OmniEgoCap: Camera-Agnostic Sequence-Level Egocentric Motion Reconstruction

Kyungwon Cho, Hanbyul Joo

Seoul National University

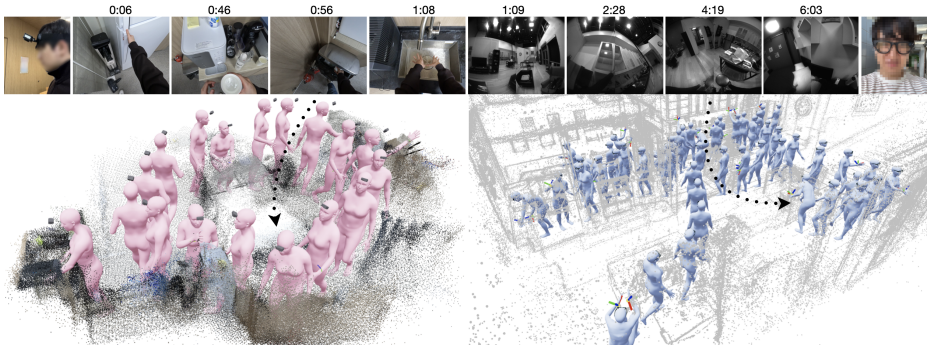

Fig. 1: OmniEgoCap is a sequence-level diffusion framework that leverages long-range physical invariants and geometry-aware augmentation to reconstruct consistent 3D full-body motion across diverse egocentric camera setups. (L) GoPro and (R) Aria.

Abstract. The proliferation of commercial egocentric devices offers a unique lens into human behavior, yet reconstructing full-body 3D motion remains difficult due to frequent self-occlusion and the "out-of-sight" nature of the wearer's limbs. While head and hand trajectories provide sparse anchor points, current methods often overfit to specific hardware optics or rely on expensive, post-hoc optimizations that compromise motion naturalness. In this paper, we present OmniEgoCap, a unified diffusion framework that scales egocentric reconstruction to diverse capture setups. By shifting from short-term windowed estimation to sequence-level inference, our method captures a global perspective and recovers invariant physical attributes, such as height and body proportions, that provide critical constraints for disambiguating head-only cues. To ensure hardware-agnostic generalization, we introduce a geometry-aware visibility augmentation strategy that treats intermittent hand appearances as principled geometric constraints rather than missing data. Our architecture jointly predicts temporally coherent motion and consistent body shape, establishing a new state-of-the-art on public benchmarks and demonstrating robust performance across diverse, in-the-wild environments.

Keywords: Human Motion Reconstruction · Egocentric Vision · Diffusion Models

1 Introduction

Egocentric vision systems are rapidly becoming commercially available, with a wide range of products including action cameras [2], VR headsets [1, 3, 7], and smart glasses [4, 5, 8, 9]. To enable effective human computer interaction, these systems need to understand 3D human motion and behavior. However, estimating full-body 3D motion from an egocentric camera remains challenging, as large portions of the wearer's body are rarely visible from the first person viewpoint.

Reconstructing 3D human motion from egocentric videos therefore relies on sparse but informative signals, primarily head trajectories and intermittently visible hands. Most approaches leverage head trajectories from SLAM or SfM as a proxy for body motion, given their strong correlation [47,76]. While effective, head-only cues are inherently ambiguous, since multiple plausible full body poses can share nearly identical head trajectories. Hand cues provide critical complementary information. Hand signals can be obtained directly from XR devices [1, 3, 4, 7] or estimated from RGB videos via robust off-the-shelf models [83]. When visible, hand positions and motions offer direct evidence of body configuration. Crucially, even their absence is informative, constraining feasible poses within the camera's field of view. Nevertheless, many prior works treat hand cues as auxiliary signals for post-hoc optimization [45, 76]. Such pipelines are computationally expensive and tend to degrade underlying motion priors, resulting in unnatural motion. Recent methods incorporate intermittent hand and head signals together for end-to-end prediction without expensive refinement [21, 32, 41, 53]. While effective under specific setups, these approaches overfit to static hand visibility boundaries dictated by particular camera FoVs. We empirically find their performance degrades significantly across varying optics or mountings.

Another crucial yet underexplored cue is human height. Height constrains plausible motions, since identical head trajectories can correspond to different actions depending on body proportions. For instance, a low head position for a tall person often indicates sitting, whereas the same head height for a shorter person might imply standing or bending. Despite its importance, existing methods often assume a fixed mean body shape [19, 32, 42, 45, 47] or predict shape independently at each frame [53, 76], leading to inconsistent body proportions over time.

In this paper, we present **OmniEgoCap**, a unified diffusion framework for sequence-level egocentric full body motion reconstruction across diverse camera setups. First, unlike prior approaches that operate on short temporal windows, our framework processes entire motion sequences. Long-range temporal reasoning enables reliable estimation of invariant attributes such as height, body proportions, and camera visibility boundaries, inferred from periods of standing or regular walking, and the spatial distribution of hand appearances. Second, to robustly leverage intermittent hand cues across various devices, we introduce a geometry-aware visibility augmentation strategy simulating diverse hand visibility boundaries and occlusion patterns. By exposing the model to diverse visibility configurations during training, we encourage it to interpret intermittent hand observations as geometric constraints rather than missing inputs. OmniEgoCap thus implicitly adapts to different fields of view and generalizes to unseen camera

setups without device-specific retraining. Finally, we design an encoder-decoder architecture with a diffusion-based decoder that jointly predicts consistent body shape and long-range 3D motion. Combined with a tailored coordinate representation and training strategy, our framework produces temporally coherent and physically plausible reconstructions. Extensive experiments on public benchmarks and in the wild data demonstrate the effectiveness of each component, establishing state-of-the-art performance for egocentric full-body motion reconstruction.

In summary, our contributions are threefold: (1) A unified sequence-level diffusion framework with a dedicated encoder–decoder architecture, where a motion-conditioned encoder and diffusion-based decoder jointly model long-range dynamics and consistent body shape for egocentric full-body reconstruction; (2) A geometry-aware visibility augmentation strategy that enables principled integration of intermittent hand cues and strong generalization across diverse camera setups; and (3) Extensive experiments and analysis across diverse devices and capture setups, demonstrating robust cross-device generalization and state-of-the-art performance on public benchmarks and in-the-wild data.

2 Related Work

Motion Capture from Body-Mounted Sensors or VR-Devices. Wearable motion capture using IMU sensors provides a camera-free alternative [6, 10], but lightweight systems with fewer sensors [36, 43, 66, 68, 78–80, 85] often require global localization cues from head-mounted cameras [33, 45, 77] to address the inherent root drift from integration errors. Another related direction leverages VR/AR headsets, which provide 6-DoF (degrees of freedom) tracking for both the head and hand-held controllers (or wrists). To tackle this highly under-constrained problem, existing methods primarily rely on physics-based simulation [73], direct deep regression [15, 16, 23, 41, 42, 86], or generative priors [19, 25, 27, 29, 62]. While EgoPoser [41] addresses intermittent tracking loss, it assumes a fixed, predefined field-of-view boundary, restricting its generalization to arbitrary setups.

Motion Capture from Egocentric Videos. There has been growing interest in capturing 3D human body pose from body-worn cameras for their ability to record self-motion anywhere without third-person observations. To maximize body visibility, a parallel line of research relies on specialized hardware, such as downward-facing or stereo fisheye cameras [11–13, 40, 44, 57, 64, 69, 71, 72, 75]. While these approaches achieve high-fidelity reconstructions, they require specific camera configurations distinct from ubiquitous everyday wearables. To enable practical applications, another line of research uses frontal-facing cameras, evolving from early chest or head-mounted setups [39, 49, 52, 65, 70, 81, 82] to ubiquitous smart glasses [4, 5]. To infer motion, these methods leverage head trajectories [47], visual features [32, 53], or intermittent hand cues [21]. However, they are strictly tied to specific device setups and fixed field-of-views, preventing camera-agnostic generalization. Although EgoAllo [76] similarly uses sparse hand cues, it heavily relies on expensive post-hoc optimization, often resulting in unnatural kinematics.

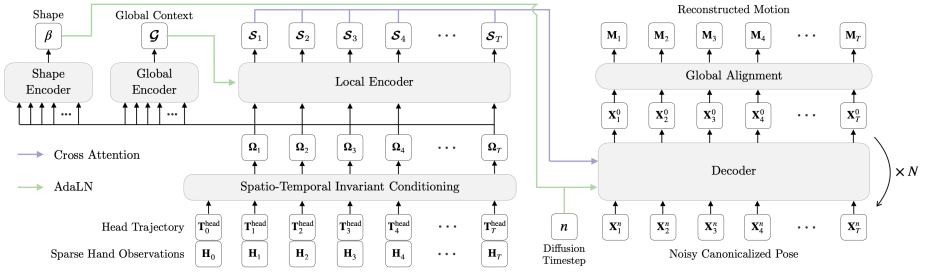

Fig. 2: Overview of our model. The model $\mathcal{F}$ takes the head trajectory $\mathbf{T}^{\text{head}}_{0:T}$ and intermittent hand observations $\mathbf{H}_{0:T}$, which are first converted into spatio-temporally invariant conditioning features $\boldsymbol{\Omega}_{1:T}$. The encoder $\mathcal{E}$ processes $\boldsymbol{\Omega}_{1:T}$ to predict a single body shape β and per-frame summary features $\boldsymbol{\mathcal{S}}_{1:T}$. A diffusion decoder $\mathcal{D}$ then conditions on these features to denoise a noisy canonicalized pose $\mathbf{X}^n_{1:T}$ and, via global alignment, reconstruct the full-body motion $\mathbf{M}_{1:T}$.

Generative Priors and Temporal Modeling. Our work builds on diffusion models as powerful generative priors for high-fidelity motion synthesis [22, 38, 63, 84]. In our context, 3D reconstruction is framed as a generative completion task [19, 20, 27, 31, 34, 59], resolving full-body motion from sparse or partial cues. While effective, scaling these completion-based models to long-horizon sequences is hampered by the $O(N^2)$ attention complexity of standard transformers. To address this, we adopt a sliding-window attention mechanism [17, 48] for efficient, sequence-level motion reconstruction, enabling long-term coherence without the computational overhead of global attention.

