [2], and 4.62 cm on UnrealEgo and 10.41 cm on UnrealEgo-RW [1]. The RW results are not directly comparable because test frame counts differ (7,936 in prior work vs. 24,012 in ours).

On an i9-13900K/RTX 4090, processing runs at 31.40 fps with per-frame detection and 58.04 fps with detection every seven frames. Between detections, the previous bounding boxes are reused. Timing ends at pose output, excluding transmission and subsequent avatar filtering and retargeting.

Figure 2 shows pose estimation and avatar retargeting on this calibrated rig, using four sequences of approximately 900 frames each from one participant wearing four different outfits. Without ground-truth 3D poses, these results demonstrate feasibility on this rig but do not establish cross-camera accuracy or robustness across users and capture conditions.

**Table 1: MPJPE on UnrealEgo (48,138 test frames) and UnrealEgo-RW (24,012), and runtime on unseen-camera recordings. All accuracy rows use identical refined 2D keypoints with per-frame detection.**

| *3D Pose Accuracy by Pipeline Stage* Method | UnrealEgo (cm)↓ | RW (cm)↓ |
|---|---|---|
| Direct triangulation | 67.55 | 83.11 |
| Constrained 3D initialization | 10.26 | 10.90 |
| + MotionAGFormerLite (Ours) | **8.79** | **9.25** |
| *Runtime* Detection setting | Avg. Time (ms)↓ | FPS↑ |
| Every frame | 31.85 | 31.40 |
| Interval = 7 | 17.23 | 58.04 |



**(a) Forward stepping.** **(b) Knee lifting.**

**(c) Crouching.** **(d) Side stepping.**

**Figure 2: Qualitative results on an unseen stereo rig: stereo 2D poses, initial and refined 3D poses, and Blender avatar retargeting.**

## Acknowledgments

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government [26ZC1100, Development of Spatial Media Technology and Interaction Technology for Convergence of the Real and Virtual World].

## References

[1] Hiroyasu Akada, Jian Wang, Vladislav Golyanik, and Christian Theobalt. 2024. 3D Human Pose Perception from Egocentric Stereo Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 767–776.

[2] Hiroyasu Akada, Jian Wang, Soshi Shimada, Masaki Takahashi, Christian Theobalt, and Vladislav Golyanik. 2022. UnrealEgo: A New Dataset for Robust Egocentric 3D Human Motion Capture. In *European Conference on Computer Vision (ECCV)*. 1–17.

[3] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. 2022. AvatarPoser: Articulated Full-Body Pose Tracking from Sparse Motion Sensing. In *European Conference on Computer Vision (ECCV)*. 443–460.

[4] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. 2023. RTMPose: Real-Time Multi-Person Pose Estimation based on MMPose. *arXiv preprint arXiv:2303.07399* (2023).

[5] Jihyun Lee, Weipeng Xu, Alexander Richard, Shih-En Wei, Shunsuke Saito, Shaojie Bai, Te-Li Wang, Minhyuk Sung, Tae-Kyun Kim, and Jason Saragih. 2025. REWIND: Real-Time Egocentric Whole-Body Motion Diffusion with Exemplar-Based Identity Conditioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7095–7104.

[6] Soroush Mehraban, Vida Adeli, and Babak Taati. 2024. MotionAGFormer: Enhancing 3D Human Pose Estimation with a Transformer-GCNFormer Network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 6920–6930.

[7] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. 2019. On the Continuity of Rotation Representations in Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5745–5753.