OmniEgoCap: Device-Agnostic Sequence-Level Egocentric Motion Reconstruction

Kyungwon Cho **Jeonghyeon Na** **Hanbyul Joo**

Seoul National University

{cscandkswon,prom317,hbjoo}@snu.ac.kr

https://kyungwoncho.github.io/OmniEgoCap/

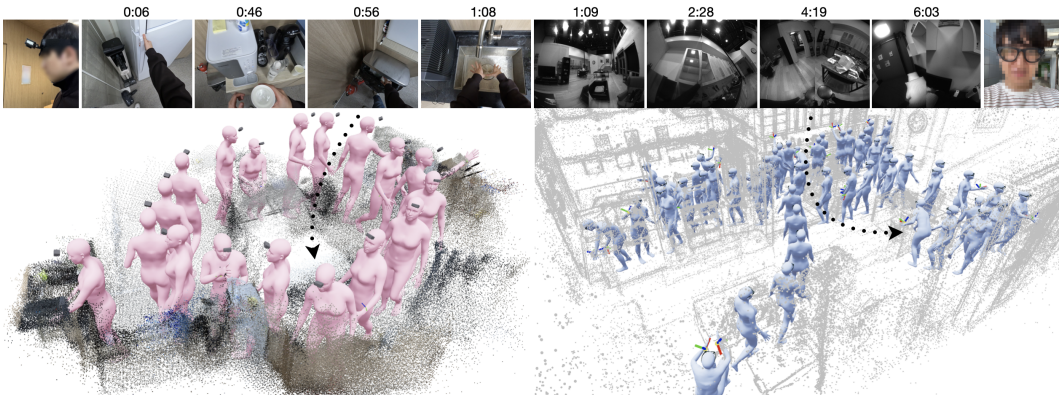

Figure 1: **OmniEgoCap** is a sequence-level diffusion framework that combines long-range body and camera-observation invariants with geometry-aware visibility augmentation to reconstruct consistent 3D full-body motion across diverse egocentric devices. (L) GoPro and (R) Aria.

Abstract

Commercial egocentric devices capture human behavior in everyday settings, yet full-body motion reconstruction must generalize across diverse cameras and mountings. This is challenging because head trajectories under-determine body motion, while hand observations are intermittent and camera-dependent: a hand may disappear simply because it has left the camera's field of view. Existing methods treat such absences as missing data, relying on fixed visibility regimes or post-hoc optimization, and fail to generalize across devices. We present OmniEgoCap, a sequence-level diffusion framework for device-agnostic egocentric full-body motion reconstruction. Our key insight is that egocentric videos contain sequence-level evidence about both the wearer and the camera-induced visibility structure, enabling OmniEgoCap to infer consistent body shape and coherent motion from sparse head and hand cues. To prevent overfitting to a single camera setup, we further introduce geometry-aware visibility augmentation, which synthesizes visibility patterns from realistic variations in camera geometry. We also present OmniEgoDB, the first mocap benchmark that pairs ground-truth motion with recordings from multiple consumer egocentric devices. Experiments on synthetic, real-device, and in-the-wild settings demonstrate superior motion reconstruction, robust cross-device generalization, and coherent in-the-wild results.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Egocentric vision systems are rapidly diversifying, from action cameras [2] and VR headsets [3, 1, 7] to smart glasses [5, 4, 8, 9]. For machines to better assist humans, these systems need to understand 3D human motion and behavior. Yet reconstructing the wearer's 3D body from egocentric observations is challenging, since the first-person view rarely captures the full body. Existing methods therefore rely on sparse cues, primarily head trajectories from SLAM or SfM [45, 70] and visible hands from XR device tracking [4, 3, 1, 7] or off-the-shelf estimators [75]. As these devices become commonplace, full-body motion capture should be democratized across the devices people actually wear, ideally without device-specific retraining. This goal is complicated by the fact that hand visibility is itself camera-dependent: hands may disappear due to occlusion, tracking failure, or simply moving outside the camera's visible region. For device-agnostic reconstruction, hand observations must therefore be understood in relation to the camera geometry that produces them.

Existing methods do not fully exploit the camera geometry encoded in hand observations. A missing hand is not merely absent evidence; across time, visibility and invisibility reveal what the camera can and cannot see. Optimization-based methods impose costly post-hoc constraints on visible hands [42, 70], but do not model sustained invisibility as part of this observation process. Recent end-to-end methods use head and hand signals directly [21, 38, 51, 31], but are typically trained under a fixed camera setup. They can therefore learn setup-specific visibility patterns and degrade when the setup changes. These failures point to a common missing ingredient, sequence-level context. An egocentric sequence contains long-range evidence about both *who the wearer is* and *how the camera observes them*. Sparse transitions between visible and invisible hand states reveal the camera's observation boundary over time. Body shape is also a sequence-level property. It differs across wearers but is fixed within a sequence, while reliable evidence for height and proportions appears unevenly over time, especially in informative poses such as standing. It is thus insufficient to rely on a fixed mean body shape [45, 42, 31, 39, 20] or to predict shape per frame [70, 51]. Thus, device-agnostic reconstruction requires inferring both the wearer's body shape and the camera's visibility context from the entire sequence.

Building on this view, we present **OmniEgoCap**, a sequence-level diffusion framework for device-agnostic egocentric full-body motion reconstruction. It uses long-range cues to estimate body shape and a visibility context that captures camera-dependent observation structure, guiding motion reconstruction from sparse head and hand signals. Our encoder separates sequence context from local motion evidence over long sequences, while a diffusion decoder reconstructs full-body motion conditioned on the inferred context and intermittent hand observations. To make the model robust to diverse camera setups, we introduce *geometry-aware visibility augmentation*. Rather than randomly dropping hand observations, we synthesize diverse yet plausible visibility patterns induced by realistic variations in egocentric camera geometry and mounting, encouraging the model to treat both hand visibility and invisibility as geometric observation cues. Evaluating this capability, however, is hindered by the fact that existing egocentric motion datasets are captured almost exclusively with a single device, most commonly Project Aria glasses [4], whose specialized hardware is not readily available to general consumers. We therefore introduce **OmniEgoDB**, the first public mocap benchmark that pairs ground-truth motion with recordings from multiple egocentric devices, including widely accessible consumer cameras and XR headsets. Experiments on public benchmarks, OmniEgoDB, and in-the-wild videos demonstrate robust generalization across device categories, mountings, and visibility patterns.

In summary, our contributions are threefold. (1) We propose a sequence-level diffusion framework that reconstructs long-range full-body motion from sparse egocentric cues while jointly inferring the wearer's body shape and a visibility context capturing camera-dependent observation structure. (2) We introduce geometry-aware visibility augmentation, which models hand visibility as a camera-induced observation process, prevents overfitting to a fixed visibility regime, and enables robust cross-device generalization. (3) We contribute a public mocap evaluation benchmark collected with multiple egocentric devices and mounting configurations, and demonstrate device-agnostic reconstruction without device-specific retraining.

2 Related Work

Sparse Observations for Egocentric Motion Capture Wearable IMU systems [10, 6, 64, 34, 73, 72, 40, 74, 77, 63] and VR/AR setups with 6-DoF head and hand tracking [39, 78, 23, 38, 15,

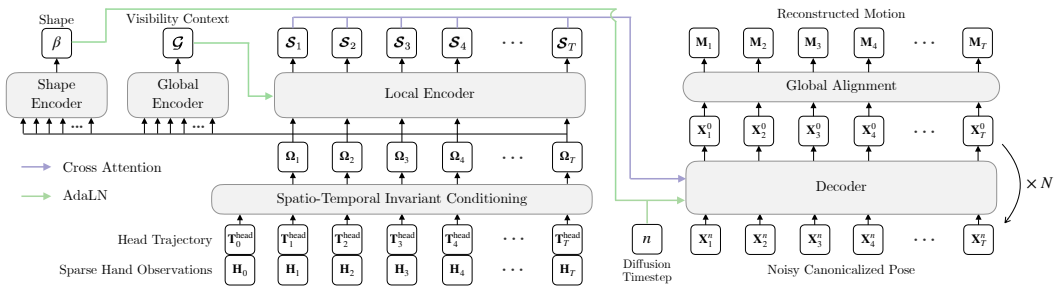

Figure 2: **Overview of our model.** OmniEgoCap takes the head trajectory $\mathbf{T}^{\text{head}}_{0:T}$ and intermittent hand observations $\mathbf{H}_{0:T}$, which are first converted into spatio-temporally invariant conditioning features $\mathbf{\Omega}_{1:T}$. The encoder $\mathcal{E}$ processes $\mathbf{\Omega}_{1:T}$ to predict a single body shape $\hat{\beta}$, a visibility context $\boldsymbol{\mathcal{G}}$, 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 $\mathbf{M}_{1:T}$.

16, 25, 27, 20, 59, 29] provide sparse yet informative body cues but rely on dedicated sensors, controllers, or global localization to reduce root drift [32, 71, 42]. Egocentric cameras further relax hardware requirements, but body visibility is severely limited. Some methods improve visibility using specialized downward-facing or stereo fisheye cameras [55, 66, 37, 62, 67, 13, 11, 69, 65, 12, 41], while more practical front-facing methods infer motion from head trajectories [45], visual features [51, 31], or intermittent hand observations [21, 70]. However, existing approaches usually assume a fixed sensing or visibility regime [21, 38], or use hand observations mainly through test-time optimization [70, 42]. This limits their ability to model how hand visibility, including sustained disappearance, changes with camera geometry. In contrast, we target *device-agnostic* reconstruction from diverse head-mounted videos, where visible hands and sustained invisibility provide *sequence-level evidence* of camera-induced visibility structure.

Generative Priors and Sequence-Level Modeling Diffusion models provide strong generative priors for high-fidelity motion synthesis [61, 36, 22, 76], making them well suited to egocentric reconstruction where sparse head and hand cues leave multiple full-body motions plausible. However, sparse local cues alone do not provide enough context to infer body shape or camera-dependent visibility structure. Prior methods often assume a fixed mean shape [45, 42, 31, 39, 20] or predict shape locally [70, 51], leading to inconsistent body proportions and ambiguous motion interpretation. As with body shape, camera-dependent visibility structure must be inferred from long-range evidence such as when and where hands appear and disappear. Modeling such long-range evidence with transformer-based diffusion is challenging because standard attention scales quadratically with sequence length. We adopt sliding-window attention [46, 17] for efficient long-sequence denoising, while aggregating sequence-level context for the wearer's body shape and the camera's observation boundary.