3 Method

3.1 Problem Formulation

Given an egocentric video input from an arbitrary device, our goal is to reconstruct the wearer's 3D full-body motion. Let $\mathbf{I}_{0:T} = \{I_t\}_{t=0}^T$ denote the egocentric image stream. From $\mathbf{I}$, we compute the head poses $\mathbf{T}^{\text{head}}_{0:T} = \{\mathbf{T}^{\text{head}}_t \in \text{SE}(3)\}_{t=0}^T$ defined in world coordinates and intermittent wrist 6D pose observations $\mathbf{H}_{0:T} = \{\mathbf{H}^{\text{lHand}}_t, \mathbf{H}^{\text{rHand}}_t\}_{t=0}^T$. Here, each $\mathbf{H}^{\text{lHand/rHand}}_t = (\mathbf{T}^{\text{lHand/rHand}}_t, v^{\text{lHand/rHand}}_t)$, where $\mathbf{T}^{\text{lHand/rHand}}_t \in \text{SE}(3)$ is its 6-DoF pose defined in the head coordinates and $v^{\text{lHand/rHand}}_t \in \{0, 1\}$ is its binary visibility state. A non-visible state can occur due to occlusion, motion blur, or the wrist moving outside of camera's field of view (FoV). Our model OmniEgoCap takes $\mathbf{T}^{\text{head}}_{0:T}$ and $\mathbf{H}_{0:T}$ as input, and produces the reconstructed human motion $\mathbf{M}_{1:T}$ as output:

$$\mathbf{M}_{1:T} = \mathit{OmniEgoCap}\left(\mathbf{T}^{\text{head}}_{0:T}, \mathbf{H}_{0:T}\right), \tag{1}$$

where $\mathbf{M}_{1:T} = \{\beta, \boldsymbol{r}_{1:T}, \boldsymbol{\Phi}_{1:T}, \boldsymbol{\Theta}_{1:T}\}$ is in SMPL format [58] with the shape parameter $\beta \in \mathbb{R}^{16}$, root translation $\boldsymbol{r}_{1:T} = \{\boldsymbol{r}_t \in \mathbb{R}^3\}_{t=1}^T$, root orientation $\boldsymbol{\Phi}_{1:T} = \{\boldsymbol{\phi}_t \in \text{SO}(3)\}_{t=1}^T$. $\boldsymbol{\Theta}_{1:T} = \{\boldsymbol{\theta}_t = (\boldsymbol{\theta}^1_t, \ldots, \boldsymbol{\theta}^{J-1}_t); \boldsymbol{\theta}^j_t \in \text{SO}(3)\}_{t=1}^T$ are the joint

angles, where J is the number of body joints in SMPL. The root poses are defined in the world coordinate system, consistent with $\mathbf{T}^{\text{head}}$. Note that we predict a single shape β for a whole sequence, contrasting prior approaches that use a fixed mean shape [42, 45, 47] or per-frame shape [53, 76]. We implement OmniEgoCap via a transformer-based diffusion architecture [35,54], which progressively denoises the motion $\mathbf{M}$ conditioned on the sparse cues from the head trajectory $\mathbf{T}^{\text{head}}$ and the detected hand cues $\mathbf{H}$. See Fig. 2 for an overview of our framework.

3.2 Preprocessing and Representation

Preprocessing. Given the egocentric images $\mathbf{I}$, we first compute camera poses $\mathbf{T}_t^{\text{cam}} \in \text{SE}(3)$[1] via off-the-shelf SLAMs. We define a world coordinate system where the z-axis aligns with gravity and the floor is at $z = 0$. Since the camera and head joints are not co-located, we compute head poses via $\mathbf{T}_t^{\text{head}} = \mathbf{T}^{\text{cam}\to\text{head}}\mathbf{T}_t^{\text{cam}}$, where $\mathbf{T}^{\text{cam}\to\text{head}}$ is a mount-specific but time-invariant rigid transform estimated via a lightweight pre-calibration step. The 6D hand observations and visibility $\mathbf{H}_t^{\text{Hand}}$ are also extracted using off-the-shelf pose estimation modules. For monocular RGB videos, we utilize HaWoR [83] without their infilling module to extract both $\mathbf{T}_t^{\text{cam}}$ and $\mathbf{H}_t^{\text{hand}}$. For Aria [28], we use the internal Aria software tools to compute them.

Coordinate and Representation. Learning motion from long, sparse signal sequences end-to-end requires spatio-temporally invariant conditioning. To this end, we convert the raw signals $\left(\mathbf{T}_t^{\text{head}}, \mathbf{H}_t^{\text{lHand}}, \mathbf{H}_t^{\text{rHand}}\right)$ into a normalized representation that is invariant to both spatial and temporal variations.

For the head trajectory, we follow the previous work [76], by defining the per-frame canonical frame at the head's floor projection, with its z-axis aligned with gravity and y-axis aligned to the forward direction of the head: $\mathbf{T}_t^{\text{cano}\to\text{world}}$. Specifically, the condition vector $\boldsymbol{\Omega}_t \in \mathbb{R}^D$ is computed by a simple neural net function Γ as follows:

$$\boldsymbol{\Omega}_t = \Gamma\left(\Delta\mathbf{T}_t^{\text{head}}, \mathbf{R}_t^{\text{head}\to\text{cano}}, h_t, \Delta\mathbf{R}_t^{\text{cano}}, \mathbf{T}_t^{\text{hand}\to\text{head}}\right), \tag{2}$$

where $\Delta\mathbf{T}_t^{\text{head}}$ is the relative head pose from $t-1$, $\mathbf{R}_t^{\text{head}\to\text{cano}}$ is the canonicalized head orientation, and h_t is the height of the head, extracted from $\mathbf{T}_t^{\text{head}}$. Unlike EgoAllo [76], we additionally include the relative rotation in the canonical coordinate $\Delta\mathbf{R}_t^{\text{cano}}$ from frame $t-1$, since our model explicitly predicts the root orientation $\boldsymbol{\Phi}$, as discussed in the next section. We also newly introduce hand conditions $\mathbf{T}_t^{\text{hand}\to\text{head}} = (\mathbf{T}_t^{\text{lHand}\to\text{head}}, \mathbf{T}_t^{\text{rHand}\to\text{head}})$, representing the relative hand poses with respect to the head coordinate. When the hand is not visible (*i.e.*, when $v_t^{\text{lHand/rHand}} = 0$), the corresponding cues are replaced with learnable null embeddings. Notably, we do not include temporal motion cues for the hands, as wrist observations are often intermittent.

[1] $\mathbf{T}_t^{\text{cam}}$ can be considered as the coordinate transformation from the camera coordinate to the world coordinate, which can be equivalently denoted as $\mathbf{T}_t^{\text{cam}\to\text{world}}$. Similarly, $\mathbf{T}_t^{\text{head}} = \mathbf{T}_t^{\text{head}\to\text{world}}$.

Global Alignment. Similar to previous work [41, 42, 76], we leverage $\mathbf{T}^{\text{head}}$ obtained from SLAM for global localization by directly deriving the root translation $\boldsymbol{r}_t$. However, unlike EgoAllo [76], we predict the relative root orientation $\mathbf{R}_t^{\text{root}\to\text{cano}}$ with respect to the canonical coordinate as the model output. Concretely, the final root orientation is computed as $\boldsymbol{\phi}_t = \mathbf{R}_t^{\text{cano}\to\text{world}}\mathbf{R}_t^{\text{root}\to\text{cano}}$. We demonstrate that our representation provides an additional 3DoF in global orientation, leading to improved motion stability and accuracy.

3.3 Architecture

We adopt a conditional diffusion model [35] to learn a plausible motion distribution from the processed conditioning features, $\boldsymbol{\Omega}_{1:T} = \{\boldsymbol{\Omega}_t\}_{t=1}^T$. For the model's architecture, we use an encoder-decoder transformer [67], which is well-suited for time-series data. To efficiently process arbitrarily long sequences, we implement both encoder $\mathcal{E}$ and decoder $\mathcal{D}$ using sliding window local attention with an attention horizon of W [17, 18, 37, 48] and Rotary Positional Embedding [61], which reduces the computational complexity from quadratic to linear.

Encoder. The encoder $\mathcal{E}$ processes conditioning features $\boldsymbol{\Omega}_{1:T}$ through three sub-components to produce the predicted shape parameter $\hat{\beta}$ and per-frame summaries $\boldsymbol{\mathcal{S}}_{1:T}$:

$$\begin{gathered}\hat{\beta} = \mathcal{E}_{\text{shape}}(\boldsymbol{\Omega}_{1:T}) \in \mathbb{R}^{16}, \quad \boldsymbol{\mathcal{G}} = \mathcal{E}_{\text{global}}(\boldsymbol{\Omega}_{1:T}) \in \mathbb{R}^{D} \\ \boldsymbol{\mathcal{S}}_{1:T} = \mathcal{E}_{\text{local}}(\boldsymbol{\Omega}_{\mathbf{1:T}}, \boldsymbol{\mathcal{G}}) \in \mathbb{R}^{T\times D}\end{gathered} \tag{3}$$

The sub-encoders $\mathcal{E}_{\text{shape}}$ and $\mathcal{E}_{\text{global}}$ capture sequence-level, frame-invariant representations, $\hat{\beta}$ and global context $\boldsymbol{\mathcal{G}}$. We implement them with a few local attention transformer layers to first aggregate local cues, which are then fed into an attention-based pooling layer [46]. Importantly, shape cues need to be inferred from a global perspective, since human height can only be reliably estimated from specific frames (e.g., upright postures), which are often ambiguous if observed within local temporal windows. This design choice marks a key distinction from prior work [76]. $\mathcal{E}_{\text{local}}$ produces summaries $\boldsymbol{\mathcal{S}}_{1:T}$ by processing conditions $\boldsymbol{\Omega}_{1:T}$ through locally attentive mechanisms that aggregate information across neighboring frames, enabling processing of long-term sequence-level input. To incorporate global cues, the global context $\boldsymbol{\mathcal{G}}$ is injected via AdaLN-Zero [54], enabling global modulation of the local aggregation process.

