EgoMo3R: Joint Egocentric Motion and Scene Reconstruction

Chaitanya Patel[1] **Yusuke Kato**[1,2] **Rin Takahira**[3] **Kazuki Kozuka**[3]
Juan Carlos Niebles[1] **Ehsan Adeli**[1]

[1]Stanford University [2]Panasonic R&D Company of America
[3]Panasonic Holdings Corporation

https://chaitanya100100.github.io/EgoMo3R/

Abstract

Understanding holistic 3D human motion and the surrounding environment from egocentric video is fundamental for applications in AR/VR and robotics. Although these tasks are inherently coupled, existing methods treat them separately. In particular, egocentric motion reconstruction methods typically assume access to metric, gravity-aligned camera trajectories from specialized inertial sensors. Recovering such trajectories from monocular egocentric video remains a significant challenge due to rapid head movements and motion blur. We present EgoMo3R, a joint optimization framework that simultaneously reconstructs 3D human motion and scene geometry from casual egocentric video. It leverages the robustness of low-level visuo-geometric primitives, including monocular depth, optical flow, and point tracks, integrated with a diffusion-based human motion prior via first-order optimization. We propose an alternating optimization scheme that establishes a synergistic feedback loop: scene reconstruction provides physical grounding and trajectory conditioning for the motion prior, while the motion prior regularizes the camera trajectory and anchors the scene in a metric, gravity-aligned world frame. We evaluate EgoMo3R on the highly dynamic sequences of the EgoExo4D and Nymeria datasets and show that it outperforms specialized baselines in both 3D scene reconstruction and egocentric motion estimation.

1 Introduction

A holistic 3D understanding of human motion and the surrounding environment from egocentric video is crucial for human-centered applications such as AR/VR, assistive technologies, and healthcare, while also offering a scalable pathway for robot learning. In the egocentric setting, 3D human motion and scene reconstruction are inherently coupled: recovering the user's full-body motion requires precise trajectory estimation of the egocentric camera that naturally follows the head motion. Traditional *visual* SLAM methods [4, 11, 47, 67, 68] typically treat camera trajectory estimation as a second-order optimization problem relying on multi-view correspondences to optimize for camera poses and sparse geometric representations. However, these methods often struggle on dynamic egocentric *locomotion* videos with narrow fields of view and motion blur. This 'inside-out' perspective results in sparse, low-quality correspondences leading to an underconstrained optimization, particularly during the rapid, rotation-heavy head movements common in daily activities [5]. While recent data-driven feedforward reconstruction methods [71, 74] offer a promising alternative, they often struggle to generalize due to the egocentric domain gap. Consequently, existing egocentric motion reconstruction methods largely bypass these geometric challenges by assuming access to a perfect metric and gravity-aligned camera trajectories from a dedicated *inertial* SLAM pipeline [80, 49, 24]. While joint modeling of humans and 3D scenes has shown promise in third-person settings [46, 8, 58], these