3 Method

3.1 Problem Formulation

Given an egocentric video 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 head poses $\mathbf{T}^{\text{head}}_{0:T} = \{\mathbf{T}^{\text{head}}_t \in \text{SE}(3)\}_{t=0}^{T}$ in world coordinates and intermittent wrist observations $\mathbf{H}_{0:T} = \{\mathbf{H}^{\text{lHand}}_t, \mathbf{H}^{\text{rHand}}_t\}_{t=0}^{T}$. Each $\mathbf{H}^{\text{lHand/rHand}}_t = (\mathbf{T}^{\text{lHand/rHand}}_t, v^{\text{lHand/rHand}}_t)$ consists of a 6-DoF wrist pose $\mathbf{T}^{\text{lHand/rHand}}_t \in \text{SE}(3)$ in head coordinates and a visibility state $v^{\text{lHand/rHand}}_t \in \{0, 1\}$. A wrist may be invisible due to occlusion, motion blur, tracking failure, or moving outside the camera field of view. Our model OmniEgoCap takes $\mathbf{T}^{\text{head}}_{0:T}$ and $\mathbf{H}_{0:T}$ as input and reconstructs human motion $\mathbf{M}_{1:T}$.

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

The motion follows the SMPL-H format [56], $\mathbf{M}_{1:T} = \{\beta, \boldsymbol{r}_{1:T}, \mathbf{\Phi}_{1:T}, \mathbf{\Theta}_{1:T}\}$, where $\beta \in \mathbb{R}^{16}$ is a sequence-level shape, $\boldsymbol{r}_{1:T}$ is the root translation, $\mathbf{\Phi}_{1:T}$ is the root orientation, and $\mathbf{\Theta}_{1:T}$ denotes joint

rotations. The root poses are defined in the world coordinate system, consistent with $\mathbf{T}^{\text{head}}$. Unlike prior approaches that use a fixed mean shape [42, 39, 45] or per-frame shape [70, 51], we predict a single consistent shape β for the whole sequence. Fig. 2 provides an overview of the framework.

3.2 Preprocessing and Representation

Preprocessing Given egocentric images $\mathbf{I}$, we first estimate camera poses $\mathbf{T}_t^{\text{cam}} \in \text{SE}(3)$[1] using off-the-shelf SLAM. We define a gravity-aligned world coordinate system with the floor at $z = 0$. Since the camera and head joints are not co-located, we compute head poses as $\mathbf{T}_t^{\text{head}} = \mathbf{T}_t^{\text{cam}}(\mathbf{T}^{\text{cam}\to\text{head}})^{-1}$, where $\mathbf{T}^{\text{cam}\to\text{head}}$ is a mount-specific but time-invariant rigid transform estimated through lightweight pre-calibration. We also extract 6-DoF wrist observations and visibility $\mathbf{H}_t^{\text{hand}}$ using off-the-shelf pose estimation modules. For monocular RGB videos, we use HaWoR [75], without its infilling module, to extract both $\mathbf{T}_t^{\text{cam}}$ and $\mathbf{H}_t^{\text{hand}}$. For Aria [28], we use Aria Machine Perception Services (MPS).

Coordinate and Conditioning Representation Learning motion from long sparse sequences requires spatio-temporally invariant conditioning. We therefore 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. For the head trajectory, following EgoAllo [70], we define a per-frame canonical frame $\mathbf{T}_t^{\text{cano}\to\text{world}}$ at the head's floor projection, with its z-axis aligned with gravity and y-axis aligned with the head's forward direction. The condition vector $\mathbf{\Omega}_t \in \mathbb{R}^D$ is computed by a neural function Γ:

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

Here, $\Delta\mathbf{T}_t^{\text{head}}$ is the relative head pose from frame $t-1$, $\mathbf{R}_t^{\text{head}\to\text{cano}}$ is the canonicalized head orientation, and h_t is the head height from $\mathbf{T}_t^{\text{head}}$. Unlike EgoAllo [70], we include the canonical relative rotation $\Delta\mathbf{R}_t^{\text{cano}}$ from frame $t-1$, because our model predicts the root orientation relative to the canonical frame. We introduce hand conditions $\mathbf{T}_t^{\text{hand}\to\text{head}} = (\mathbf{T}_t^{\text{lHand}\to\text{head}}, \mathbf{T}_t^{\text{rHand}\to\text{head}})$, which encode relative hand poses in head coordinates. When a hand is invisible, its cue is replaced with a learnable null embedding. We do not include temporal hand motion cues, as wrist observations are often intermittent.

Global Alignment Following prior work [38, 39, 70], we use $\mathbf{T}^{\text{head}}$ for global localization and directly derive the root translation $\boldsymbol{r}_t$. Unlike EgoAllo [70], however, we predict the relative root orientation $\mathbf{R}_t^{\text{root}\to\text{cano}}$ in the canonical coordinate. The final global root orientation is computed as $\mathbf{\Phi}_t = \mathbf{R}_t^{\text{cano}\to\text{world}}\mathbf{R}_t^{\text{root}\to\text{cano}}$. This representation provides an additional 3DoF in global orientation, improving motion stability and accuracy.

3.3 Sequence-Level Diffusion Architecture

We adopt a conditional diffusion model [33] conditioned on the processed conditioning features $\mathbf{\Omega}_{1:T} = \{\mathbf{\Omega}_t\}_{t=1}^T$. The architecture separates *sequence-level context* from local motion evidence. A global pathway infers the wearer's body shape and camera-dependent visibility structure from long-range cues, while a local pathway summarizes per-frame head and hand evidence. The two pathways are fused so that the same local evidence can be interpreted differently depending on the inferred body and observation context. To train and infer on long sequences end-to-end without splitting them into independent subsequences, both the encoder $\mathcal{E}$ and decoder $\mathcal{D}$ use sliding-window local attention with horizon W [18, 35, 17, 46] and Rotary Positional Embedding [58], reducing attention complexity from quadratic to linear.

Sequence-Level Encoder The encoder $\mathcal{E}$ maps the sparse conditioning sequence $\mathbf{\Omega}_{1:T}$ into a sequence-level shape $\hat{\beta}$, a visibility context $\boldsymbol{\mathcal{G}}$, and per-frame summaries $\boldsymbol{\mathcal{S}}_{1:T}$.

$$\hat{\beta} = \mathcal{E}_{\text{shape}}(\mathbf{\Omega}_{1:T}), \quad \boldsymbol{\mathcal{G}} = \mathcal{E}_{\text{global}}(\mathbf{\Omega}_{1:T}), \quad \boldsymbol{\mathcal{S}}_{1:T} = \mathcal{E}_{\text{local}}(\mathbf{\Omega}_{1:T}, \boldsymbol{\mathcal{G}}). \tag{3}$$

$\mathcal{E}_{\text{shape}}$ infers the wearer's body proportions by pooling long-range cues such as head-height variations and the spatial extent of visible hand positions. $\mathcal{E}_{\text{global}}$ infers visibility context from visible hand

[1] $\mathbf{T}_t^{\text{cam}}$ denotes the transformation from the camera to world coordinates, equivalently $\mathbf{T}_t^{\text{cam}\to\text{world}}$. Similarly, $\mathbf{T}_t^{\text{head}} = \mathbf{T}_t^{\text{head}\to\text{world}}$.

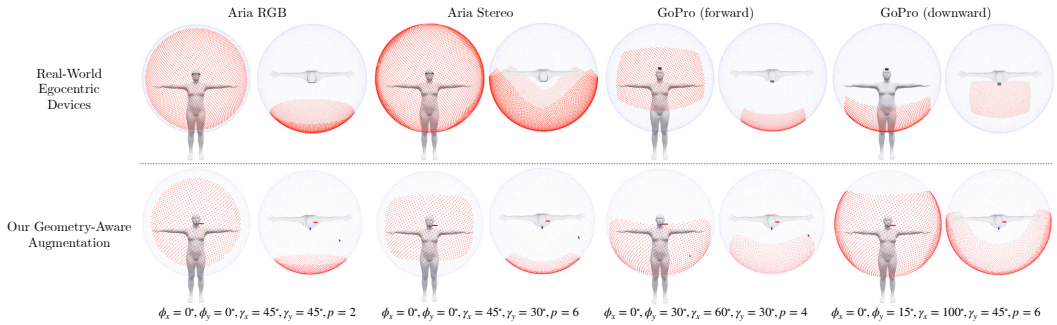

Figure 3: **Real and simulated visibility boundaries.** The top row shows calibrated FoVs from real-world devices (*e.g.*, Aria, GoPro). The bottom row illustrates visibility patterns sampled by our geometry-aware augmentation. Red regions indicate the visible field of view.

poses and the patterns of when and where hands disappear. Both branches first aggregate local temporal features with sliding-window attention and then pool information across the full sequence using learnable cross-attention queries [43]. The local branch $\mathcal{E}_{\text{local}}$ produces per-frame summaries from neighboring conditioning features, modulated by $\boldsymbol{\mathcal{G}}$ through AdaLN-Zero [52]. This lets the same local head and hand evidence induce different poses depending on the inferred body shape and visibility context.

Decoder The decoder $\mathcal{D}$ denoises the canonicalized pose representation $\mathbf{X}^0_{1:T} = \{\mathbf{R}^{\text{root}\to\text{cano}}_{1:T}, \boldsymbol{\Theta}_{1:T}\}$ rather than the final global motion directly. The root translation $\boldsymbol{r}$ and orientation $\boldsymbol{\Phi}$ are then recovered as described in the Global Alignment section. The decoder is implemented as a DiT [52] conditioned on the encoded summaries $\boldsymbol{\mathcal{S}}_{1:T}$ and body shape $\hat{\beta}$.