Decoder. The decoder $\mathcal{D}$ does not directly denoise final motion $\mathbf{M}$. Instead, it denoises the canonicalized pose representation, denoted as $\mathbf{X}_{1:T}^0 = \{\mathbf{R}_{1:T}^{\text{root}\to\text{cano}}, \boldsymbol{\Theta}_{1:T}\}$. The root translation $\boldsymbol{r}$ and orientation $\boldsymbol{\Phi}$ are then computed as described in the Global Alignment section. This decoder $\mathcal{D}$ is implemented as a DiT architecture [54], conditioned on the encoded summary feature $\boldsymbol{\mathcal{S}}_{1:T}$ and body shape $\hat{\beta}$:

$$\hat{\mathbf{X}}_{1:T}^0 = \mathcal{D}(\mathbf{X}_{1:T}^n, n, \hat{\beta}, \boldsymbol{\mathcal{S}}_{1:T}), \tag{4}$$

where $n \in [1, N]$ is the diffusion timestep and $\hat{\mathbf{X}}_{1:T}^0$ is the predicted $\mathbf{X}_{1:T}^0$. The noised input $\mathbf{X}_{1:T}^n$ is defined by the DDPM [35] forward process $q(\mathbf{X}_{1:T}^n|\mathbf{X}_{1:T}^0) =$

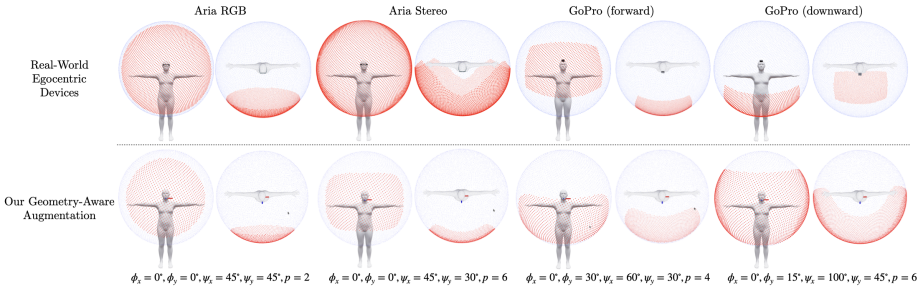

Fig. 3: Visualization of real-world and simulated visibility boundaries. The top row shows calibrated FoVs from real-world devices (*e.g.*, Aria, GoPro). The bottom row illustrates diverse visibility patterns generated by our geometry-aware augmentation through parameter sampling. Red regions indicate the visible field of view.

$\mathcal{N}(\mathbf{X}^n_{1:T}; \sqrt{\bar{\alpha}_n}\mathbf{X}^0_{1:T}, (1-\bar{\alpha}_n)\mathbf{I})$, where $\bar{\alpha}_n$ is the noise schedule. We inject global conditions n and $\hat{\beta}$ via AdaLN, while the per-frame summary $\boldsymbol{\mathcal{S}}_{1:T}$ is injected via cross-attention, using a local attention mask.

3.4 Training

Hand Visibility Augmentation. Recent motion-annotated egocentric datasets [50] are often tied to specific hardware, limiting their generalization to arbitrary devices. To build a truly device-agnostic model, we leverage diverse AMASS [51] motions and simulate intermittent hand visibilities via augmentation. Specifically, we introduce a geometry-aware augmentation to simulate real-world visibility boundary complexities, such as arbitrary field-of-views (FoV), camera tilt, varying aspect ratios, lens masks, and optical distortions. This addresses ideal pinhole models [21, 41], which assume a front-facing camera with a fixed symmetric FoV. We first compute the wrist's yaw ψ_x and pitch ψ_y in head coordinates. We then define the visiblity boundary using five parameters: center offset (ϕ_x, ϕ_y) for tilt, half-angles (γ_x, γ_y) for field-of-view size and aspect ratio, and a power p for lens distortion and shape. A wrist is visible when its angles satisfy the following generalized ellipse equation:

$$\left|\frac{\psi_x - \phi_x}{\gamma_x}\right|^p + \left|\frac{\psi_y - \phi_y}{\gamma_y}\right|^p \leq 1 \tag{5}$$

By sampling these parameters from realistic distributions during training, we expose the model to diverse conditions, improving its generalization. (See Fig. 3 for examples)

Finally, to ensure robustness against imperfect in-the-wild tracking, we additionally apply stochastic temporal signal drops and pose perturbations. Instead of simple random masking [14], we model the heavy-tailed nature of real-world occlusions using Poisson and Log-Normal distributions. Furthermore, we inject Gaussian noise into the visible wrist poses to account for the inherent jitter of off-the-shelf estimators.

Loss Functions. Our training objective $\mathcal{L}$ combines $\mathcal{L}_{\text{simple}}$, $\mathcal{L}_{\text{shape}}$, and $\mathcal{L}_{\text{aux}}$:

$$\mathcal{L} = \mathcal{L}_{\text{simple}} + \lambda_{\text{shape}}\mathcal{L}_{\text{shape}} + \mathcal{L}_{\text{aux}} \tag{6}$$

$\mathcal{L}_{\text{simple}}$ is the standard DDPM [35] objective:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{n,\mathbf{X}^0,\boldsymbol{\Omega}}[\|\mathbf{X}^0 - \hat{\mathbf{X}}^0\|^2] \tag{7}$$

To enforce shape consistency without directly penalizing the PCA-derived β, which lacks physical meaning, $\mathcal{L}_{\text{shape}}$ minimizes the 3D T-pose joint error:

$$\mathcal{L}_{\text{shape}} = \|\text{FK}(\mathbf{0}, \beta) - \text{FK}(\mathbf{0}, \hat{\beta})\|^2 \tag{8}$$

, where FK is the forward kinematics function.
Finally, $\mathcal{L}_{\text{aux}}$ regularizes the reconstructed motion $\hat{\mathbf{M}}_{1:T}$ to align with 3D physical constraints [63] via joint position $\mathcal{L}_{\text{pos}}$ and foot skating $\mathcal{L}_{\text{skat}}$ losses:

$$\begin{aligned} \mathcal{L}_{\text{pos}} &= \frac{1}{T}\sum_{t=1}^{T} \|\text{FK}(\mathbf{M}_t, \beta) - \text{FK}(\hat{\mathbf{M}}_t, \hat{\beta})\|^2 \\ \mathcal{L}_{\text{skat}} &= \frac{1}{T-1}\sum_{t=1}^{T-1} \|\text{FK}(\hat{\mathbf{M}}_{t+1}, \hat{\beta}) - \text{FK}(\hat{\mathbf{M}}_t, \hat{\beta})\|^2 \cdot c_t \end{aligned} \tag{9}$$

, where c_t is the ground-truth binary contact label. These losses are scaled by $\bar{\alpha}_n$ to enforce physical accuracy primarily when the signal level is high:

$$\mathcal{L}_{\text{aux}} = \bar{\alpha}_n(\lambda_{\text{pos}}\mathcal{L}_{\text{pos}} + \lambda_{\text{skat}}\mathcal{L}_{\text{skat}}) \tag{10}$$

We find that $\mathcal{L}_{\text{aux}}$ is crucial for satisfying the hand condition $\mathbf{H}$ and generating stable and accurate motion. We set the weights $\lambda_{\text{shape}} = 2.0$, $\lambda_{\text{pos}} = 0.25$, and $\lambda_{\text{skat}} = 0.4$.

3.5 Inference

We use DDIM [60] sampling for motion generation. While the feed-forward denoiser's prediction is accurate, we integrate test-time guidance optimization at each sampling step to better satisfy the sparse hand constraints $\mathbf{H}$. The objective is to find a refined motion $\tilde{\mathbf{M}}^0$ by minimizing $\mathcal{L}_{\text{opt}}$:

$$\mathcal{L}_{\text{opt}} = \sum_{t=1}^{T}\sum_{j}\left[\sqrt{\bar{\alpha}_n}\|\text{FK}_j(\tilde{\mathbf{M}}_t^0, \hat{\beta}) - \text{FK}_j(\hat{\mathbf{M}}_t^0, \hat{\beta})\|^2 + sv_t^j\sqrt{1-\bar{\alpha}_n}\|\text{FK}_j(\tilde{\mathbf{M}}_t^0, \hat{\beta}) - p_t^j\|^2\right] \tag{11}$$

, where FK_j denotes the position of hand joint $j \in \{\text{lHand}, \text{rHand}\}$, p_t^j is the observed wrist position in world coordinates derived from $\mathbf{H}$, and $s = 30.0$ is the guidance scale. We optimize only the arm joints [42]. The objective balances two terms. The first term (prior) acts as a regularizer, using the denoiser's prior to ensure motion plausibility. The second term (constraint) acts as a perturbation, pulling the wrist toward the target p_t^j. This dynamically leverages $\mathcal{D}$'s denoising ability, as the constraint dominates in early steps, while the prior dominates in late steps, ensuring the motion remains on the manifold. This frame-independent guidance is highly parallelizable, avoiding slow post-hoc optimization.

4 Experiments

4.1 Experiment Setting

Dataset. We use the AMASS [51] dataset for training and simulated evaluation. Sequences are resampled to 30fps and preprocessed following HuMoR [56], adjusting the floor and annotating foot contact labels. We train on the AMASS train split with stochastic augmentation (Sec. 3.4). We evaluate on AMASS validation and test splits. The validation split is included specifically to enable long-sequence evaluation, as the standard test split alone is relatively short.

Training Details. We train using AdamW optimizer with a learning rate 10^{-4}, weight decay 10^{-4}, and a batch size of 32 for 16 hours on 8 A5000 GPUs. Max training sequence length is 512 with attention horizon $W = \pm 63$, implemented via FlexAttention [26]. Test-time optimization uses Theseus [55] Levenberg-Marquardt optimizer. Additional details are provided in the supplementary material.