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

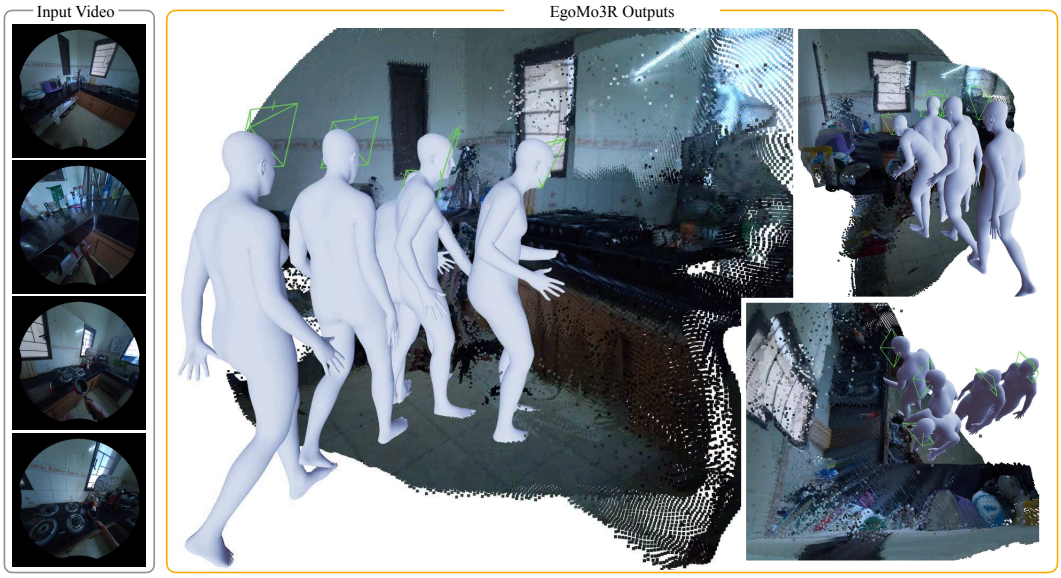

Figure 1: Given an egocentric video, EgoMo3R optimizes for camera poses, 3D pointmaps, and human motion in a common gravity-aligned metric reference frame. We also show two additional views of our scene and motion reconstruction on the side.

benefits do not trivially translate to the egocentric domain, where the human body remains largely invisible.

In this paper, we take the first step towards the joint modeling of egocentric motion and scene geometry with EgoMo3R (see Fig. 1). We demonstrate that EgoMo3R enables the synergistic reconstruction of both 3D human motion and the surrounding environment from a casually captured egocentric video. At its core, our approach formulates this task as a first-order joint optimization problem that recovers camera poses, scene geometry, and human motion in a single reference frame by leveraging data-driven low-level geometry and motion priors.

Our first insight is that while feedforward 3D reconstruction models often struggle with the egocentric domain gap, low-level visuo-geometric primitives, such as monocular depth, optical flow, and point tracks, exhibit significantly higher robustness. Although these cues can be individually noisy and mutually inconsistent, they provide complementary geometric constraints that, when aggregated within a global optimization framework, yield a coherent reconstruction. Specifically, our formulation utilizes monocular depth as a geometric base and point tracking for long-term temporal correspondence. To ensure convergence, we leverage optical flow to initialize the camera trajectory, effectively guiding our first-order optimization away from the poor local minima typical of rapid, rotation-heavy egocentric locomotion sequences.

Our second key insight is that egocentric human motion and scene reconstruction can be mutually reinforcing. Accurate 3D scene reconstruction with a precise camera trajectory directly informs head motion and offers explicit physical grounding for the body within the 3D environment. Conversely, 3D human motion serves as a unique prior for camera trajectory estimation, beyond simple constraints such as constant velocity or smoothness, common in traditional SLAM. Human motion priors also anchor the optimization in a canonical upright (metric, gravity-aligned, grounded) world frame, thereby improving the reconstruction on challenging egocentric videos.

However, unlike exocentric scenarios, jointly optimizing human motion with scene geometry is non-trivial in the egocentric domain, where the body remains largely out of view. To address this, we employ a motion diffusion generative model as a robust motion prior. In particular, we use a motion diffusion model conditioned on $\mathrm{SE}(3)$-invariant egocentric trajectory and formulate an alternating optimization framework that iteratively refines the 3D scene and human motion. This setup creates a synergistic feedback loop: improved camera pose provides more accurate conditioning for the motion prior, which in turn improves physically grounded reconstruction. We evaluate EgoMo3R on EgoExo4D [18] and Nymeria [42] datasets and show that it outperforms strong baselines in both motion and scene reconstruction. Finally, we provide an extensive ablation of our

optimization framework, including the technical design and integration of neural motion priors into our reconstruction pipeline.

In summary, we make the following contributions: (1) We present a joint first-order optimization framework that simultaneously recovers metric camera poses, scene geometry, and human motion from a single egocentric video. (2) We show that integrating complementary low-level visuo-geometric primitives and motion prior mitigates the challenging egocentric domain gap. (3) Our alternating optimization incorporates a diffusion-based motion prior to provide physical grounding and gravity alignment for both the human motion and the reconstructed scene. (4) EgoMo3R outperforms strong specialized baselines in both 3D scene reconstruction and egocentric motion estimation.