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

Here, $n \in \{1, \ldots, N\}$ is the diffusion timestep and $\hat{\mathbf{X}}^0_{1:T}$ is the predicted $\mathbf{X}^0_{1:T}$. The noised input $\mathbf{X}^n_{1:T}$ is defined by the DDPM [33] forward process $q(\mathbf{X}^n_{1:T}|\mathbf{X}^0_{1:T}) = \mathcal{N}(\mathbf{X}^n_{1:T}; \sqrt{\bar{\alpha}_n}\mathbf{X}^0_{1:T}, (1 - \bar{\alpha}_n)\mathbb{I})$, 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

Geometry-Aware Visibility Augmentation For $\mathcal{E}_{\text{global}}$ to learn visibility context rather than memorize a single visibility regime, training must expose the model to diverse camera-induced visibility patterns. Motion-annotated egocentric datasets, such as Nymeria [47], obtain accurate motion using specialized capture hardware, but their hand visibility patterns are tied to specific camera optics and mountings. Training directly on these patterns encourages a device-specific prior. We instead synthesize a continuous family of camera-dependent visibility boundaries.

We model boundary variations caused by field of view, camera tilt, aspect ratio, lens masks, and optical distortions. We rotate head coordinates by the center offsets (ϕ_x, ϕ_y) into a camera frame, compute the wrist yaw ψ_x and pitch ψ_y in this frame, and define the boundary with half-angles (γ_x, γ_y) and a power p controlling boundary shape. A wrist is visible when its angles satisfy

$$|\psi_x \,/\, \gamma_x|^p + |\psi_y \,/\, \gamma_y|^p \leq 1. \tag{5}$$

Sampling these parameters from realistic ranges exposes the model to diverse camera setups, as shown in Fig. 3. Unlike random masking [14], this augmentation makes both visibility and invisibility depend on camera geometry, allowing $\mathcal{E}_{\text{global}}$ to learn visibility patterns as sequence-level context. We further handle imperfect in-the-wild tracking by applying stochastic temporal signal drops and Gaussian perturbations to visible wrist poses. We sample signal drops from Poisson and Log-Normal distributions to model the heavy-tailed nature of occlusions.

Loss Functions Our training objective combines the DDPM simple loss, a shape consistency loss, and auxiliary physical losses.

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

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{n,\mathbf{X}^0,\boldsymbol{\Omega}}\left[\|\mathbf{X}^0 - \hat{\mathbf{X}}^0\|^2\right], \quad \mathcal{L}_{\text{shape}} = \|\text{FK}(\mathbf{0},\beta) - \text{FK}(\mathbf{0},\hat{\beta})\|^2. \tag{7}$$

Here, $\mathcal{L}_{\text{simple}}$ is the standard DDPM [33] denoising objective. To enforce shape consistency, $\mathcal{L}_{\text{shape}}$ minimizes the 3D T-pose joint error instead of directly penalizing the PCA-derived β, which lacks physical meaning. FK denotes the forward kinematics function.

Finally, $\mathcal{L}_{\text{aux}}$ regularizes the reconstructed motion $\hat{\mathbf{M}}_{1:T}$ with joint position $\mathcal{L}_{\text{pos}}$ and foot-skating $\mathcal{L}_{\text{skat}}$ losses [61]:

$$\mathcal{L}_{\text{pos}}=\frac{1}{T}\sum_{t=1}^{T}\|\text{FK}(\mathbf{M}_t,\beta)-\text{FK}(\hat{\mathbf{M}}_t,\hat{\beta})\|^2,\ \mathcal{L}_{\text{skat}}=\frac{1}{T-1}\sum_{t=1}^{T-1}\|\text{FK}(\hat{\mathbf{M}}_{t+1},\hat{\beta})-\text{FK}(\hat{\mathbf{M}}_t,\hat{\beta})\|^2\cdot c_t \tag{8}$$

$$\mathcal{L}_{\text{aux}} = \bar{\alpha}_n\left(\lambda_{\text{pos}}\mathcal{L}_{\text{pos}} + \lambda_{\text{skat}}\mathcal{L}_{\text{skat}}\right). \tag{9}$$

Here, c_t is the ground-truth binary contact label, and scaling by $\bar{\alpha}_n$ emphasizes physical accuracy when the signal level is high. We find that $\mathcal{L}_{\text{aux}}$ is crucial for satisfying the hand condition $\mathbf{H}$ and generating stable and accurate motion. We set $\lambda_{\text{shape}} = 2.0$, $\lambda_{\text{pos}} = 0.25$, and $\lambda_{\text{skat}} = 0.4$.

3.5 Inference

We use DDIM [57] sampling for motion generation. Because the denoiser is already conditioned on hand observations and sequence-level context, feed-forward prediction alone produces accurate and plausible motion without post-hoc fitting. Optionally, we introduce timestep-adaptive guidance optimization that exploits the diffusion sampling process to improve visible wrist alignment while preserving motion plausibility. We provide further details in the supplementary material.

4 Experiments

4.1 Experimental Setting

Dataset We train on the AMASS [48] train split with the proposed stochastic augmentation (Sec. 3.4). All sequences are resampled to 30 FPS and preprocessed following HuMoR [54], including floor alignment and foot-contact annotation. We evaluate on AMASS validation/test splits for simulation, our OmniEgoDB for real-world quantification, and in-the-wild videos for qualitative analysis.

Training Details We train with AdamW using a learning rate of 10^{-4}, weight decay 10^{-4}, and batch size 32 for 16 hours on 8 A5000 GPUs. The maximum training sequence length is 512 frames, with attention horizon $W = \pm 63$ implemented using FlexAttention [26]. Test-time optimization uses the Theseus [53] Levenberg–Marquardt optimizer. Additional details are provided in the supplementary material.

Metrics We evaluate motion quality using 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. For hand reconstruction, we report Hand Position Error (**Hand PE**, mm) and Visible Hand Position Error (**Vis Hand PE**, mm), computed on visible hand frames. For body shape, we report **Height Error** and **Span Error** (cm) for accuracy, and **Height Std** and **Span Std** (cm) for temporal consistency, where height and span are the maximum vertical and horizontal vertex distances in a T-pose. **Runtime** (sec) measures average execution time per sequence on an RTX 3090Ti, including optimization setup. For the OmniEgoDB benchmark, SMPL-H predictions are evaluated against Xsens ground truth on 15 corresponding major joints, excluding ambiguous spine, collar, and toe joints.

4.2 Overall Performance under Diverse Simulated Visibility

Setup We evaluate on simulated hand-visibility patterns generated by our augmentation and fixed across all methods. We compare against EgoAllo [70], retrained on AMASS for a fair comparison under the same training distribution, with and without test-time optimization. We also retrain EgoAllo† and EgoPoser† [38] with our augmentation to isolate architectural differences, where EgoAllo† denotes our hand-conditioned EgoAllo extension. For shape metrics, baselines use mean shape for spatial errors and raw per-frame estimates for consistency.

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†	117.04	58.24	0.972	244.39	117.45	4.65	7.52	2.97	3.52	**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



Figure 4: **Qualitative results under diverse simulated settings. (Left)** EgoAllo produces unnatural, unstable motion during dancing, while ours remains plausible. **(Right)** During a kick, EgoAllo hallucinates the invisible hand back into the FoV, whereas ours predicts it outside the view.

Results Table 1 shows that our model achieves the best overall motion, hand, and shape metrics while predicting a single consistent body shape across the sequence. This supports our sequence-level formulation: body shape helps disambiguate full-body motion from head trajectories, while visibility context helps interpret intermittent hand observations. Although EgoAllo† with optimization attains the lowest visible-hand error, its higher overall Hand PE reveals failures during invisible periods. As shown in Fig. 4, EgoAllo often hallucinates unobserved hands inside the FoV and then snaps them to the observations when they reappear. Its aggressive post-hoc optimization can also introduce artifacts such as body penetration or implausible joint angles when enforcing sparse hand constraints. In contrast, by conditioning on intermittent hands and sequence-level visibility context, ours keeps out-of-FoV hand trajectories plausible and avoids such discontinuities. Moreover, baselines with per-frame shape estimates can change body proportions even under the same motion state, amplifying ambiguities such as a shorter wearer standing versus a taller wearer sitting, whereas our sequence-level shape remains fixed. Our lightweight guidance further improves hand alignment while preserving smooth motion and running over $9\times$ faster than EgoAllo. Finally, our improvement over EgoPoser† and EgoAllo† shows that augmented training alone is insufficient without our sequence-level architecture.

4.3 Generalization across Camera Setups

Setup We evaluate robustness to camera-geometry shifts using five synthetic visibility setups: $90°$, $120°$, and $180°$ FoV, $120°$ FoV with $30°$ downward tilt, and $120°$ hFoV $\times$ $60°$ vFoV. To isolate geometry effects, all models are trained and tested without temporal drops or pose perturbations. Baselines marked (120) are trained under EgoPoser's fixed $120°$ pinhole visibility regime, whereas OmniEgoCap uses only geometry-aware visibility augmentation. EgoAllo†† is our hand-conditioned

Table 2: **Quantitative results across varying camera setups.** PE and H-PE denote MPJPE and Hand PE.

Method	90° FoV			120° FoV			180° FoV			120° + 30° Tilt			120° × 60°		
	PE↓	Jerk↓	H-PE↓	PE↓	Jerk↓	H-PE↓	PE↓	Jerk↓	H-PE↓	PE↓	Jerk↓	H-PE↓	PE↓	Jerk↓	H-PE↓
EgoPoser (120)	100.80	0.519	**164.19**	95.62	0.505	**136.04**	88.95	0.427	100.93	98.02	0.526	150.97	104.63	0.525	**183.79**
EgoAllo†† (120)	110.23	0.298	212.93	107.10	0.293	178.94	113.14	0.284	156.06	115.86	0.287	183.02	113.94	0.299	237.91
EgoAllo (w/ opt)	112.06	0.162	239.87	104.80	0.165	192.48	86.70	0.159	75.56	91.36	0.156	108.28	115.13	0.161	259.06
Ours (120)	101.56	0.129	207.62	91.11	0.126	152.14	92.07	0.115	130.77	95.64	0.117	160.50	111.58	0.131	266.18
Ours (Random)	151.73	0.158	278.73	131.97	0.154	246.58	94.78	0.139	137.02	103.51	0.142	172.89	161.32	0.159	295.27
Ours	93.65	**0.119**	199.35	85.70	**0.117**	158.33	72.22	**0.114**	88.50	75.70	**0.115**	110.64	98.00	**0.119**	223.59
Ours (w/ opt)	**93.21**	0.120	193.40	**85.32**	0.119	149.65	**71.67**	0.117	**70.96**	**75.36**	0.117	**96.23**	**97.48**	0.120	217.47

Table 3: **Quantitative results on OmniEgoDB.** PE, VE, and H-PE denote MPJPE, MPJVE, and Hand PE. # denotes methods evaluated with their official checkpoints. – denotes entries not applicable to UniEgoMotion, which needs RGB frames and predicts at 10 FPS.