Metrics. We use Mean Per Joint Position Error (**MPJPE**, mm), Mean Per Joint Velocity Error (**MPJVE**, cm/s), and **Jerk** (km/s^3) (the third derivative of position) to evaluate motion quality. To assess hand observations $\mathbf{H}$, we use Hand Position Error (**Hand PE**, mm) and Visible Hand Position Error (**Vis Hand PE**, mm, Hand PE on visible frames). For shape, we use **Height Error** (cm) and **Span Error** (cm) for accuracy and Height Standard Deviation (**Height Std**, cm) and Span Standard Deviation (**Span Std**, cm) for consistency. Height and Span are defined as the maximum vertical and horizontal vertex distances in a T-pose. **Runtime** (sec) measures the average execution time per sequence, inclusive of optimization setup on an RTX3090Ti GPU.

4.2 Evaluation under Diverse Realistic Simulations

Baselines. To assess overall performance and validate our architectural advantages, we evaluate baselines across diverse simulated visibility settings using fixed, pre-generated augmentations (Sec. 3.4). Unlike FoV-specific methods [21, 41], our primary baseline, EgoAllo [76], handles arbitrary hand visibility via test-time optimization. We retrain it on the AMASS train set to prevent data leakage. For a fair comparison, we introduce variants retrained with our augmentation: EgoAllo† (diffusion) and EgoPoser† [41] (regressor). Notably, we extend EgoAllo†'s diffusion prior to condition on intermittent hands. For EgoPoser†, we enforce a cold-start by duplicating the first frame to pad its input window. For shape evaluation, we use the mean shape for spatial metrics and raw per-frame values for consistency.

Results. Table 1 demonstrates that our method clearly outperforms the baselines in motion quality, shape accuracy, and consistency. While our feed-forward Vis Hand PE is slightly higher than EgoAllo's optimized version, our overall Hand PE is significantly lower. This indicates that our approach robustly localizes hands even during prolonged occlusions. EgoAllo relies on a head-only prior that,

Table 1: Quantitative results under diverse simulated settings. w/o opt denotes evaluation without optimization; w/ opt denotes evaluation with optimization.

Method	MPJPE↓	MPJVE↓	Jerk↓	Hand PE↓	Vis Hand PE↓	Height↓	Span↓	Height Std↓	Span Std↓	Runtime↓
EgoAllo (w/o opt)	122.99	44.02	0.350	318.69	298.01	3.80	6.24	1.87	2.43	2.49
EgoAllo (w/ opt)	99.32	31.66	0.152	164.80	41.49	3.86	6.19	1.84	2.35	56.09
EgoPoser†	203.25	81.75	1.478	522.86	500.15	6.47	9.41	3.22	4.03	**0.09**
EgoAllo† (w/o opt)	103.55	35.57	0.290	212.95	122.89	3.62	6.10	1.70	2.06	2.55
EgoAllo† (w/ opt)	94.00	26.13	0.120	142.24	**25.74**	3.78	6.25	1.71	2.08	59.26
Ours (w/o opt)	81.35	25.81	**0.117**	146.71	56.98	**2.40**	**4.61**	**0.00**	**0.00**	2.07
Ours (w/ opt)	**80.45**	**25.48**	0.120	**132.74**	29.34	**2.40**	**4.61**	**0.00**	**0.00**	5.97

Fig. 4: Qualitative results under diverse simulated settings. (Left) In a dance sequence, EgoAllo produces unnatural arm poses and unstable motion, while ours remains plausible and smooth. **(Right)** During a kick, EgoAllo's prior hallucinates the invisible hand back into FoV. Ours correctly predicts the hand moving outside the view.

due to its inherent ambiguity, frequently hallucinates invisible hands back into the FoV (Fig. 4). Because its test-time optimization cannot fully correct this error, it causes arms to unnaturally snap to observed positions upon re-entry. By conditioning on hand cues and the visibility-encoding global context $\boldsymbol{\mathcal{G}}$, our prior avoids this issue and accurately infers the position of unseen hands outside the FoV. Furthermore, our lightweight, timestep-adaptive optimization preserves physical plausibility by refining poses on the motion manifold. This prevents the severe artifacts, such as body penetration and implausible joint angles (Fig. 4), caused by EgoAllo's aggressive post-hoc optimization. This naturally satisfies intermittent constraints and achieves superior smoothness without an explicit smoothness loss, operating over 9× faster. Moreover, our model predicts a single, accurate, and consistent body shape (β) by aggregating sequence-level cues via attention-based pooling. In contrast, baselines rely on fluctuating per-frame predictions that destabilize the head-device geometry and cause severe motion ambiguity (*e.g.*, a small person standing vs. a large person sitting under identical inputs; Fig. 7). By ensuring a constant β, our model completely eliminates this

Table 2: Quantitative results across varying camera setups. Metrics are denoted as PE (MPJPE), VE (MPJVE), H-PE (Hand PE), and V-PE (Vis Hand PE).

Method	90° FoV					120° FoV					180° FoV				
	PE↓	VE↓	Jerk↓	H-PE↓	V-PE↓	PE↓	VE↓	Jerk↓	H-PE↓	V-PE↓	PE↓	VE↓	Jerk↓	H-PE↓	V-PE↓
EgoPoser (120)	100.80	41.44	0.519	**164.19**	82.24	95.62	39.60	0.505	**136.04**	64.84	88.95	36.52	0.427	100.93	72.59
EgoAllo†† (120)	110.23	37.95	0.298	212.93	69.68	107.10	36.87	0.293	178.94	71.38	113.14	39.26	0.284	156.06	109.17
EgoAllo (w/ opt)	112.06	34.74	0.162	239.87	**11.06**	104.80	32.72	0.165	192.48	**11.10**	86.70	26.94	0.159	75.56	**10.56**
Ours (120)	101.56	31.49	0.129	207.62	68.96	91.11	28.94	0.126	152.14	59.05	92.07	28.06	0.115	130.77	75.89
Ours ($p = 0.2$)	151.73	45.72	0.158	278.73	75.39	131.97	41.41	0.154	246.58	60.44	94.78	30.75	0.139	137.02	42.60
Ours	93.65	28.67	**0.119**	199.35	59.94	85.70	26.80	**0.117**	158.33	52.21	72.22	23.54	**0.114**	88.50	46.57
Ours (w/ opt)	**93.21**	**28.51**	0.120	193.40	28.19	**85.32**	**26.61**	0.119	149.65	25.26	**71.67**	**23.09**	0.117	**70.96**	23.19

Method	90° FoV + 30° Downward Tilt					120° hFoV × 60° vFoV					Average				
	PE↓	VE↓	Jerk↓	H-PE↓	V-PE↓	PE↓	VE↓	Jerk↓	H-PE↓	V-PE↓	PE↓	VE↓	Jerk↓	H-PE↓	V-PE↓
EgoPoser (120)	98.02	40.54	0.526	150.97	96.47	104.63	42.59	0.525	**183.79**	94.00	97.60	40.14	0.500	147.18	83.33
EgoAllo†† (120)	115.86	40.02	0.287	183.02	106.77	113.94	39.06	0.299	237.91	77.34	112.05	38.63	0.292	193.77	86.87
EgoAllo (w/ opt)	91.36	28.15	0.156	108.28	**10.64**	115.13	35.48	0.161	259.06	**11.18**	102.01	31.61	0.161	175.05	**10.91**
Ours (120)	95.64	29.15	0.117	160.50	80.55	111.58	33.49	0.131	266.18	78.85	98.39	30.23	0.124	183.44	72.66
Ours ($p = 0.2$)	103.51	32.70	0.142	172.89	40.13	161.32	47.28	0.159	295.27	88.26	128.66	39.57	0.150	226.10	61.36
Ours	75.70	24.28	**0.115**	110.64	43.26	98.00	29.60	**0.119**	223.59	69.26	85.05	26.58	**0.117**	156.08	54.25
Ours (w/ opt)	**75.36**	**23.95**	0.117	**96.23**	22.05	**97.48**	**29.46**	0.120	217.47	31.47	**84.61**	**26.32**	0.119	**145.54**	26.03

inconsistency. Finally, our robust performance stems from our unique structural design rather than the augmented training data alone. Retraining the baselines with our augmentations confirms their inherent limitations. The regression-based EgoPoser† collapses to the mean pose, merely adjusting global orientation and causing high jitter. While the diffusion-based EgoAllo† learns a better prior, its feed-forward predictions remain inaccurate, and its optimized performance still trails ours (see Sec. 4.5 for a detailed architectural analysis).

4.3 Evaluation across Varying Camera Setups

Baselines. To evaluate robustness across varying camera setups, we compare our model against EgoPoser [41] (regressor), EgoAllo†† (extended hand-conditioned diffusion), and EgoAllo [76] with test-time optimization. To isolate geometry effects, all models are trained and tested without temporal drops or pose perturbations. During training, baselines assume EgoPoser's fixed 120° pinhole FoV, whereas our model relies solely on the proposed geometry-aware augmentation for hand visibility. For evaluation, we synthesize hand visibility masks using EgoPoser's formulation across five setups: 120°, 90°, and 180° FoV, 120° hFoV × 60° vFoV and 90° FoV with a 30° downward pitch.

Results. Table 2 shows that our camera-agnostic framework consistently outperforms 120°-trained specialists (EgoPoser, EgoAllo††) across all setups, including their native 120° domain. This demonstrates robust generalization under geometry shifts, achieved by synergizing a diffusion prior with geometry-aware augmentation. Unlike regression models that inherently collapse to a learned mean pose when hands become invisible, our diffusion prior ensures natural, continuous arm trajectories. While EgoPoser successfully tracks visible joints as observations expand, its competitively low Hand PE is misleading. Instead, it collapse to the mean pose when hands are invisible, which conservatively

Fig. 5: Qualitative results across varying camera setups. (Left) Sudden arm movements in EgoPoser and arms pushed out of bounds in EgoAllo†† at 90° FoV. **(Right)** Misaligned tracking of EgoAllo†† at 180° FoV.