2 Related Works

Egocentric Motion Reconstruction: The front-facing nature of egocentric RGB cameras [15] presents significant challenges for motion reconstruction due to limited body visibility. To compensate for this, many works use simulation-based physical motion priors [41, 84, 83], data-driven motion diffusion priors [49, 32, 80, 24, 19, 6], or autoregressive motion priors [72, 20] learned from large-scale motion datasets. However, these priors critically depend on an accurate, metric, and gravity-aligned egocentric camera trajectory. Several recent works [49, 80, 9, 19] sidestep this challenge by assuming access to such trajectories from highly optimized inertial SLAM systems [15] with multi-modal sensors, which are not always available in casual capture scenarios. Although classical monocular SLAM or VIO [67, 68] methods can provide an $\mathrm{SE}(3)$-invariant trajectory, these estimates are typically inaccurate and prone to significant jitter in egocentric setup. To remedy this, some methods [32, 41, 84, 83] attempt to learn camera pose estimation directly from optical flow features. We demonstrate that such flow-based learned pose estimation models struggle with egocentric locomotion in cluttered scenes, which subsequently degrades downstream motion reconstruction. In contrast, EgoMo3R presents a robust optimization framework that simultaneously recovers camera poses, 3D scene geometry, and human motion from casual egocentric videos in a synergistic manner.

Camera Pose Estimation & 3D Reconstruction: Classic SfM [61, 62, 48, 1] and SLAM [4, 11, 47] frameworks typically formulate camera pose and 3D scene estimation as a second-order bundle adjustment problem, relying on multi-view correspondences from locally invariant keypoint descriptors [39, 59]. To improve robustness, various approaches integrate data-driven components ranging from learned sparse features [60, 12] to dense correspondence learning [68, 67, 65]. Recent methods [87, 70, 64, 35, 23, 73] further leverage robust priors, such as long-term point tracking [25, 14, 26], monocular depth [75, 79], video depth [7], and semantic segmentation [29, 56], to handle dynamic scenes and sparse views. While several works have adapted similar strategies to the egocentric setting [85, 31, 82], they struggle on egocentric locomotion videos with rapid, rotation-heavy camera movement. Although recent end-to-end feedforward models [77, 30, 86, 71, 74] are promising due to their simplicity, their performance often degrades on challenging egocentric inputs. Closing this gap through finetuning is difficult since existing egocentric datasets lack large-scale, dense, high-quality metric 3D groundtruth. EgoMo3R addresses these challenges by combining robust visuo-geometric primitives with a diffusion human motion prior, ensuring accurate, physically grounded, and gravity-aligned reconstruction.

Joint Human and Scene Reconstruction: Several works have leveraged the presence of humans within a scene to overcome the limitations of classical reconstruction approaches. For instance, Ma et al. [43] and Xu et al. [78] use 3D human pose estimation and person re-identification, respectively, to define virtual correspondences for geometry estimation in extreme views. Recent methods [51, 46, 58, 38] perform joint 3D reconstruction of humans and scenes by optimizing global objectives on top of independent predictions. More recently, Human3R [8] builds upon CUT3R [74] using visual prompting to efficiently train for simultaneous 3D human mesh and scene reconstruction in a feedforward manner. These approaches demonstrate that joint reconstruction is not only achievable but that the two tasks can be mutually beneficial [88, 8]. However, such methods rely on explicit human-related visual cues within the image. In contrast, the user's body remains largely invisible in egocentric viewpoints. In this setting, human-specific cues are not encoded in the raw pixels, but rather in the camera motion that generated them. EgoMo3R leverages these implicit motion cues for the synergistic reconstruction of humans and scenes in the egocentric domain.