Method	Aria (Stereo) 21 seqs, 67.1 min				Aria (RGB) 21 seqs, 67.1 min				GoPro 20 seqs, 68.1 min				Quest 9 seqs, 42.0 min			
	PE↓	VE↓	Jerk↓	H-PE↓	PE↓	VE↓	Jerk↓	H-PE↓	PE↓	VE↓	Jerk↓	H-PE↓	PE↓	VE↓	Jerk↓	H-PE↓
EgoAllo# (w/o opt)	138.22	24.70	0.146	261.10	138.22	24.70	0.146	261.10	136.86	24.11	0.143	260.80	152.17	22.02	0.140	282.93
EgoAllo# (w/ opt)	115.48	25.06	0.122	142.00	128.71	32.26	0.259	182.86	125.65	28.65	0.189	186.80	122.87	22.97	0.121	135.15
EgoPoser#	126.54	43.56	0.659	172.75	135.99	51.65	0.906	197.39	134.89	47.30	0.776	209.21	130.04	35.17	0.521	153.07
EgoPoser (120)	124.95	58.15	1.265	185.71	132.22	65.00	1.551	198.19	130.40	65.67	1.580	213.15	125.46	46.07	0.979	161.43
UniEgoMotion#	–	–	–	–	151.61	–	–	243.02	161.10	–	–	258.63	–	–	–	–
Ours	106.61	**21.29**	**0.077**	142.28	**118.11**	**23.25**	**0.083**	173.83	**111.68**	**21.81**	**0.073**	166.20	**110.49**	**18.16**	**0.062**	135.86
Ours (w/ opt)	**106.23**	21.81	0.093	**134.56**	119.12	25.99	0.178	**171.28**	112.21	23.30	0.125	**164.39**	111.50	18.91	0.080	**128.35**

EgoAllo extension trained under this fixed 120° regime. Table 2 also includes visibility-modeling variants of OmniEgoCap, analyzed in Sec. 4.5.

Results Table 2 shows that our model generalizes consistently across camera geometries, achieving the best full-body accuracy and smoothness across all five setups. The fixed-regime baselines reveal two complementary failure modes. EgoPoser (120) can obtain low Hand PE by collapsing invisible hands toward a conservative mean pose, but this conservative strategy leads to higher full-body error, jerk, and discontinuities near FoV transitions. Although EgoAllo†† uses a diffusion prior, it overfits to the fixed 120° visibility boundary and even worsens at 180° despite increased visibility, showing that diffusion priors alone do not ensure camera-geometry generalization. Supplementary Fig. 8 shows this boundary bias, where EgoAllo†† pushes invisible hands toward its learned boundary and fails to adapt to the actual FoV. Ours also outperforms DSPoser [21] in its own 90° setting (Supplementary Table 8). Overall, geometry-aware augmentation and sequence-level visibility context allow ours to adapt to unseen FoVs and mountings without device-specific retraining while preserving plausible full-body motion.

4.4 Real-World Evaluation on OmniEgoDB and Further In-the-Wild Results

Existing egocentric motion datasets are limited for evaluating device-agnostic reconstruction, as they are typically captured with a single device, most commonly Project Aria glasses [4], and rely on specialized hardware whose mountings induce device-specific visibility patterns [47]. We introduce **OmniEgoDB**, a new mocap benchmark that pairs Xsens bodysuit ground-truth motion with recordings from multiple consumer egocentric devices, including Aria, GoPro, and Quest, across 50 diverse sequences from four subjects totaling 177 minutes. This enables direct evaluation of cross-device generalization under real-world conditions. Capture details are provided in the supplementary material.

Table 3 reports quantitative results on OmniEgoDB. Aria (Stereo) uses the device's hand tracking, and Aria (RGB) uses HaWoR [75] on the RGB stream of the same recordings. We run the official EgoPoser [38] at its native 60 FPS on inputs upsampled from 30 FPS and the official UniEgoMotion [51] at its native 10 FPS, and resample their predictions to 30 FPS for evaluation. The official

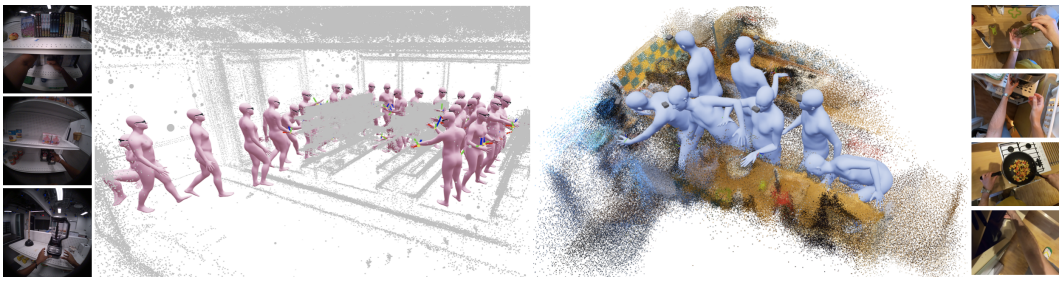

Figure 5: **Qualitative results on in-the-wild data.** Reconstructions from (Left) Aria RGB (monocular) and (Right) GoPro captures.

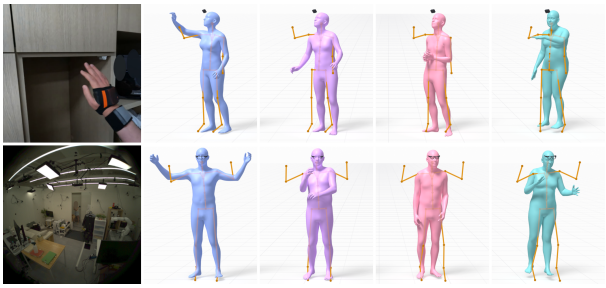

Figure 6: **Qualitative results on OmniEgoDB.** GoPro (top) and Aria (RGB) (bottom). From left, input, OmniEgoCap, EgoAllo# (w/ opt), EgoPoser (120), and UniEgoMotion#, with the Xsens ground truth in orange.



Figure 7: **Qualitative results on EgoExo4D.** From top, input, EgoAllo (w/ opt), and ours (w/ opt). EgoAllo's shape inconsistency worsens under real-world noise, causing fluctuations during static cooking, while ours remains stable throughout.

EgoPoser does not estimate body shape, so we evaluate it with the mean shape. Unlike simulated settings, OmniEgoDB contains real-world device recordings with longer trajectories and larger spatial coverage, while also introducing diverse optics and mountings, imperfect wrist visibility estimates, and noisy 6-DoF wrist pose estimates. Across all devices and both Aria settings, OmniEgoCap achieves the lowest MPJPE, MPJVE, and Jerk, and with guidance the lowest Hand PE, showing that the learned body shape and visibility context transfer to real-world device recordings without device-specific retraining (Fig. 6). No method, including ours, is trained on OmniEgoDB. We also report zero-shot results on Aria Digital Twin [50] in the supplementary material.

Beyond OmniEgoDB, we evaluate OmniEgoCap on public in-the-wild egocentric videos from EgoExo4D [30], Reading in the Wild [68], EPIC-KITCHENS [24], and custom captures. As shown in Fig. 1, Fig. 7, and Fig. 5, OmniEgoCap reconstructs coherent full-body motion across diverse camera configurations without device-specific retraining. Baselines often exhibit jitter, abrupt pose changes, or implausible arm motion in these noisy settings, while ours maintains stable global motion and consistent body shape. This supports real-world deployment from ordinary egocentric videos.

4.5 Ablation Studies

We conduct ablation studies to validate the two components central to our formulation: geometry-aware visibility augmentation and sequence-level context. Additional ablations on the auxiliary loss and canonical root-orientation representation are provided in the supplementary material.

Necessity of Geometry-Aware Augmentation Table 2 evaluates two visibility-modeling variants: **Ours (Random)**, which uses random masking with probability 0.2 [14, 38], and **Ours (120)**, trained with a fixed 120° visibility boundary. Random masking performs poorly because it removes hand observations without preserving camera-induced spatial structure. The fixed-boundary variant behaves like a camera-specific specialist, performing competitively near its native setup but degrading under out-of-distribution FoVs and mountings. In contrast, our geometry-aware augmentation exposes the model to a continuous family of visibility boundaries, enabling it to learn transferable visibility

Table 4: **Quantitative results on ablation studies.** Comparison of variants on diverse simulated hand observations. All variants are 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
Default	**81.35**	25.81	0.117	**146.71**	**56.98**

Table 5: **Linear probing of the visibility context.** Visible-region IoU (top) and stature error in cm (bottom) on held-out AMASS motions.

Probe input (frozen)	Region IoU↑
Shuffled labels	0.447
Mean camera	0.582
Visibility statistics	0.662
$\mathcal{G}$ (ours)	**0.740**
Probe input (frozen)	Stature err.↓
$\mathcal{G}$	9.19
$\hat{\beta}$	**3.05**

context. Notably, geometry-aware augmentation improves full-body accuracy over the fixed-boundary variant even at the nominal 120° setting, while maintaining comparable hand accuracy.

Importance of Sequence-Level Context Table 4 ablates our sequence-level shape and visibility context. **No Shape/Global** removes both branches and uses the mean shape without $\mathcal{G}$. **No Shape Cond** predicts $\hat{\beta}$ but removes decoder shape conditioning. **No Global** removes $\mathcal{G}$, and **Concat** replaces AdaLN with concatenation. Removing all sequence-level context (No Shape/Global) substantially degrades MPJPE, while shape conditioning alone (No Global) recovers most full-body accuracy and adding $\mathcal{G}$ (Default) further improves hand accuracy. Supplementary qualitative results support this interpretation, showing that $\mathcal{G}$ captures camera-induced visibility boundaries and helps maintain FoV-consistent invisible-hand trajectories. These results demonstrate that body shape and visibility context are most effective when used to modulate the denoising process, rather than being injected as local conditioning alone.