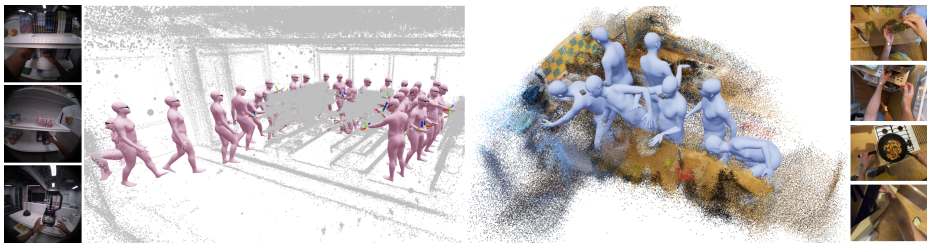

Fig. 6: Qualitative results on in-the-wild data. Reconstructions from (Left) Aria RGB (monocular) and (Right) GoPro captures.

bounds absolute error but destroys kinematic plausibility, causing abrupt motion discontinuities when observations drop out (Fig. 5). Furthermore, exposing our network to a continuous spectrum of simulated visibilities teaches it a robust relationship between motion and camera-induced boundaries. This prevents the severe geometry overfitting exhibited by the diffusion-based EgoAllo††. Paradoxically, EgoAllo††'s MPJPE worsens in 180° and tilted setups despite increased visibility. As Fig. 5 illustrates, it unnaturally pushes invisible hands to its learned 120° boundary in 90° FoVs and fails to exploit valid observations beyond this limit in 180° FoVs, causing misaligned tracking. Consequently, our framework generalizes remarkably well to unseen, strictly bounded pinhole FoVs ($p \to \infty$). By effectively incorporating available visual cues, the model dynamically adapts to arbitrary camera optics without requiring any setup-specific retraining.

4.4 Qualitative Results in the Wild

To demonstrate real-world generalization, we evaluate OmniEgoCap on EgoExo4D [30], Reading in the wild [74], EPIC-KITCHENS [24], and custom captures. As shown in Fig. 1, 7, and 6, our unified model reconstructs coherent full-body motion across diverse camera configurations without hardware-specific

Table 3: Quantitative results on ablation studies. Comparison of variants on diverse simulated hand observation. All variants evaluated without optimization.

Method	MPJPE↓	MPJVE↓	Jerk↓	Hand PE↓	Vis Hand PE↓
No Shape/Global	97.06	**25.17**	**0.104**	161.25	79.70
No Shape Cond	86.55	26.59	0.111	156.09	65.05
No Global	83.53	26.82	0.116	152.73	65.99
Concat	81.60	26.13	0.112	149.92	61.42
No Auxiliary	83.44	27.56	0.163	153.17	60.57
No Root Ori	87.68	26.95	0.120	154.81	66.17
Head	82.13	25.43	0.119	148.71	58.17
Default	**81.35**	25.81	0.117	**146.71**	**56.98**

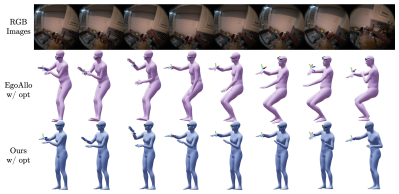

Fig. 7: Qualitative results on EgoExo4D. EgoAllo's shape inconsistency worsens under real-world noise, causing fluctuations during static cooking, while ours remains stable.

retraining and efficiently processes untrimmed videos exceeding 6 minutes on a single RTX 3090Ti GPU. The combination of our perturbation-aware prior and manifold-preserving refinement protects the framework against monocular tracking noise [83], preventing unnatural poses while ensuring consistent body shape and global stability. This establishes OmniEgoCap as a device-agnostic solution generalizing across diverse devices and configurations in the wild.

4.5 Ablation Studies

We conduct ablation studies to validate our key design choices, comparing our full model against variants with specific components removed.

Necessity of Geometry-Aware Augmentation. We validate our geometry-aware augmentation against two baselines: **Ours (Random)** using $p = 0.2$ random masking [14, 41], and **Ours (120)** trained on a fixed 120° FoV. The **Ours (Random)** variant performs worst. By ignoring spatial boundaries, it fails to learn motion-FoV correlations, making it highly vulnerable to realistic, contiguous occlusions. Alternatively, the **Ours (120)** variant acts as a specialist. It learns a strong native prior but suffers from severe geometry overfitting; like EgoAllo†† (Sec. 4.3), its performance degrades under out-of-distribution camera setups. In contrast, our continuous visibility spectrum teaches robust camera-to-motion relationships, yielding a true device-agnostic generalist.

Importance of Sequence-Level Context. To validate our sequence-level context $\boldsymbol{\mathcal{G}}$ and β, we define four baselines: **No Shape/Global** removes all sequence-level contexts, using mean shape; **No Shape Cond** omits β as a decoder condition; **No Global** predicts β but removes $\boldsymbol{\mathcal{G}}$; and **Concat** injects $\boldsymbol{\mathcal{G}}$ via concatenation. First, removing all context (**No Shape/Global**) degrades MPJPE to 97.06mm, while reintroducing $\mathcal{E}_{\text{shape}}$ (**No Global**) improves it to 83.53mm, proving an identity-preserving shape β is crucial. Furthermore, adding $\boldsymbol{\mathcal{G}}$ via AdaLN (**Default**) further improves hand accuracy, reducing Hand PE and Vis Hand PE, confirming $\boldsymbol{\mathcal{G}}$ is essential for generating FoV-consistent hand trajectories. By implicitly encoding FoV boundaries through the spatial distribution of visibility transitions, the model resolves hand location ambiguities local cues cannot address. Without this context, the model fails to infer out-of-FoV

positions, unnaturally hallucinating invisible hands into the camera's view (examples in the supplementary material). Validating our injection strategy, the degradation in **No Shape Cond** (81.35mm to 86.55mm) implies the denoiser must be conditioned on body shape β to predict correct joint orientations for varying bone lengths. Additionally, **Concat** fails to match our AdaLN's hand accuracy, suggesting global modulation better incorporates FoV information.

Auxiliary Loss. We validate the auxiliary loss $\mathcal{L}_{\text{aux}}$. Although many diffusion-based models [19, 47, 53, 76] rely solely on $\mathcal{L}_{\text{simple}}$, which operates purely on joint orientations, this is insufficient. Small orientation errors accumulate along the kinematic chain, causing large position errors at end-effectors like hands. We introduce $\mathcal{L}_{\text{aux}}$ to penalize 3D position errors. Removing $\mathcal{L}_{\text{aux}}$ (**No Auxiliary**) degrades MPJPE and end-effector accuracy (Vis Hand PE 56.98mm to 60.57mm). This confirms $\mathcal{L}_{\text{aux}}$ is critical for satisfying the hand condition $\mathbf{H}$. It also acts as a 3D geometric regularizer. Its removal causes a 39% Jerk spike. Thus, $\mathcal{L}_{\text{aux}}$ is indispensable for hand-aware models to generate accurate and smooth motion.

Root Representation. We validate root orientation parameterization. The **No Root Ori** variant, mimicking EgoAllo [76] by directly stitching predicted body to $\mathbf{T}^{\text{head}}$, performs worst. This approach removes 3DoF, eliminating flexibility by making the root's orientation entirely dependent on the head. Alternatively, the **Head** variant predicts root orientation in head coordinate system ($\mathbf{R}_t^{\text{root}\to\text{head}}$). While restoring flexibility, it forces the model to learn a difficult 3D mapping relative to the volatile head coordinate. In contrast, our gravity-aligned canonical frame provides a stable 2D mapping that is easier to learn. This strikes the best balance between flexibility and stability, yielding top performance.

5 Discussion

We have presented **OmniEgoCap**, a unified sequence-level diffusion framework for robust 3D full-body motion reconstruction across diverse egocentric setups. By leveraging long-range temporal reasoning and geometry-aware visibility augmentation, our model treats intermittent hand observations as geometric constraints, enabling device-agnostic reconstruction that generalizes to unseen optics and mounting configurations without hardware-specific retraining. Finally, predicting a consistent body shape (β) and sequence-level context ($\boldsymbol{\mathcal{G}}$) provides physical anchors that resolve head-trajectory ambiguity and capture visibility boundaries beyond local windows, yielding stable and kinematically plausible motion.

While our framework is robust in practice, it depends on the reliability of upstream SLAM and hand-tracking modules. Moreover, our sequence-level architecture emphasizes global context for stable motion reconstruction, but it is currently optimized for offline processing rather than low-latency, real-time inference. Our experiments also assume a flat ground plane and do not explicitly model detailed finger articulation or hand-object interactions. Improving robustness to tracking failures and relaxing these environmental assumptions, such as by incorporating non-planar scene geometry and richer models for hand-object interaction, are promising directions for future research.

Acknowledgements

This work was supported by KT (Korea Telecom). H. Joo is the corresponding author.