3 Method

3.1 Problem Formulation

Given an egocentric video sequence $\boldsymbol{I}_{1:N} \in \mathbb{R}^{N\times H\times W\times 3}$ captured by a head-mounted camera, our goal is to jointly estimate camera trajectory, 3D scene geometry, and the underlying 3D human motion in a *physically grounded* canonical world frame with known gravity direction and ground plane. We denote the camera poses as extrinsics $\boldsymbol{C} \in \mathbb{R}^{N\times 4\times 4}$ representing camera-to-world $SE(3)$ transformations. The 3D scene geometry is represented by a sequence of dense pointmaps [77] $\boldsymbol{P} \in \mathbb{R}^{N\times H\times W\times 3}$, where each element $\boldsymbol{P}_{n,i,j} \in \mathbb{R}^3$ corresponds to the 3D coordinates of pixel (i, j) at time n in the local camera coordinate frame. The global 3D scene representation $\boldsymbol{P}^W$ in the world frame is obtained by transforming local pointmaps $\boldsymbol{P}$ using the estimated camera poses $\boldsymbol{C}$. The 3D human motion $\boldsymbol{M}_{1:N}$ is defined by SMPL-X parameters [50] $M_n = (\boldsymbol{R}^r_n, \boldsymbol{t}^r_n, \theta_n)$ comprising root rotation $R^r_n \in \mathbb{R}^3$, translation $t^r_n \in \mathbb{R}^3$ (at pelvis), and body pose angles $\theta_n \in \mathbb{R}^{21\times 3}$.

3.2 Visual Geometry Priors

While end-to-end egocentric reconstruction often struggles with egocentric domain shift, low-level visuo-geometric primitives offer a more robust basis for optimization. We formulate a first-order optimization framework that treats these primitives as soft constraints. Rather than treating 3D points as free variables, we adopt per-frame relative depth $\boldsymbol{D}^r \in \mathbb{R}^{N\times H\times W}$, predicted by an off-the-shelf monocular depth estimator [76], as our base geometry. While modern monocular estimators [75, 76, 79] provide robust relative depth, their metric scale consistency often degrades in egocentric settings. To address this, we optimize per-frame scale factors $\boldsymbol{S} \in \mathbb{R}^N$ to derive metric depth estimates $D_n = S_n D^r_n$. 2D pixels in $\boldsymbol{I}$ can now be backprojected into 3D using depth $\boldsymbol{D}$ and known camera intrinsics $K \in \mathbb{R}^{3\times 3}$ to obtain camera-frame pointmaps $\boldsymbol{P}$ as $P_{n,i,j} = K^{-1}[i, j, D_{n,i,j}]^T$.

Additionally, we use an off-the-shelf optical flow model [66] to predict dense optical flow $\boldsymbol{O} \in \mathbb{R}^{N\times H\times W\times 2}$, where $\boldsymbol{O}_{n,i,j} = (\delta u, \delta v)$ implies that pixels $\boldsymbol{I}_{n,i,j}$ and $\boldsymbol{I}_{n+1,i+\delta u,j+\delta v}$ correspond to the same 3D point. Formally, this induces the constraint $C_{n+1}P_{n+1,i+\delta u,j+\delta v} = C_n P_{n,i,j}$. While these constraints are theoretically sufficient to estimate $\boldsymbol{C}$, they only provide local supervision between consecutive frames. In egocentric settings, monocular depth and optical flow often degrade due to motion blur and narrow fields of view. Consequently, small errors compound over time, leading to significant trajectory drift and distorted 3D reconstructions. Traditional SLAM pipelines [11, 67, 23, 35] typically employ a two-tiered system to correct such drift via loop closure. Instead, we leverage an off-the-shelf 2D point tracking model [26] to provide robust, long-range correspondences $\boldsymbol{T}$ to maintain reconstruction alignment [87, 64, 73, 82]. We demonstrate in our experiments that relying on visuo-geometric priors already yields strong performance in camera pose estimation and 3D scene reconstruction. However, these purely visual constraints remain insufficient for recovering the underlying human motion, gravity direction, or floor grounding.

3.3 Motion Prior

The egocentric setup is unique in that the camera trajectory $\boldsymbol{C}$ directly follows the human head motion. Prior works [32, 80, 49] have leveraged this to perform egocentric human motion reconstruction, often without access to paired datasets, achieving impressive results even when the body largely remains out-of-view. These methods typically formulate motion reconstruction as sampling from $p(\boldsymbol{M}|\boldsymbol{C})$, where $\boldsymbol{C}$ is assumed to be a metric, gravity-aligned, and floor-grounded camera trajectory. This conditional distribution $p(\boldsymbol{M}|\boldsymbol{C})$ is usually modeled via a conditional motion diffusion model $f_\phi(\boldsymbol{M}_t, \boldsymbol{C})$ with learnable parameters ϕ, where $\boldsymbol{M}_t$ denotes the noisy motion at diffusion timestep t.