Probing the Visibility Context To test whether $\mathcal{G}$ encodes camera geometry, we freeze the encoder and fit a ridge-regression probe from $\mathcal{G}$ to the camera's visible region on a sphere of head-relative directions, using held-out AMASS test motions each paired with 12 cameras sampled from our augmentation. The IoU between predicted and true regions is computed over the directions the wrists visit, with folds grouped by motion (details in the supplementary material). $\mathcal{G}$ reaches 0.740 IoU (Table 5), well above shuffled labels (0.447), the mean training camera (0.582), and coarse per-sequence visibility statistics such as the fraction of frames with visible wrists and the visibility transition rate (0.662), so it encodes more than how often the wrists are seen. In contrast, $\hat{\beta}$ from the same encoder predicts wearer stature (3.05 cm error, versus 9.19 cm from $\mathcal{G}$), indicating that the two branches separate camera geometry from body shape.

5 Discussion

We present **OmniEgoCap**, a sequence-level diffusion framework for 3D full-body motion reconstruction across diverse egocentric camera setups. By inferring consistent body shape and visibility context from long-range cues, OmniEgoCap resolves head-trajectory ambiguity and treats intermittent hand visibility as geometric evidence rather than missing data. Together with geometry-aware visibility augmentation, this enables device-agnostic reconstruction without device-specific retraining. We also introduce **OmniEgoDB**, a multi-device mocap benchmark for quantitative cross-device evaluation with real-world egocentric recordings and ground-truth full-body motion.

Our framework needs no device-specific retraining, but each device still requires calibration and preprocessing that provide a mount-specific camera-to-head transform, a gravity-aligned metric head trajectory, and wrist observations with visibility flags. Devices without native tracking obtain these from SLAM and off-the-shelf hand estimators, so reconstruction quality depends on these upstream components. OmniEgoCap is designed for offline sequence-level reconstruction, not low-latency real-time inference. It also assumes a flat ground plane and does not model finger articulation or hand-object interactions. OmniEgoDB covers only four subjects and a limited range of capture locations and activities. Future work may improve robustness to tracking failures, incorporate non-planar scene geometry, and model finger articulation and richer hand-object interactions for more general egocentric motion understanding.

Acknowledgments and Disclosure of Funding

We thank Gunhee Kim and Lee Chae-Yeon for their help with the benchmark capture. This work was supported by KT(Korea Telecom)-Seoul National University AICT R&D Center. This work was also supported by NRF grant funded by the Korea government (MSIT) (No. RS-2022-NR070498), IITP grant funded by the Korea government (MSIT) [No. RS-2024-00439854, No. 2022-0-00156, and No. RS-2025-25441838], and Institute of Engineering Research at Seoul National University. We thank Meta for providing the Project Aria research glasses used in this work. H. Joo is the corresponding author.

References

[1] Apple vision pro. https://www.apple.com/kr/apple-vision-pro/. 2

[2] Gopro. https://gopro.com. 2, 19

[3] Meta quest vr headsets. https://www.meta.com/quest/. 2, 19, 22

[4] Project aria. https://www.projectaria.com/. 2, 8, 22

[5] Ray-ban meta smart glasses. https://www.meta.com/smart-glasses. 2

[6] Rokoko smartsuit pro. https://www.rokoko.com/products/smartsuit-pro. 2

[7] Samsung galaxy xr. https://www.samsung.com/us/xr/galaxy-xr/galaxy-xr/. 2

[8] Spectacles by snap inc. https://www.spectacles.com/. 2

[9] Vuzix smart glasses. https://www.vuzix.com/pages/smart-glasses. 2

[10] Xsens mvn link. https://www.movella.com/products/motion-capture/xsens-mvn-link. 2

[11] Hiroyasu Akada, Jian Wang, Vladislav Golyanik, and Christian Theobalt. 3d human pose perception from egocentric stereo videos. In *CVPR*, 2024. 3

[12] Hiroyasu Akada, Jian Wang, Vladislav Golyanik, and Christian Theobalt. Bring your rear cameras for egocentric 3d human pose estimation. In *ICCV*, 2025. 3

[13] Hiroyasu Akada, Jian Wang, Soshi Shimada, Masaki Takahashi, Christian Theobalt, and Vladislav Golyanik. Unrealego: A new dataset for robust egocentric 3d human motion capture. In *ECCV*, 2022. 3

[14] Sadegh Aliakbarian, Pashmina Cameron, Federica Bogo, Andrew Fitzgibbon, and Thomas J Cashman. Flag: Flow-based 3d avatar generation from sparse observations. In *CVPR*, 2022. 5, 9

[15] Sadegh Aliakbarian, Fatemeh Saleh, David Collier, Pashmina Cameron, and Darren Cosker. Hmd-nemo: Online 3d avatar motion generation from sparse observations. In *ICCV*, 2023. 2

[16] German Barquero, Nadine Bertsch, Manojkumar Marramreddy, Carlos Chacón, Filippo Arcadu, Ferran Rigual, Nicky Sijia He, Cristina Palmero, Sergio Escalera, Yuting Ye, et al. From sparse signal to smooth motion: Real-time motion generation with rolling prediction models. In *CVPR*, 2025. 3

[17] German Barquero, Sergio Escalera, and Cristina Palmero. Seamless human motion composition with blended positional encodings. In *CVPR*, 2024. 3, 4

[18] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. 4

[19] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, RISHI HAZRA, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollar, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *ICLR*, 2026. 20

[20] Angela Castillo, Maria Escobar, Guillaume Jeanneret, Albert Pumarola, Pablo Arbeláez, Ali Thabet, and Artsiom Sanakoyeu. Bodiffusion: Diffusing sparse observations for full-body human motion synthesis. In *ICCV*, 2023. 2, 3, 17

[21] Seunggeun Chi, Pin-Hao Huang, Enna Sachdeva, Hengbo Ma, Karthik Ramani, and Kwonjoon Lee. Estimating ego-body pose from doubly sparse egocentric video data. In *NeurIPS*, 2024. 2, 3, 8, 18

[22] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *CVPR*, 2023. 3

[23] Peng Dai, Yang Zhang, Tao Liu, Zhen Fan, Tianyuan Du, Zhuo Su, Xiaozheng Zheng, and Zeming Li. Hmd-poser: On-device real-time human motion tracking from scalable sparse observations. In *CVPR*, 2024. 2

[24] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *IJCV*, 2022. 9

[25] Andrea Dittadi, Sebastian Dziadzio, Darren Cosker, Ben Lundell, Thomas J Cashman, and Jamie Shotton. Full-body motion from a single head-mounted device: Generating smpl poses from partial observations. In *ICCV*, 2021. 3

[26] Juechu Dong, Boyuan Feng, Driss Guessous, Yanbo Liang, and Horace He. FlexAttention: A programming model for generating fused attention variants. In *MLSys*, 2025. 6

[27] Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, and Artsiom Sanakoyeu. Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In *CVPR*, 2023. 3

[28] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, Cheng Peng, Chris Sweeney, Cole Wilson, Dan Barnes, Daniel DeTone, David Caruso, Derek Valleroy, Dinesh Ginjupalli, Duncan Frost, Edward Miller, Elias Mueggler, Evgeniy Oleinik, Fan Zhang, Guruprasad Somasundaram, Gustavo Solaira, Harry Lanaras, Henry Howard-Jenkins, Huixuan Tang, Hyo Jin Kim, Jaime Rivera, Ji Luo, Jing Dong, Julian Straub, Kevin Bailey, Kevin Eckenhoff, Lingni Ma, Luis Pesqueira, Mark Schwesinger, Maurizio Monge, Nan Yang, Nick Charron, Nikhil Raina, Omkar Parkhi, Peter Borschowa, Pierre Moulon, Prince Gupta, Raul Mur-Artal, Robbie Pennington, Sachin Kulkarni, Sagar Miglani, Santosh Gondi, Saransh Solanki, Sean Diener, Shangyi Cheng, Simon Green, Steve Saarinen, Suvam Patra, Tassos Mourikis, Thomas Whelan, Tripti Singh, Vasileios Balntas, Vijay Baiyya, Wilson Dreewes, Xiaqing Pan, Yang Lou, Yipu Zhao, Yusuf Mansour, Yuyang Zou, Zhaoyang Lv, Zijian Wang, Mingfei Yan, Carl Ren, Renzo De Nardi, and Richard Newcombe. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. 4, 19

[29] Han Feng, Wenchao Ma, Quankai Gao, Xianwei Zheng, Nan Xue, and Huijuan Xu. Stratified avatar generation from sparse observations. In *CVPR*, 2024. 3

[30] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, David Crandall, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C. V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *CVPR*, 2024. 9

[31] Vladimir Guzov, Yifeng Jiang, Fangzhou Hong, Gerard Pons-Moll, Richard Newcombe, C. Karen Liu, Yuting Ye, and Lingni Ma. Hmd^2: Environment-aware motion generation from single egocentric head-mounted device. In *3DV*, 2025. 2, 3

[32] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *CVPR*, 2021. 3

[33] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 4, 5, 6, 16

[34] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J. Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser learning to reconstruct human pose from sparse inertial measurements in real time. *ACM TOG*, 2018. 2

[35] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiV:2310.06825*, 2023. 4

[36] Chiyu Max Jiang, Andre Cornman, Cheolho Park, Ben Sapp, Yin Zhou, and Dragomir Anguelov. Motiondiffuser: Controllable multi-agent motion prediction using diffusion. In *CVPR*, 2023. 3

[37] Hao Jiang and Vamsi Krishna Ithapu. Egocentric pose estimation from human vision span. In *ICCV*, 2021. 3

[38] Jiaxi Jiang, Paul Streli, Manuel Meier, and Christian Holz. Egoposer: Robust real-time egocentric pose estimation from sparse and intermittent observations everywhere. In *ECCV*, 2024. 2, 3, 4, 6, 8, 9

[39] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *ECCV*, 2022. 2, 3, 4, 17, 18

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

[41] Jihyun Lee, Weipeng Xu, Alexander Richard, Shih-En Wei, Shunsuke Saito, Shaojie Bai, Te-Li Wang, Minhyuk Sung, Tae-Kyun Kim, and Jason Saragih. Rewind: Real-time egocentric whole-body motion diffusion with exemplar-based identity conditioning. In *CVPR*, 2025. 3