References

1. Apple vision pro. https://www.apple.com/kr/apple-vision-pro/
2. Gopro. https://gopro.com
3. Meta quest vr headsets. https://www.meta.com/quest/
4. Project aria. https://www.projectaria.com/
5. Ray-ban meta smart glasses. https://www.meta.com/smart-glasses
6. Rokoko smartsuit pro. https://www.rokoko.com/products/smartsuit-pro
7. Samsung galaxy xr. https://www.samsung.com/us/xr/galaxy-xr/galaxy-xr/
8. Spectacles by snap inc. https://www.spectacles.com/
9. Vuzix smart glasses. https://www.vuzix.com/pages/smart-glasses
10. Xsens mvn link. https://www.movella.com/products/motion-capture/xsens-mvn-link
11. Akada, H., Wang, J., Golyanik, V., Theobalt, C.: 3d human pose perception from egocentric stereo videos. In: CVPR (2024)
12. Akada, H., Wang, J., Golyanik, V., Theobalt, C.: Bring your rear cameras for egocentric 3d human pose estimation. In: ICCV (2025)
13. Akada, H., Wang, J., Shimada, S., Takahashi, M., Theobalt, C., Golyanik, V.: Unrealego: A new dataset for robust egocentric 3d human motion capture. In: ECCV (2022)
14. Aliakbarian, S., Cameron, P., Bogo, F., Fitzgibbon, A., Cashman, T.J.: Flag: Flow-based 3d avatar generation from sparse observations. In: CVPR (2022)
15. Aliakbarian, S., Saleh, F., Collier, D., Cameron, P., Cosker, D.: Hmd-nemo: Online 3d avatar motion generation from sparse observations. In: ICCV (2023)
16. Barquero, G., Bertsch, N., Marramreddy, M., Chacón, C., Arcadu, F., Rigual, F., He, N.S., Palmero, C., Escalera, S., Ye, Y., et al.: From sparse signal to smooth motion: Real-time motion generation with rolling prediction models. In: CVPR (2025)
17. Barquero, G., Escalera, S., Palmero, C.: Seamless human motion composition with blended positional encodings. In: CVPR (2024)
18. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150 (2020)
19. Castillo, A., Escobar, M., Jeanneret, G., Pumarola, A., Arbeláez, P., Thabet, A., Sanakoyeu, A.: Bodiffusion: Diffusing sparse observations for full-body human motion synthesis. In: ICCV (2023)
20. Chi, H.g., Ha, M.H., Chi, S., Lee, S.W., Huang, Q., Ramani, K.: Infogcn: Representation learning for human skeleton-based action recognition. In: CVPR (2022)
21. Chi, S., Huang, P.H., Sachdeva, E., Ma, H., Ramani, K., Lee, K.: Estimating ego-body pose from doubly sparse egocentric video data. In: NeurIPS (2024)
22. Dabral, R., Mughal, M.H., Golyanik, V., Theobalt, C.: Mofusion: A framework for denoising-diffusion-based motion synthesis. In: CVPR (2023)
23. Dai, P., Zhang, Y., Liu, T., Fan, Z., Du, T., Su, Z., Zheng, X., Li, Z.: Hmd-poser: On-device real-time human motion tracking from scalable sparse observations. In: CVPR (2024)

24. Damen, D., Doughty, H., Farinella, G.M., Furnari, A., Ma, J., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. IJCV (2022)
25. Dittadi, A., Dziadzio, S., Cosker, D., Lundell, B., Cashman, T.J., Shotton, J.: Full-body motion from a single head-mounted device: Generating smpl poses from partial observations. In: ICCV (2021)
26. Dong, J., Feng, B., Guessous, D., Liang, Y., He, H.: Flex attention: A programming model for generating optimized attention kernels. arXiv preprint arXiv:2412.05496 (2024)
27. Du, Y., Kips, R., Pumarola, A., Starke, S., Thabet, A., Sanakoyeu, A.: Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In: CVPR (2023)
28. Engel, J., Somasundaram, K., Goesele, M., Sun, A., Gamino, A., Turner, A., Talattof, A., Yuan, A., Souti, B., Meredith, B., Peng, C., Sweeney, C., Wilson, C., Barnes, D., DeTone, D., Caruso, D., Valleroy, D., Ginjupalli, D., Frost, D., Miller, E., Mueggler, E., Oleinik, E., Zhang, F., Somasundaram, G., Solaira, G., Lanaras, H., Howard-Jenkins, H., Tang, H., Kim, H.J., Rivera, J., Luo, J., Dong, J., Straub, J., Bailey, K., Eckenhoff, K., Ma, L., Pesqueira, L., Schwesinger, M., Monge, M., Yang, N., Charron, N., Raina, N., Parkhi, O., Borschowa, P., Moulon, P., Gupta, P., Mur-Artal, R., Pennington, R., Kulkarni, S., Miglani, S., Gondi, S., Solanki, S., Diener, S., Cheng, S., Green, S., Saarinen, S., Patra, S., Mourikis, T., Whelan, T., Singh, T., Balntas, V., Baiyya, V., Dreewes, W., Pan, X., Lou, Y., Zhao, Y., Mansour, Y., Zou, Y., Lv, Z., Wang, Z., Yan, M., Ren, C., Nardi, R.D., Newcombe, R.: Project aria: A new tool for egocentric multi-modal ai research. arXiv preprint arXiv:2308.13561 (2023)
29. Feng, H., Ma, W., Gao, Q., Zheng, X., Xue, N., Xu, H.: Stratified avatar generation from sparse observations. In: CVPR (2024)
30. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., Byrne, E., Chavis, Z., Chen, J., Cheng, F., Chu, F.J., Crane, S., Dasgupta, A., Dong, J., Escobar, M., Forigua, C., Gebreselasie, A., Haresh, S., Huang, J., Islam, M.M., Jain, S., Khirodkar, R., Kukreja, D., Liang, K.J., Liu, J.W., Majumder, S., Mao, Y., Martin, M., Mavroudi, E., Nagarajan, T., Ragusa, F., Ramakrishnan, S.K., Seminara, L., Somayazulu, A., Song, Y., Su, S., Xue, Z., Zhang, E., Zhang, J., Castillo, A., Chen, C., Fu, X., Furuta, R., Gonzalez, C., Gupta, P., Hu, J., Huang, Y., Huang, Y., Khoo, W., Kumar, A., Kuo, R., Lakhavani, S., Liu, M., Luo, M., Luo, Z., Meredith, B., Miller, A., Oguntola, O., Pan, X., Peng, P., Pramanick, S., Ramazanova, M., Ryan, F., Shan, W., Somasundaram, K., Song, C., Southerland, A., Tateno, M., Wang, H., Wang, Y., Yagi, T., Yan, M., Yang, X., Yu, Z., Zha, S.C., Zhao, C., Zhao, Z., Zhu, Z., Zhuo, J., Arbelaez, P., Bertasius, G., Crandall, D., Damen, D., Engel, J., Farinella, G.M., Furnari, A., Ghanem, B., Hoffman, J., Jawahar, C.V., Newcombe, R., Park, H.S., Rehg, J.M., Sato, Y., Savva, M., Shi, J., Shou, M.Z., Wray, M.: Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In: CVPR (2024)
31. Guo, C., Mu, Y., Javed, M.G., Wang, S., Cheng, L.: Momask: Generative masked modeling of 3d human motions. In: CVPR (2024)
32. Guzov, V., Jiang, Y., Hong, F., Pons-Moll, G., Newcombe, R., Liu, C.K., Ye, Y., Ma, L.: Hmd^2: Environment-aware motion generation from single egocentric head-mounted device. In: 3DV (2025)

33. Guzov, V., Mir, A., Sattler, T., Pons-Moll, G.: Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In: CVPR (2021)
34. Harvey, F.G., Yurick, M., Nowrouzezahrai, D., Pal, C.: Robust motion in-betweening. ACM Transactions on Graphics (TOG) (2020)
35. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
36. Huang, Y., Kaufmann, M., Aksan, E., Black, M.J., Hilliges, O., Pons-Moll, G.: Deep inertial poser learning to reconstruct human pose from sparse inertial measurements in real time. ACM TOG (2018)
37. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b. arXiv preprint arXiV:2310.06825 (2023)
38. Jiang, C.M., Cornman, A., Park, C., Sapp, B., Zhou, Y., Anguelov, D.: Motiondiffuser: Controllable multi-agent motion prediction using diffusion. In: CVPR (2023)
39. Jiang, H., Grauman, K.: Seeing invisible poses: Estimating 3d body pose from egocentric video. In: CVPR (2017)
40. Jiang, H., Ithapu, V.K.: Egocentric pose estimation from human vision span. In: ICCV (2021)
41. Jiang, J., Streli, P., Meier, M., Holz, C.: Egoposer: Robust real-time egocentric pose estimation from sparse and intermittent observations everywhere. In: ECCV (2024)
42. Jiang, J., Streli, P., Qiu, H., Fender, A., Laich, L., Snape, P., Holz, C.: Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In: ECCV (2022)
43. Jiang, Y., Ye, Y., Gopinath, D., Won, J., Winkler, A.W., Liu, C.K.: Transformer inertial poser: Real-time human motion reconstruction from sparse imus with simultaneous terrain generation. In: SIGGRAPH Asia (2022)
44. Lee, J., Xu, W., Richard, A., Wei, S.E., Saito, S., Bai, S., Wang, T.L., Sung, M., Kim, T.K., Saragih, J.: Rewind: Real-time egocentric whole-body motion diffusion with exemplar-based identity conditioning. In: CVPR (2025)
45. Lee, J., Joo, H.: Mocap everyone everywhere: Lightweight motion capture with smartwatches and a head-mounted camera. In: CVPR (2024)
46. Lee, J., Lee, Y., Kim, J., Kosiorek, A.R., Choi, S., Teh, Y.W.: Set transformer: A framework for attention-based permutation-invariant neural networks. In: ICML (2019)
47. Li, J., Liu, K., Wu, J.: Ego-body pose estimation via ego-head pose estimation. In: CVPR (2023)
48. Li, J., Cao, J., Zhang, H., Rempe, D., Kautz, J., Iqbal, U., Yuan, Y.: Genmo: A generalist model for human motion. In: ICCV (2025)
49. Luo, Z., Hachiuma, R., Yuan, Y., Kitani, K.: Dynamics-regulated kinematic policy for egocentric pose estimation. In: NeurIPS (2021)
50. Ma, L., Ye, Y., Hong, F., Guzov, V., Jiang, Y., Postyeni, R., Pesqueira, L., Gamino, A., Baiyya, V., Kim, H.J., Bailey, K., Fosas, D.S., Liu, C.K., Liu, Z., Engel, J., Nardi, R.D., Newcombe, R.: Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In: ECCV (2024)
51. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: ICCV (2019)
52. Merel, J., Tunyasuvunakool, S., Ahuja, A., Tassa, Y., Hasenclever, L., Pham, V., Erez, T., Wayne, G., Heess, N.: Catch & carry: reusable neural controllers for vision-guided whole-body tasks. ACM TOG (2020)