However, incorporating a strong diffusion-based motion prior into our joint optimization framework is non-trivial for several reasons. First, unlike prior works, we do not have *a priori* access to a gravity- and floor-aligned camera trajectory $\boldsymbol{C}$. Furthermore, training unconditional diffusion models remains inherently difficult for complex distributions such as images [33, 22] and human motion [55, 81]. The lack of direct body observations makes it difficult to guide the denoising process to adhere to a suboptimal, unaligned trajectory $\boldsymbol{C}$ that is itself undergoing optimization (See the Appendix for more discussion).

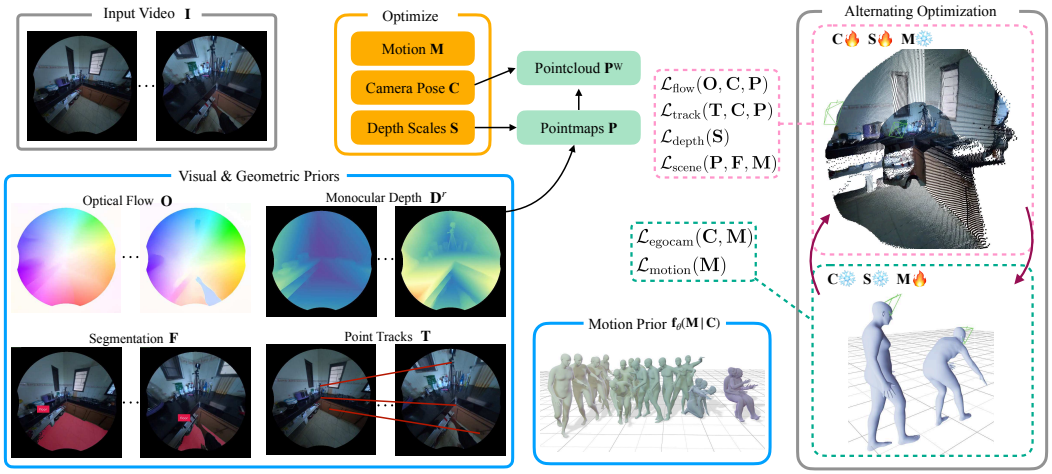

Figure 2: Given an input video $\boldsymbol{I}$, we leverage a monocular depth estimator to obtain relative depth $\boldsymbol{D}^r$ as a geometric base, optimizing only the per-frame scales $\boldsymbol{S}$ to derive metric depth $\boldsymbol{D}$. Our framework jointly optimizes for camera poses $\boldsymbol{C}$, which transform metric depth into a global world-frame pointmaps $\boldsymbol{P}^W$, and 3D human motion $\boldsymbol{M}$. We employ an alternating optimization strategy: (1) **Scene Refinement** updates $\boldsymbol{C}$ and $\boldsymbol{S}$ using optical flow, point tracks, and a scene-alignment loss to anchor the reconstruction in a canonical frame provided by motion; (2) **Motion Alignment** refines $\boldsymbol{M}$ by leveraging a diffusion motion prior conditioned on $\boldsymbol{C}$. This synergistic loop prevents optimization collapse and ensures physically grounded joint 3D scene and motion reconstruction.

To remedy this, we instead model the distribution $p(\boldsymbol{M}|\mathcal{A}(\boldsymbol{C}))$, where $\mathcal{A}(\boldsymbol{C})$ is an SE(3)-invariant canonicalized camera trajectory defined as $\mathcal{A}(\boldsymbol{C})_i = C_0^{-1}C_i$. We train a conditional diffusion model $f_\phi(\boldsymbol{M}_t, \mathcal{A}(\boldsymbol{C}))$ to generate physically grounded motion $\boldsymbol{M}$ conditioned on this relative trajectory. This design is motivated by the observation that the SE(3)-invariant trajectory remains highly informative of the underlying body motion. Crucially, it allows for the stable integration of the motion prior during the early stages of optimization when $\boldsymbol{C}$ is not yet optimal or canonically aligned. It not only provides unique prior on the camera trajectory $\boldsymbol{C}$ but also improves the overall 3D reconstruction by implicitly enforcing gravity direction and ground-plane priors.