[42] Jiye Lee and Hanbyul Joo. Mocap everyone everywhere: Lightweight motion capture with smartwatches and a head-mounted camera. In *CVPR*, 2024. 2, 3, 4

[43] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *ICML*, 2019. 5, 16

[44] Junyoung Lee, Inhee Lee, Kyungwon Cho, Jeonghwan Kim, Mingi Choi, Sookwan Han, Lee Chae-Yeon, Wonjung Woo, Gunhee Kim, Jisoo Kim, Jeonghyeon Na, and Hanbyul Joo. Omnirobothome: A multi-camera home platform for real-time human-robot interaction. *arXiv preprint arXiv:2604.28197*, 2026. 20

[45] Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *CVPR*, 2023. 2, 3, 4, 17

[46] Jiefeng Li, Jinkun Cao, Haotian Zhang, Davis Rempe, Jan Kautz, Umar Iqbal, and Ye Yuan. Genmo: A generalist model for human motion. In *ICCV*, 2025. 3, 4

[47] Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, Kevin Bailey, David Soriano Fosas, C. Karen Liu, Ziwei Liu, Jakob Engel, Renzo De Nardi, and Richard Newcombe. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *ECCV*, 2024. 5, 8, 21

[48] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *ICCV*, 2019. 6

[49] Linfei Pan, Johannes Lutz Schönberger, and Marc Pollefeys. Global structure-from-motion meets feedforward reconstruction. In *CVPR*, 2026. 20

[50] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng (Carl) Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *ICCV*, 2023. 9, 21

[51] Chaitanya Patel, Hiroki Nakamura, Yuta Kyuragi, Kazuki Kozuka, Juan Carlos Niebles, and Ehsan Adeli. Uniegomotion: A unified model for egocentric motion reconstruction, forecasting, and generation. In *ICCV*, 2025. 2, 3, 4, 8, 17

[52] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 5, 16

[53] Luis Pineda, Taosha Fan, Maurizio Monge, Shobha Venkataraman, Paloma Sodhi, Ricky TQ Chen, Joseph Ortiz, Daniel DeTone, Austin Wang, Stuart Anderson, Jing Dong, Brandon Amos, and Mustafa Mukadam. Theseus: A library for differentiable nonlinear optimization. In *NeurIPS*, 2022. 6

[54] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *ICCV*, 2021. 6

[55] Helge Rhodin, Christian Richardt, Dan Casas, Eldar Insafutdinov, Mohammad Shafiei, Hans-Peter Seidel, Bernt Schiele, and Christian Theobalt. Egocap: egocentric marker-less motion capture with two fisheye cameras. *ACM TOG*, 2016. 3

[56] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM TOG*, 2017. 3

[57] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 6, 16

[58] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024. 4

[59] Jiangnan Tang, Jingya Wang, Kaiyang Ji, Lan Xu, Jingyi Yu, and Ye Shi. A unified diffusion framework for scene-aware human motion estimation from sparse signals. In *CVPR*, 2024. 3

[60] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. In *NeurIPS*, 2021. 22

[61] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*, 2023. 3, 6

[62] Denis Tome, Thiemo Alldieck, Patrick Peluse, Gerard Pons-Moll, Lourdes Agapito, Hernan Badino, and Fernando De la Torre. Selfpose: 3d egocentric pose estimation from a headset mounted camera. *IEEE TPAMI*, 2020. 3

[63] Tom Van Wouwe, Seunghwan Lee, Antoine Falisse, Scott Delp, and C Karen Liu. Diffusionposer: Real-time human motion reconstruction from arbitrary sparse sensors using autoregressive diffusion. In *CVPR*, 2024. 2

[64] Timo Von Marcard, Bodo Rosenhahn, Michael J Black, and Gerard Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer Graphics Forum*, 2017. 2

[65] Jian Wang, Zhe Cao, Diogo Luvizon, Lingjie Liu, Kripasindhu Sarkar, Danhang Tang, Thabo Beeler, and Christian Theobalt. Egocentric whole-body motion capture with fisheyevit and diffusion-based motion refinement. In *CVPR*, 2024. 3

[66] Jian Wang, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, and Christian Theobalt. Estimating egocentric 3d human pose in global space. In *ICCV*, 2021. 3

[67] Jian Wang, Diogo Luvizon, Weipeng Xu, Lingjie Liu, Kripasindhu Sarkar, and Christian Theobalt. Scene-aware egocentric 3d human pose estimation. In *CVPR*, 2023. 3

[68] Charig Yang, Samiul Alam, Shakhrul Iman Siam, Michael Proulx, Lambert Mathias, Kiran Somasundaram, Luis Pesqueira, James Fort, Sheroze Sheriffdeen, Omkar Parkhi, Carl Ren, Mi Zhang, Yuning Chai, Richard Newcombe, and Hyo Jin Kim. Reading recognition in the wild. In *NeurIPS*, 2025. 9

[69] Chenhongyi Yang, Anastasia Tkach, Shreyas Hampali, Linguang Zhang, Elliot J Crowley, and Cem Keskin. Egoposeformer: A simple baseline for stereo egocentric 3d human pose estimation. In *ECCV*, 2024. 3

[70] Brent Yi, Vickie Ye, Maya Zheng, Yunqi Li, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. Estimating body and hand motion in an ego-sensed world. In *CVPR*, 2025. 2, 3, 4, 6, 17, 22

[71] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Vladislav Golyanik, Shaohua Pan, Christian Theobalt, and Feng Xu. Egolocate: Real-time motion capture, localization, and mapping with sparse body-mounted sensors. *ACM TOG*, 2023. 3

[72] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In *CVPR*, 2022. 2

[73] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM TOG*, 2021. 2

[74] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Physical non-inertial poser (pnp): Modeling non-inertial effects in sparse-inertial human motion capture. In *SIGGRAPH*, 2024. 2

[75] Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. Hawor: World-space hand motion reconstruction from egocentric videos. In *CVPR*, 2025. 2, 4, 8, 20, 21, 22

[76] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE TPAMI*, 2024. 3

[77] Yu Zhang, Songpengcheng Xia, Lei Chu, Jiarui Yang, Qi Wu, and Ling Pei. Dynamic inertial poser (dynaip): Part-based motion dynamics learning for enhanced human pose estimation with sparse inertial sensors. In *CVPR*, 2024. 2

[78] Xiaozheng Zheng, Zhuo Su, Chao Wen, Zhou Xue, and Xiaojie Jin. Realistic full-body tracking from sparse observations via joint-level modeling. In *ICCV*, 2023. 2

A Implementation Details

In this section, we provide additional details to supplement the descriptions in the main text. Code, pre-trained checkpoints, and OmniEgoDB are publicly available on our project page.

Network Architecture Tab. 6 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 [43] use 16 learnable queries that attend to the input sequence, are refined by self-attention, then mean-pooled and passed through an MLP to produce the branch output, *i.e.* $\hat{\beta}$ for the shape branch and $\boldsymbol{\mathcal{G}}$ for the global branch. Throughout the network, all transformer blocks adopt pre-Layer Normalization and GELU activations. In the decoder, the global conditioning vector for the AdaLN-Zero [52] 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 [33] training timesteps, while inference is accelerated using 30 DDIM [57] 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 to downward), we sample spatial parameters uniformly. The sampling ranges are defined as follows (angles in radians):

$$\begin{aligned} \gamma_x \in [0.35, 2.15],\ \gamma_y \in [0.35, 1.35] \\ \phi_x \in [-0.15, 0.15],\ \phi_y \in [0.0, 1.5],\ p \in [2.0, 10.0] \end{aligned} \tag{10}$$

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

$$0.4 \leq \frac{\gamma_y}{\gamma_x} \leq 1.1, \quad \phi_y \leq 1.5 - (\gamma_y - 0.35) \cdot 1.1. \tag{11}$$

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:

$$\begin{aligned} &\text{Short: } \rho \sim \mathcal{U}(0.0, 0.1),\ m = 2.0,\ s = 1.0 \\ &\text{Long: } \rho \sim \mathcal{U}(0.0, 0.2),\ m = 28.0,\ s = 25.0 \end{aligned} \tag{12}$$

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 systematic errors of off-the-shelf wrist estimators, we offset the wrist poses of each hand by a small random rigid transformation in head coordinates, sampled once per sequence, with rotation angles drawn from $\mathcal{N}(0, (\pi/180)^2)$ along each axis (in radians) and translation drawn from $\mathcal{N}(0, 0.005^2)$ (in meters).

Table 6: **Architectural hyperparameters.** Configurations of the transformer modules 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 7: **Additional ablation studies.** Auxiliary loss and root-orientation representation on diverse simulated hand observations. All variants are evaluated without optimization.

Method	MPJPE↓	MPJVE↓	Jerk↓	Hand PE↓	Vis Hand PE↓
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**

Table 8: **Comparison with DSPoser.** Results under DSPoser's 90° circular FoV protocol and AvatarPoser split. Since DSPoser does not model body shape, ours uses a fixed mean shape here. DSPoser numbers are from its paper.

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

C Timestep-Adaptive Guidance Optimization

We provide the objective for the optional timestep-adaptive guidance used during DDIM sampling. This guidance is used only as a lightweight refinement for visible wrist alignment. At each sampling step, we minimize $\mathcal{L}_{\text{opt}}$.

$$\mathcal{L}_{\text{opt}} = \sum_{t=1}^{T} \sum_{j} \left[\sqrt{\bar{\alpha}_n}\, d\big(\tilde{\mathbf{W}}_t^j, \hat{\mathbf{W}}_t^j\big) + s \cdot v_t^j \cdot \sqrt{1-\bar{\alpha}_n}\, d\big(\tilde{\mathbf{W}}_t^j, \mathbf{P}_t^j\big) \right] \tag{13}$$

Here, $\tilde{\mathbf{W}}_t^j = \text{FK}_j(\tilde{\mathbf{M}}_t^0, \hat{\beta})$ and $\hat{\mathbf{W}}_t^j = \text{FK}_j(\hat{\mathbf{M}}_t^0, \hat{\beta})$ are the SE(3) poses of hand joint $j \in \{\text{lHand}, \text{rHand}\}$ in the refined motion $\tilde{\mathbf{M}}^0$ and the denoiser prediction $\hat{\mathbf{M}}^0$, $\mathbf{P}_t^j$ is the observed wrist pose in world coordinates from $\mathbf{H}$, $d(\mathbf{A}, \mathbf{B}) = \| \log(\mathbf{A}^{-1}\mathbf{B})^\vee \|^2$ is the squared distance between two rigid transformations, and $s = 30.0$ is the guidance scale. We optimize only the arm joints [39]. The first term keeps the motion close to the denoiser prediction, while the second pulls visible wrists toward the observations. The weights $\sqrt{\bar{\alpha}_n}$ and $\sqrt{1-\bar{\alpha}_n}$ make the constraint dominate early steps and the prior dominate late steps, keeping the motion on the manifold. Because the denoiser already provides a strong hand-aware motion prior, a few LM iterations are sufficient to refine visible wrist alignment from a plausible initialization. This frame-independent guidance is highly parallelizable and avoids slow sequence-level post-hoc fitting.