53. Patel, C., Nakamura, H., Kyuragi, Y., Kozuka, K., Niebles, J.C., Adeli, E.: Uniegomotion: A unified model for egocentric motion reconstruction, forecasting, and generation. In: ICCV (2025)
54. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV (2023)
55. Pineda, L., Fan, T., Monge, M., Venkataraman, S., Sodhi, P., Chen, R.T., Ortiz, J., DeTone, D., Wang, A., Anderson, S., Dong, J., Amos, B., Mukadam, M.: Theseus: A library for differentiable nonlinear optimization. In: NeurIPS (2022)
56. Rempe, D., Birdal, T., Hertzmann, A., Yang, J., Sridhar, S., Guibas, L.J.: Humor: 3d human motion model for robust pose estimation. In: ICCV (2021)
57. Rhodin, H., Richardt, C., Casas, D., Insafutdinov, E., Shafiei, M., Seidel, H.P., Schiele, B., Theobalt, C.: Egocap: egocentric marker-less motion capture with two fisheye cameras. ACM TOG (2016)
58. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. arXiv preprint arXiv:2201.02610 (2022)
59. Sinha, A., Choi, C., Ramani, K.: Deephand: Robust hand pose estimation by completing a matrix imputed with deep features. In: CVPR (2016)
60. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
61. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing (2024)
62. Tang, J., Wang, J., Ji, K., Xu, L., Yu, J., Shi, Y.: A unified diffusion framework for scene-aware human motion estimation from sparse signals. In: CVPR (2024)
63. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., Bermano, A.H.: Human motion diffusion model. In: ICLR (2023)
64. Tome, D., Alldieck, T., Peluse, P., Pons-Moll, G., Agapito, L., Badino, H., De la Torre, F.: Selfpose: 3d egocentric pose estimation from a headset mounted camera. IEEE TPAMI (2020)
65. Tran, M., Mao, H., Chen, Q., Kim, Y.: Head2body: Body pose generation from multi-sensory head-mounted inputs. In: ICCV (2025)
66. Van Wouwe, T., Lee, S., Falisse, A., Delp, S., Liu, C.K.: Diffusionposer: Real-time human motion reconstruction from arbitrary sparse sensors using autoregressive diffusion. In: CVPR (2024)
67. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
68. Von Marcard, T., Rosenhahn, B., Black, M.J., Pons-Moll, G.: Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In: Comput. Graph. Forum (2017)
69. Wang, J., Cao, Z., Luvizon, D., Liu, L., Sarkar, K., Tang, D., Beeler, T., Theobalt, C.: Egocentric whole-body motion capture with fisheyevit and diffusion-based motion refinement. In: CVPR (2024)
70. Wang, J., Dabral, R., Luvizon, D., Cao, Z., Liu, L., Beeler, T., Theobalt, C.: Ego4o: Egocentric human motion capture and understanding from multi-modal input. In: CVPR (2025)
71. Wang, J., Liu, L., Xu, W., Sarkar, K., Theobalt, C.: Estimating egocentric 3d human pose in global space. In: ICCV (2021)
72. Wang, J., Luvizon, D., Xu, W., Liu, L., Sarkar, K., Theobalt, C.: Scene-aware egocentric 3d human pose estimation. In: CVPR (2023)
73. Winkler, A., Won, J., Ye, Y.: Questsim: Human motion tracking from sparse sensors with simulated avatars. In: SIGGRAPH Asia (2022)
74. Yang, C., Alam, S., Siam, S.I., Proulx, M., Mathias, L., Somasundaram, K., Pesqueira, L., Fort, J., Sheriffdeen, S., Parkhi, O., Ren, C., Zhang, M., Chai, Y., Newcombe, R., Kim, H.J.: Reading recognition in the wild. In: NeurIPS (2025)

75. Yang, C., Tkach, A., Hampali, S., Zhang, L., Crowley, E.J., Keskin, C.: Egoposeformer: A simple baseline for stereo egocentric 3d human pose estimation. In: ECCV (2024)
76. Yi, B., Ye, V., Zheng, M., Li, Y., Müller, L., Pavlakos, G., Ma, Y., Malik, J., Kanazawa, A.: Estimating body and hand motion in an ego-sensed world. In: CVPR (2025)
77. Yi, X., Zhou, Y., Habermann, M., Golyanik, V., Pan, S., Theobalt, C., Xu, F.: Egolocate: Real-time motion capture, localization, and mapping with sparse body-mounted sensors. ACM TOG (2023)
78. Yi, X., Zhou, Y., Habermann, M., Shimada, S., Golyanik, V., Theobalt, C., Xu, F.: Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In: CVPR (2022)
79. Yi, X., Zhou, Y., Xu, F.: Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. ACM TOG (2021)
80. Yi, X., Zhou, Y., Xu, F.: Physical non-inertial poser (pnp): Modeling non-inertial effects in sparse-inertial human motion capture. In: SIGGRAPH (2024)
81. Yuan, Y., Kitani, K.: 3d ego-pose estimation via imitation learning. In: ECCV (2018)
82. Yuan, Y., Kitani, K.: Ego-pose estimation and forecasting as real-time pd control. In: ICCV (2019)
83. Zhang, J., Deng, J., Ma, C., Potamias, R.A.: Hawor: World-space hand motion reconstruction from egocentric videos. In: CVPR (2025)
84. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. IEEE TPAMI (2024)
85. Zhang, Y., Xia, S., Chu, L., Yang, J., Wu, Q., Pei, L.: Dynamic inertial poser (dynaip): Part-based motion dynamics learning for enhanced human pose estimation with sparse inertial sensors. In: CVPR (2024)
86. Zheng, X., Su, Z., Wen, C., Xue, Z., Jin, X.: Realistic full-body tracking from sparse observations via joint-level modeling. In: ICCV (2023)

Supplementary Material for OmniEgoCap: Camera-Agnostic Sequence-Level Egocentric Motion Reconstruction

Kyungwon Cho, Hanbyul Joo

Seoul National University

A Implementation Details

In this section, we provide additional details to supplement the descriptions in the main text. To ensure full reproducibility, we will release our code and pre-trained checkpoints upon acceptance.

Network Architecture. Tab. 1 summarizes the capacity of each sub-network. While the core structural design is introduced in the main text, we detail the specific module configurations here. Prior to the transformer blocks, raw condition variables are projected into their respective hidden dimensions via Multi-Layer Perceptrons (MLPs). For the shape and global branches, the attention poolers [6] consist of a single cross-attention layer with 16 learnable queries, followed by a self-attention layer and MLPs. Throughout the network, all transformer blocks adopt pre-Layer Normalization and GELU activations. In the decoder, the global conditioning vector for the AdaLN-Zero [7] modules is constructed by directly adding the Fourier-embedded diffusion timestep n and the MLP-projected shape parameter $\hat{\beta}$.

Training and Inference. We apply an Exponential Moving Average (EMA) decay of 0.9999 to stabilize training. For the diffusion process, our model is parameterized to directly predict the clean canonicalized pose $\mathbf{X}^0$ ($\mathbf{X}_0$-prediction). We utilize a cosine noise schedule with 1,000 DDPM [4] training timesteps, while inference is accelerated using 30 DDIM [8] steps. For the test-time guidance optimization introduced in the main text, we specifically optimize only the 3D local rotations of the arm joints (*e.g.*, shoulder, elbow, wrist). To ensure computational efficiency, the Levenberg-Marquardt solver is constrained to a maximum of 5 iterations per diffusion step and is executed in a fully vectorized manner.

B Augmentation Details

We present the detailed hyperparameters for the augmentation strategy introduced in Sec. 3.4.

Geometry-Aware Augmentation. To simulate diverse capture setups, ranging from monocular and stereo to fisheye systems with varying orientations (forward

Table 1: Architectural Hyperparameters. Detailed configurations of the transformer modules used in OmniEgoCap.

Module	Hidden Dim	Out Dim	Layers	Heads	Dropout
Shape Encoder	384	384	2	6	0.1
Shape Pooler	384	16	1	6	0.1
Global Encoder	384	384	2	6	0.1
Global Pooler	384	512	1	6	0.1
Local Encoder	512	512	6	8	0.1
Decoder	512	132	8	8	0.1

Table 2: Comparison with DSPoser. Tested on circular 90° FoV.

Method	MPJPE↓	MPJVE↓
DSPoser	55.1	24.19
Ours	**52.15**	**17.58**

to downward), we sample spatial parameters uniformly. The sampling ranges are defined as follows (angles in radians):

$$\begin{aligned} \gamma_x \in [0.35, 2.15],\ \gamma_y \in [0.35, 1.35] \\ \phi_x \in [-0.15, 0.15],\ \phi_y \in [0.0, 1.5],\ p \in [2.0, 10.0] \end{aligned} \tag{1}$$

To preserve realistic aspect ratios and tilt orientations, we enforce the constraints:

$$0.4 \leq \frac{\gamma_y}{\gamma_x} \leq 1.1, \quad \phi_y \leq 0.4 + (\gamma_y - 0.35) \cdot 1.1. \tag{2}$$

Temporal Drops and Pose Perturbation. To simulate realistic tracking failures like motion blur or prolonged occlusions, we employ a two-stage stochastic masking strategy. We dynamically sample the number of drop events $K \sim \text{Poisson}(T \cdot \rho / \mathbb{E}[D])$ for a sequence of length T to match a target drop ratio ρ. The duration D of each drop is drawn from a heavy-tailed Log-Normal distribution with mean $m = \mathbb{E}[D]$ and standard deviation s. We define two independent masking modes:

$$\begin{aligned} \text{Short: } \rho \sim \mathcal{U}(0.0, 0.1),\ m = 2.0,\ s = 1.0 \\ \text{Long: } \rho \sim \mathcal{U}(0.0, 0.2),\ m = 28.0,\ s = 25.0 \end{aligned} \tag{3}$$

The final visibility mask is the union of these two modes, clipping the long-drop durations to a minimum of 5 frames.