3.4 Optimization

Given visual primitives and motion prior, we formulate the following optimization problem:

$$(\boldsymbol{C}^*, \boldsymbol{M}^*, \boldsymbol{S}^*) = \underset{\boldsymbol{C},\boldsymbol{M},\boldsymbol{S}}{\operatorname{argmin}} \; \mathcal{L}_{\text{flow}} + \mathcal{L}_{\text{track}} + \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{scene}} + \mathcal{L}_{\text{egocam}} + \mathcal{L}_{\text{motion}} \tag{1}$$

The term $\mathcal{L}_{\text{flow}} = \sum_{n,i,j} \|O_{n,i,j} - \Pi_K(C_{n+1}^{-1} C_n P_{n,i,j})\|_2^2$ minimizes the reprojection error between the observed optical flow and the flow induced by the estimated camera motion, where Π_K denotes the camera projection using intrinsics K. Similarly, $\mathcal{L}_{\text{track}}$ enforces long-range geometric consistency across non-consecutive frames using the point tracks $\boldsymbol{T}$. We use Median absolute deviation filtering on $\mathcal{L}_{\text{flow}}$ and $\mathcal{L}_{\text{track}}$ to mitigate the influence of bad correspondences from dynamic objects. To regularize the depth estimation, we define $\mathcal{L}_{\text{depth}} = \sum_n \|S_n - S_n^m\|_2^2$, which constrains the optimized per-frame scales $\boldsymbol{S}$ to remain near the initial metric scales $\boldsymbol{S}^m$ provided by an off-the-shelf estimator [76]. While $\boldsymbol{S}^m$ may suffer from monocular ambiguity in egocentric setting, this term prevents the optimization from converging to degenerate solutions in the presence of noisy or no correspondences.

The term $\mathcal{L}_{\text{scene}} = \sum_{n,i,j} F_{n,i,j} \cdot d(\boldsymbol{M}, P_{n,i,j}^W)$ ensures the canonical world frame alignment between points $\boldsymbol{P}^W$ and motion $\boldsymbol{M}$, where $\boldsymbol{F}$ is a binary ground segmentation mask [37, 56] and $d(\boldsymbol{M}, a)$ computes the point-to-plane distance between a 3D point a and the ground plane implicitly defined by the human motion $\boldsymbol{M}$. The coupling between the camera and the human motion is enforced via $\mathcal{L}_{\text{egocam}} = \|\boldsymbol{C} - \mathcal{H}(\boldsymbol{M})\|_2^2$, which aligns the camera trajectory $\boldsymbol{C}$ with the head trajectory $\mathcal{H}(\boldsymbol{M})$ derived from the SMPL-X parameters. Finally, we incorporate the generative motion prior via a denoising score-matching loss $\mathcal{L}_{\text{motion}} = \mathbb{E}_{t\sim\mathcal{U}[0,1]}\left[\|f_\phi(\boldsymbol{M}_t, \mathcal{A}(\boldsymbol{C})) - \boldsymbol{M}\|_2^2\right]$. It drives the motion $\boldsymbol{M}$ toward the manifold of physically plausible human movements conditioned on $\mathcal{A}(\boldsymbol{C})$.