D Additional Ablation Studies

We provide additional ablations for training losses and root-orientation representation, complementing the main ablations in Sec. 4.5.

Auxiliary Loss We validate the auxiliary loss $\mathcal{L}_{\text{aux}}$ used in Sec. 3.4. Many diffusion-based motion models [70, 45, 51, 20] optimize only the denoising loss on joint rotations, but small rotation errors can accumulate along the kinematic chain and lead to large end-effector errors. Our auxiliary loss directly regularizes 3D joint positions and foot contacts, making it particularly important for satisfying intermittent hand observations. Removing $\mathcal{L}_{\text{aux}}$ degrades both full-body and end-effector accuracy, increasing Vis Hand PE from 56.98 mm to 60.57 mm, and causes a 39% increase in Jerk. This confirms that $\mathcal{L}_{\text{aux}}$ acts as a geometric regularizer that improves hand alignment and motion smoothness.

Root Representation We also ablate the root-orientation representation. The **No Root Ori** variant mimics EgoAllo [70] by directly stitching the predicted body to $\mathbf{T}^{\text{head}}$, making the root orientation



Figure 8: **Qualitative results across varying camera setups. (Left)** Sudden arm movements in EgoPoser (120) and arms pushed out of bounds in EgoAllo†† at 90° FoV. **(Right)** Misaligned tracking of EgoAllo†† at 180° FoV.

fully determined by the head trajectory. This removes 3DoF of flexibility and leads to unstable or inaccurate body orientation. The **Head** variant restores flexibility by predicting $\mathbf{R}_t^{\text{root}\to\text{head}}$ in head coordinates, but this requires learning motion relative to a volatile head frame. In contrast, our gravity-aligned canonical representation predicts $\mathbf{R}_t^{\text{root}\to\text{cano}}$, providing a stable reference frame while preserving root-orientation flexibility. This yields the best balance between stability and expressiveness.

E Comparison with DSPoser

While DSPoser [21] 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 [39] 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. 8, our device-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.

F Additional Qualitative Results for Camera-Setup Generalization

Fig. 8 provides qualitative examples complementing the camera-setup generalization results in Sec. 4.3. The figure visualizes how fixed-setup specialists fail when the visibility boundary differs from their training regime. In narrow-FoV settings, EgoAllo†† tends to pull invisible hands toward its learned 120° boundary, while in wider-FoV settings it fails to fully exploit hand observations outside this boundary. EgoPoser (120) can produce low hand-position error in some frames by collapsing invisible hands toward a conservative mean pose, but this leads to discontinuous arm motion when observations appear or disappear. In contrast, OmniEgoCap adapts to the changed camera geometry and maintains plausible hand trajectories and smooth full-body motion. These examples support the quantitative trends in Table 2, showing that robust generalization requires modeling camera-induced visibility rather than memorizing a fixed observation regime.



Figure 9: **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 visibility context. In contrast, our full model correctly infers that the invisible hands are outside the FoV.

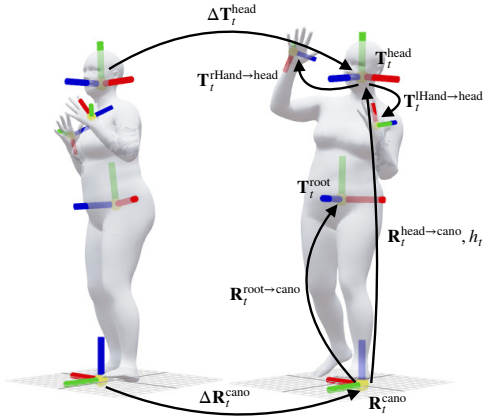

Figure 10: **Representation Visualization.** Illustration of the gravity-aligned canonical frame and the relative geometric transformations used to construct our model's features.

G Qualitative Analysis of Sequence-Level Context

As discussed in Sec. 4.5, Fig. 9 highlights the role of the visibility context $\mathcal{G}$ in resolving out-of-FoV hand ambiguities. In the initial frames, 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 wrists later enter the frontal FoV, the full model uses this sequence-level evidence to infer that the initially unobserved hands were outside the camera's observation boundary. In contrast, **No Global** lacks this broader temporal context and hallucinates invisible hands within the visible region. This supports our claim that $\mathcal{G}$ captures the visibility context from visibility transitions across the full sequence.

H Linear Probing of the Visibility Context

We pair each of the 452 held-out AMASS test motions with 12 cameras sampled independently from our augmentation, giving 5,424 samples, and extract $\mathcal{G}$ with the frozen encoder. The probe target is the camera's visible region over head-relative directions, discretized into 2,592 cells on the unit sphere. We fit a ridge regression from $\mathcal{G}$ to the per-cell visibility and report the area-weighted IoU between the predicted and true regions, computed over the cells within 5° of a direction visited by the wrists (19.4% of the sphere). We use 5-fold cross-validation with folds grouped by motion, and select the ridge strength and the IoU threshold on a motion-grouped split of the training folds. The baselines follow the same protocol. Shuffled labels fits the probe to targets shuffled across samples, mean training camera predicts the visible region of the average training camera, and coarse visibility statistics replaces $\mathcal{G}$ with four per-sequence features, namely the visible fraction, its square, the rate of visibility transitions, and the log sequence length. For stature, we fit the same ridge probe from $\hat{\beta}$ or $\mathcal{G}$ to the wearer's height with leave-one-subject-out cross-validation over the 36 test subjects and report the mean absolute error.

I OmniEgoDB Benchmark Details

In Sec. 4.4, we introduce OmniEgoDB as a real-device mocap benchmark for evaluating device-agnostic egocentric motion reconstruction. Here, we provide additional details on the capture setup, processing and calibration, and qualitative analysis.

Capture Setup OmniEgoDB pairs synchronized egocentric videos with Xsens bodysuit ground-truth motion. We capture natural daily activities using Project Aria [28], GoPro [2], and Quest [3].



Figure 11: **Overview of OmniEgoDB.** The left side shows participants equipped with three different egocentric devices and Xsens mocap suits. The right side shows representative egocentric frames from OmniEgoDB, illustrating diverse viewpoints, scenes, and activities.

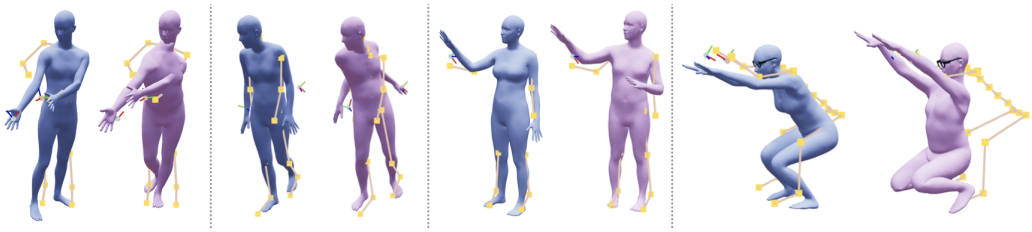

Figure 12: **Qualitative Analysis on OmniEgoDB.** From left to right, we show a GoPro depth-direction wrist error, a Quest false-positive wrist observation, temporarily missing Aria wrist observations, and an Aria case with inconsistent body shape estimation. Blue, purple, and yellow denote OmniEgoCap, EgoAllo with optimization, and ground truth, respectively.

These devices represent different egocentric sensing regimes. Aria provides a smart-glasses setup with built-in perception services, Quest provides an XR headset setup with onboard VIO and real-time hand tracking, and GoPro provides an RGB-only action-camera setup without native head or hand tracking. The benchmark contains 50 sequences from four subjects totaling 177.2 minutes, with 21 Aria sequences (67.1 min), 20 GoPro sequences (68.1 min), and 9 Quest sequences (42.0 min). The sequences are captured at three indoor locations (an office, a hallway, and the OmniRobotHome platform [44]) and two outdoor locations (a plaza and a parking lot). The activities cover locomotion, object manipulation, and dynamic motions, including walking, jumping, cooking, dancing, and exercising. For GoPro captures, we do not fix the head-mount tilt angle, allowing natural variation in camera mounting.

Processing and Calibration OmniEgoDB is processed to provide the input signals required by OmniEgoCap, including head trajectories, intermittent hand visibility, and 6-DoF wrist observations. The processing pipeline differs across devices depending on the available tracking signals. For Aria, we use Project Aria's Machine Perception Services to obtain the head trajectory and hand tracking signals. For Quest, we use the headset's VIO-based head tracking and real-time hand tracker. For GoPro, which does not provide native head or hand tracking, we estimate the camera trajectory with GlueMap [49], recover metric scale from the wearer's camera height, and estimate wrist observations with HaWoR [75]. For Aria (RGB), we undistort the RGB stream to a pinhole camera and estimate wrist observations with the same pipeline, while keeping the MPS head trajectory.

All captures are performed with Xsens suits and gloves to obtain ground-truth full-body motion. The gloves can partially cover the back of the hands and reduce the reliability of image-based hand detection. This is especially challenging for GoPro and Aria (RGB), where hand visibility and wrist observations must be estimated from RGB video alone. We therefore use SAM 3 [19], prompted with "hand", to obtain hand bounding boxes for HaWoR.