Finally, to mimic the inherent jitter of off-the-shelf pose estimators, we inject Gaussian noise into the visible wrist poses. Specifically, we apply additive noise sampled from $\mathcal{N}(0, 0.005^2)$ for the 3D translations (in meters) and $\mathcal{N}(0, (\pi/180)^2)$ for the rotation angles (in radians) along each axis.

C Comparison with DSPoser

While DSPoser [3] is closely related to our work, its official implementation is not publicly available, precluding a direct evaluation under the diverse camera setups investigated in our main text. To ensure a fair and comprehensive comparison, we provide additional results by training and evaluating our model strictly

following DSPoser's official protocol and comparing it against their reported metrics. Specifically, we evaluate our framework under their fixed 90° circular field-of-view (FoV) setting using the AvatarPoser [5] train/test split. Furthermore, since DSPoser does not model varying body proportions, we disable our shape prediction module ($\mathcal{E}_{\text{shape}}$) and enforce a fixed mean body shape during both training and inference to ensure an exact experimental match.

As shown in Tab. 2, our camera-agnostic framework outperforms the 90°-trained specialist (DSPoser) even within its native 90° domain. This is consistent with our findings in Sec. 4.3, where our model also outperformed FoV-specific specialist baselines even in their native settings. By exposing our network to a continuous spectrum of simulated visibilities, the model effectively mitigates FoV-specific bias and reduces the geometry overfitting typically exhibited by specialist models trained on fixed configurations.

D Qualitative Analysis on Ablation Studies

As discussed in Sec. 4.5 of the main text, Fig. 1 provides further qualitative analysis to highlight the critical impact of the global context ($\boldsymbol{\mathcal{G}}$) in resolving out-of-FoV ambiguities. In the initial frames of the sequence, our full model correctly reconstructs the hands outside the camera's view, whereas the **No Global** baseline incorrectly places them in front of the body. As the motion progresses and the subject swings their arms forward, the wrists enter the frontal field-of-view (FoV) and become clearly observed. By aggregating sequence-level cues via $\boldsymbol{\mathcal{G}}$, our full model leverages this later visibility evidence to correctly infer that the initially unobserved hands must have been outside the FoV. In contrast, the **No Global** model, constrained by a local attention horizon, lacks this broader temporal context and hallucinates the invisible hands within the camera's view. These results further support our claim that the global context implicitly encodes FoV boundaries, serving as an important constraint for accurately disambiguating egocentric motions.

E Representation Visualization

We provide visual illustrations of the coordinate representations detailed in Sec. 3.2 of the main text. Following EgoAllo [10], the canonical frame serves as a gravity-aligned local reference. It is defined by projecting the head coordinate onto the ground plane, aligning its z-axis with gravity and its y-axis with the head's forward direction. Based on this frame, Fig. 2 visualizes the spatio-temporally invariant relative transformations used to construct our model's features. These include the relative head trajectory ($\Delta\mathbf{T}_t^{\text{head}}$), wrist positions relative to the head ($\mathbf{T}_t^{\text{hand}\to\text{head}}$), canonicalized head orientation ($\mathbf{R}_t^{\text{head}\to\text{cano}}$), head height ($h_t$), relative canonical rotation ($\Delta\mathbf{R}_t^{\text{cano}}$), and relative root orientation ($\mathbf{R}_t^{\text{root}\to\text{cano}}$).

Fig. 1: Qualitative Analysis of Ablation Studies. In a jumping and running sequence, the **No Global** model incorrectly places invisible hands within the visible zone due to the lack of global context. In contrast, our full model correctly infers that the invisible hands are outside the FoV.

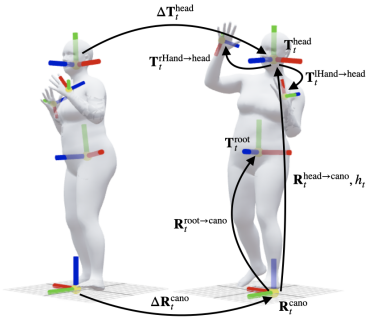

Fig. 2: Representation Visualization. Illustration of the gravity-aligned canonical frame and the relative geometric transformations used to construct our model's features.

F Calibration and World Alignment

We present the detailed calibration and alignment procedures used to map raw tracking data into the gravity-aligned world coordinate system, expanding upon Sec. 3.2 of the main text. Importantly, these steps are lightweight one-time preprocessing procedures and do not require any device-specific retraining.

Camera-to-Head Calibration. To map the camera trajectory to the head joint, we must determine the rigid transformation $\mathbf{T}^{\text{cam}\to\text{head}}$. For Aria [2] and Quest [1], which track the center eye frame, we use a precomputed transformation based on a mean body shape. For monocular cameras, estimating the absolute metric translation between the camera and the head joint is ill-posed. Therefore, we decouple the calibration of translation and orientation. We approximate the translation offset based on the device mounting configuration (*e.g.*, using an external reference image). For the orientation, we use a brief pre-calibration phase where the user stands straight and looks forward. Since this posture approximately aligns with our predefined canonical head frame, we can estimate the relative rotation from the camera to the head joint from this initialization.

Ground Plane Fitting and Scale Correction. To establish the world coordinate system, we determine the ground plane by extracting the global 3D point cloud—generated either by the device's native SLAM system (*e.g.*, Aria) or the DROID-SLAM [9] backend within a modified HaWoR pipeline [11]. We apply RANSAC to this point cloud to fit the ground plane, using its normal to define the gravity-aligned z-axis.

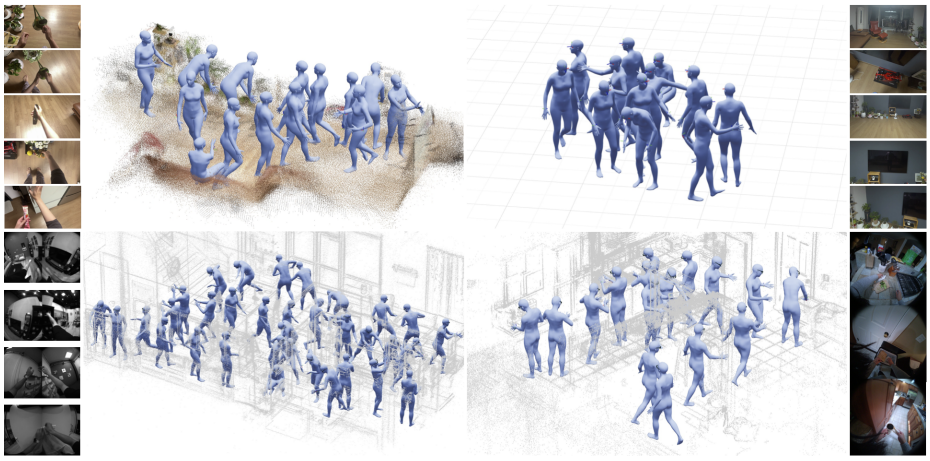

Fig. 3: Additional qualitative results on in-the-wild data. Reconstructions from (Upper Left) iPhone 15, (Upper Right) Quest, (Lower Left) Aria Stereo (SLAM), and (Lower Right) Aria Monocular (RGB).

For monocular in-the-wild captures, we further refine the global scale for qualitative visualization. Since the initial metric scale from HaWoR is often inaccurate, we estimate a single global scale correction factor during the static standing phase of the pre-calibration by comparing the estimated camera height against an approximate physical camera height based on the user's height and device placement. This correction is used only as a preprocessing step to resolve the metric ambiguity of monocular SLAM for qualitative in-the-wild examples.

G Additional Qualitative Results in the Wild

In Fig. 3, we provide additional in-the-wild reconstruction results captured across diverse setups, including an iPhone 15 (26mm lens), Meta Quest 3 [1], and Project Aria [2]. Notably, for the Aria RGB sequences, we determine hand visibility by projecting the device's 3D hand tracking outputs onto the 2D image plane. For the Quest captures, the visualized front RGB image is shown only for reference; hand tracking is estimated by the headset's onboard camera system rather than being restricted to the displayed RGB view. Complementing the main text, these diverse examples further highlight OmniEgoCap's robust and device-agnostic generalization capabilities across unconstrained camera setups. We encourage readers to view the supplementary video for animated visualizations.

H Failure Cases

We further investigate representative failure cases in Fig. 4, which are shared by head-centric egocentric reconstruction methods, including both EgoAllo [10] and

Fig. 4: Representative failure cases. Gray corresponds to the ground truth, blue to Ours (w/ opt), and purple to EgoAllo (w/ opt).

our framework. First, seated poses remain inherently ambiguous in an egocentric view, as similar head and hand trajectories may correspond to squatting, sitting on the floor, or sitting on a chair. Second, near-lying poses may introduce minor foot-floating artifacts, as the head-centric canonicalization becomes less stable in such extreme configurations. Additionally, our reconstruction quality may also degrade when the upstream SLAM or hand tracking signals become unreliable.

References

1. Meta quest vr headsets. `https://www.meta.com/quest/`
2. Project aria. `https://www.projectaria.com/`
3. Chi, S., Huang, P.H., Sachdeva, E., Ma, H., Ramani, K., Lee, K.: Estimating ego-body pose from doubly sparse egocentric video data. In: NeurIPS (2024)
4. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
5. Jiang, J., Streli, P., Qiu, H., Fender, A., Laich, L., Snape, P., Holz, C.: Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In: ECCV (2022)
6. Lee, J., Lee, Y., Kim, J., Kosiorek, A.R., Choi, S., Teh, Y.W.: Set transformer: A framework for attention-based permutation-invariant neural networks. In: ICML (2019)
7. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV (2023)
8. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
9. Teed, Z., Deng, J.: Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. In: NeurIPS (2021)
10. Yi, B., Ye, V., Zheng, M., Li, Y., Müller, L., Pavlakos, G., Ma, Y., Malik, J., Kanazawa, A.: Estimating body and hand motion in an ego-sensed world. In: CVPR (2025)
11. Zhang, J., Deng, J., Ma, C., Potamias, R.A.: Hawor: World-space hand motion reconstruction from egocentric videos. In: CVPR (2025)