3.5 Alternating Optimization

We observe that directly minimizing Eq. 1 via gradient descent often yields suboptimal results and poor alignment between the camera trajectory $\boldsymbol{C}$ and human motion $\boldsymbol{M}$. This instability stems primarily from two factors. First, tracking failures can introduce significant outlier correspondences that bias the optimization. While we employ Median Absolute Deviation (MAD) filtering – analogous to RANSAC [16] in traditional SLAM – a poor initialization can lead to the erroneous rejection of valid correspondences, causing the optimization to converge to a local minimum. To mitigate this, we initialize the camera trajectory by optimizing $\boldsymbol{C}$ solely with the $\mathcal{L}_{\text{flow}}$ objective. As discussed in Sec. 3.2, $\mathcal{L}_{\text{flow}}$ is a 2D reformulation of the 3D constraint $C_{n+1}P_{n+1,i+\delta u,j+\delta v} = C_n P_{n,i,j}$, which has a closed-form solution using Procrustes analysis [17]. This warm-start enables more accurate outlier detection and stabilizes subsequent optimization steps. Second, the simultaneous optimization of $\boldsymbol{C}$ and $\boldsymbol{M}$ introduces noisy gradients from $\mathcal{L}_{\text{motion}}$, as our diffusion prior $f_\phi(\boldsymbol{M}_t, \mathcal{A}(\boldsymbol{C}))$ is itself a function of the evolving camera trajectory $\boldsymbol{C}$ [34]. To resolve this, we propose an alternating optimization scheme that iterates between two distinct modes:

1. Camera and Scene Refinement: We optimize the camera poses $\boldsymbol{C}$ and depth scales $\boldsymbol{S}$ using $\mathcal{L}_{\text{flow}}$, $\mathcal{L}_{\text{track}}$, $\mathcal{L}_{\text{depth}}$, and $\mathcal{L}_{\text{scene}}$, while keeping the human motion $\boldsymbol{M}$ fixed. This stage refines the 3D scene reconstruction while leveraging the gravity and ground-plane priors from fixed motion $\boldsymbol{M}$.

2. Motion and Prior Alignment: We optimize only the motion parameters $\boldsymbol{M}$ using $\mathcal{L}_{\text{motion}}$ and $\mathcal{L}_{\text{egocam}}$, while keeping $(\boldsymbol{C}, \boldsymbol{S})$ fixed. This stage enforces the generative motion prior and eliminates physical implausibilities in the camera trajectory, such as jitters or sudden jumps. It also defines the canonical frame to improve 3D reconstruction. This mode can be interpreted as sampling from the posterior $p(\boldsymbol{M}|\mathcal{A}(\boldsymbol{C}))$ with guidance [13] provided by the camera-head alignment loss $\mathcal{L}_{\text{egocam}}$.

We demonstrate empirically that this alternating strategy significantly improves joint reconstruction results compared to direct optimization. Please refer to the Appendix for more details.

4 Experiments

4.1 Dataset

We evaluate our approach on EgoExo4D [18] and Nymeria [42] datasets. Both datasets provide multimodal egocentric data captured via Project Aria smartglasses [15]. The Aria MPS kit provides high-quality groundtruth camera poses obtained through an extensive offline VIO and SLAM system that fuses data from multiple sensors [15]. All frames are undistorted to a linear pinhole camera following the official Project Aria guidelines [15], at 512×512 resolution with a focal length of 150 px and without cropping. All methods, including EgoMo3R, receive these identical rectified frames. To evaluate 3D reconstruction, we filter the MPS SLAM point clouds using 2D observation metadata to retain only high-confidence points. For 3D human motion, Nymeria provides groundtruth motion annotations, while for EgoExo4D, we use the 3D human motion fits provided by UniEgoMotion [49] as groundtruth. For our testing, we sample video clips uniformly at random from both datasets. To further assess performance in challenging locomotion scenarios, we additionally include segments where groundtruth camera poses exhibit high movement. We report metrics on this combined test set in the main paper, while separate metrics on the uniform and locomotion-heavy subsets are provided in the Appendix.