Table 9: **Quantitative results on Aria Digital Twin.** # denotes methods evaluated with their official checkpoints. – denotes entries not applicable to UniEgoMotion, which needs RGB frames and predicts at 10 FPS.

Method	ADT (Stereo) 34 seqs, 52.5 min				ADT (RGB) 34 seqs, 52.5 min			
	MPJPE↓	MPJVE↓	Jerk↓	Hand PE↓	MPJPE↓	MPJVE↓	Jerk↓	Hand PE↓
EgoAllo# (w/o opt)	159.06	18.36	0.117	293.74	159.06	18.36	0.117	293.74
EgoAllo# (w/ opt)	145.78	19.61	0.075	212.97	150.04	26.08	0.256	204.78
EgoPoser#	127.50	32.22	0.429	236.36	126.86	46.84	0.950	206.69
EgoPoser (120)	131.06	40.20	0.857	243.56	130.56	61.49	1.511	213.93
UniEgoMotion#	–	–	–	–	141.99	–	–	232.31
Ours	127.14	**17.49**	**0.062**	193.10	125.86	**17.66**	**0.067**	199.55
Ours (w/ opt)	**125.40**	17.91	0.074	**189.80**	**125.63**	19.76	0.144	**197.32**

We temporally synchronize each egocentric recording with the corresponding Xsens motion sequence by estimating the temporal offset that best aligns the magnitudes of the egocentric device and Xsens head angular velocities. Following the calibration protocol used in Nymeria [47], we estimate a fixed hand-eye transform between each egocentric device trajectory and the Xsens head frame. Since Xsens motion can accumulate drift over time, we use the egocentric camera trajectory as the reference coordinate system and align the Xsens motion to this camera/world frame. This yields a common coordinate system for evaluating reconstructed SMPL-H motion against Xsens ground truth.

Analysis of Benchmark Results Table 3 shows consistent gains across Aria, GoPro, and Quest, with different error patterns across baselines. EgoAllo is sensitive to inaccurate or jittery wrist constraints, especially on GoPro, where both camera motion and wrist observations must be estimated from RGB video. Its optimization can propagate these errors to the arms or whole body, producing implausible configurations such as body interpenetration or twisted arms. In contrast, OmniEgoCap relies on a hand-aware diffusion prior and sequence-level visibility context, maintaining stable full-body motion while optional guidance mainly improves Hand PE.

Figure 12 further compares OmniEgoCap with EgoAllo under noisy or unreliable wrist observations. In the GoPro example, a large depth-direction wrist error causes EgoAllo to overfit the erroneous constraint and produce body interpenetration, whereas OmniEgoCap yields a plausible reaching motion. In the Quest example, a transient false-positive wrist observation is corrected in the following frames, and OmniEgoCap uses this temporal context to suppress the outlier, while EgoAllo remains over-constrained by the erroneous signal. The two Aria examples show complementary failures of EgoAllo. When wrist observations are temporarily missing, EgoAllo hallucinates unstable arm motion. In another case, inconsistent body shape estimation distorts the whole-body configuration and leads to an incorrect squat motion. These examples highlight that OmniEgoCap is more robust to transient tracking errors by leveraging sequence-level visibility context and a consistent body shape.

J Evaluation on Aria Digital Twin

We evaluate all methods zero-shot on Aria Digital Twin (ADT) [50], which pairs Aria recordings with OptiTrack body motion. We use 34 of the 35 single-person recordings, excluding one whose MPS hand tracking is unavailable, and score the 15 joints shared by SMPL-H and the OptiTrack skeleton at 30 FPS. As for OmniEgoDB, ADT (Stereo) uses the MPS head trajectory and hand tracking, and ADT (RGB) uses HaWoR [75] wrists from the RGB stream with the same head trajectory. Because the OptiTrack head joint lies at a different point than the SMPL-H head, we shift the head trajectory given to our model and EgoPoser by a fixed offset estimated once from three recordings. Table 9 shows that OmniEgoCap achieves the lowest error on all four metrics in both settings, and EgoAllo with optimization remains 18–24 mm behind in MPJPE. The official EgoPoser comes close in MPJPE and is more accurate in recordings where the wearer is mostly seated, such as desk work, which we attribute to its regression toward the conditional mean pose fitting static seated postures. Its motion is, however, far less smooth, with roughly twice the MPJVE and about seven times the Jerk of ours in ADT (Stereo), whereas OmniEgoCap produces smooth and natural motion.

K Representation Visualization

We provide visual illustrations of the coordinate representations detailed in Sec. 3.2 of the main text. Following EgoAllo [70], 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. 10 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 ($\mathbf{R}_t^{\text{root}\to\text{cano}}$).

L 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 [4] and Quest [3], 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 [60] backend within a modified HaWoR pipeline [75]. We apply RANSAC to this point cloud to fit the ground plane, using its normal to define the gravity-aligned z-axis.

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.

M Additional Qualitative Results in the Wild

In Fig. 13, we provide additional in-the-wild reconstruction results captured across diverse setups, including an iPhone 15 (26mm lens), Meta Quest 3 [3], and Project Aria [4]. 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 visit our project page for animated visualizations.

N Failure Cases

We further investigate representative failure cases in Fig. 14, which are shared by head-centric egocentric reconstruction methods, including both EgoAllo [70] and our framework. First, seated poses remain inherently ambiguous in an egocentric view, as similar head and hand trajectories may

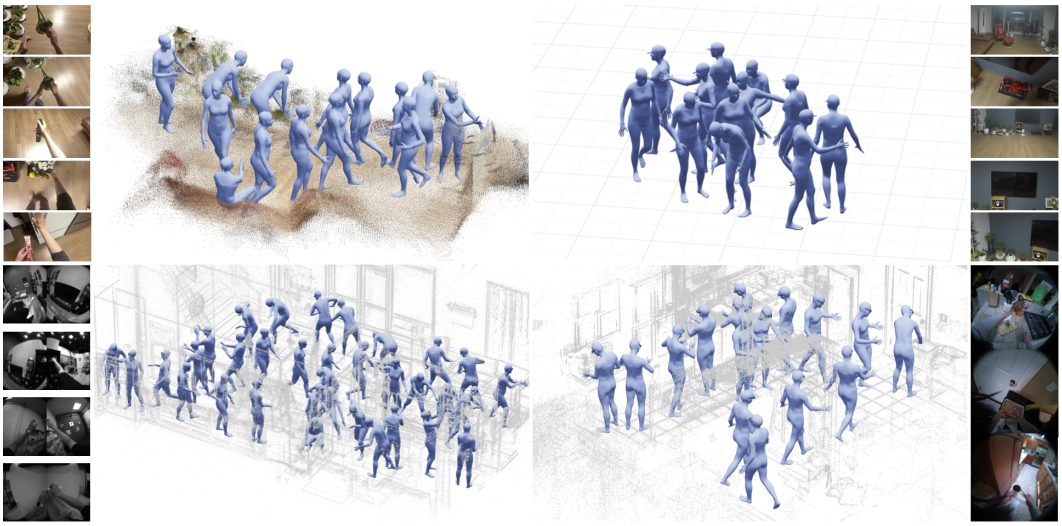

Figure 13: **Additional qualitative results on in-the-wild data.** Reconstructions from (Upper Left) iPhone 15, (Upper Right) Quest, (Lower Left) Aria Stereo (SLAM), and (Lower Right) Aria Monocular (RGB).

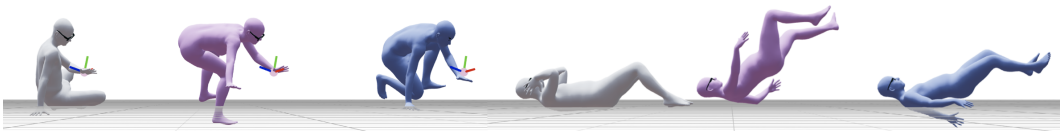

Figure 14: **Representative failure cases.** Gray corresponds to the ground truth, blue to Ours (w/ opt), and purple to EgoAllo (w/ opt).

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.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction clearly reflect the contribution of the paper and are backed by the experimental results in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper addresses the limitation of the method briefly in the discussion section and discusses detailed failure cases in Sec. N of supplementary materials.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper is primarily empirical and does not include theoretical results requiring formal assumptions or proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides detailed explanation of the dataset preprocessing scheme, model architecture, training hyperparameters, and settings used to conduct quantitative evaluation.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The training and evaluation code, the pre-trained checkpoint, and OmniEgoDB are publicly available on our project page, with instructions to reproduce the main results. AMASS must be obtained separately under its own license.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the experimental settings such as how the data were split and preprocessed, training hyperparameters, and optimizer details in Sec. 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper does not report error bars or confidence intervals, as there was no repetition of the experiments multiple times. That said, our evaluation is conducted on fixed test sets of 452 AMASS sequences, 50 OmniEgoDB sequences, and 34 Aria Digital Twin recordings, and we report averaged results accompanied by detailed evaluation protocols to ensure reproducibility.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper specifies the computing resources used and execution time in Sec. 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research respects the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

Justification: Our paper addresses egocentric motion reconstruction, and the dataset we release is anonymized and access-controlled (see Safeguards), and is not applicable for generating fake profiles, surveillance and spreading disinformation. We believe our research is not directly related to malicious downstream applications, thereby making broader societal impact of our research minimal.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This work does not involve any models or datasets with a high potential for misuse. Before releasing OmniEgoDB, we replaced names with IDs, removed audio, eye-tracking, and GPS data, and blurred bystander faces and license plates. Access requires agreeing to non-commercial use (CC BY-NC 4.0) and to no re-identification.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We use the existing datasets AMASS, Aria Digital Twin, Ego-Exo4D, Reading in the Wild, and EPIC-KITCHENS, the SMPL-H body model, and the released code and checkpoints of EgoAllo, EgoPoser, UniEgoMotion, HaWoR, SAM 3, and GlueMap. We cite their original papers and use each under its license.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We introduce two new assets, our model and the OmniEgoDB benchmark. Both are released with documentation, the code under the MIT License, the checkpoint under CC BY-NC 4.0, and OmniEgoDB under CC BY-NC 4.0.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our paper does not have any study participant or subject nor crowdsourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Our core methodology does not contain any usage of LLMs, and LLMs were used only for writing, editing and formatting purposes.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.