4.2 3D Reconstruction & Camera Pose Estimation Results

Baselines. We compare EgoMo3R's 3D reconstruction and camera pose estimation against several competitive baselines. VGGT [71] and CUT3R [74] represent purely data-driven, general feedforward reconstruction methods trained on large-scale 3D data to directly regress camera poses and pointmaps. EgoMono4D [82] is a similar feedforward model trained on egocentric datasets. ViPE [23] (similar to MegaSaM [35]) is a modern bundle-adjustment-based SLAM system with loop closure that optimizes 3D points as free variables while leveraging robust video depth, point tracks, and segmentation priors. EgoEgo [32] is the only baseline in the egocentric motion estimation literature that does not assume access to a canonical trajectory. Instead, it employs a dedicated camera pose estimation module on top of an off-the-shelf SLAM system [67] and optical flow features. We train it on EgoExo4D

Table 1: **3D Reconstruction**: Comparison across metric and scale-invariant 3D reconstruction. Lower values (↓) indicate better performance. Our method outperforms all baselines by a significant margin.

		Metric Scale				Scale Invariant			
		Acc		Comp		Acc		Comp	
	Method	Mean	Median	Mean	Median	Mean	Median	Mean	Median
EgoExo4D	VGGT [71]	–	–	–	–	0.470	0.418	0.440	0.226
	CUT3R [74]	0.433	0.397	0.665	0.506	0.410	0.350	0.341	0.226
	ViPE [23]	0.905	0.657	0.819	0.573	0.425	0.366	0.352	0.194
	EgoMono4D [82]	0.619	0.544	0.707	0.546	0.469	0.423	0.593	0.413
	Ours	**0.357**	**0.287**	**0.187**	**0.124**	**0.309**	**0.245**	**0.158**	**0.098**
Nymeria	VGGT [71]	–	–	–	–	0.337	0.290	0.309	0.151
	CUT3R [74]	0.346	0.298	0.362	0.229	0.338	0.286	0.293	0.180
	ViPE [23]	1.742	1.520	0.762	0.566	0.345	0.298	0.271	0.141
	EgoMono4D [82]	0.489	0.417	0.556	0.390	0.395	0.355	0.544	0.375
	Ours	**0.247**	**0.189**	**0.131**	**0.073**	**0.223**	**0.171**	**0.118**	**0.065**

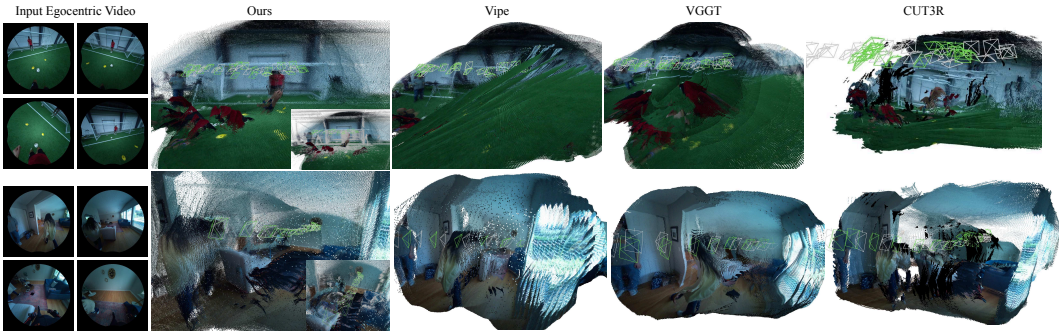

Figure 3: **3D Reconstruction**: We visualize the predicted pointmaps and camera trajectories (green) against the groundtruth trajectories (grey). The inset shows our reconstructed scene overlaid with sparse groundtruth points (grey), demonstrating the metric accuracy and alignment of EgoMo3R. Please refer to our supplementary video for a better qualitative comparison.

training split and evaluate on both EgoExo4D and Nymeria to test its generalization capabilities. Notably, all other methods, including EgoMo3R, are evaluated without dataset-specific training.

Metrics. For 3D reconstruction, we report standard Accuracy and Completion distances where lower values indicate superior performance. To evaluate camera pose estimation, we report Absolute Trajectory Error (ATE), which assesses the global accuracy after alignment with the ground truth. We also report Relative Pose Error (RPE) for both rotation and translation to evaluate local consistency. Since VGGT outputs are scale-agnostic while other methods (including ours) operate in metric scale, we report all metrics following both scale-preserving and scale-invariant alignment.

Results. Tables 1 and 2 present the results for 3D reconstruction and camera pose estimation, respectively. EgoMo3R outperforms all baselines across all 3D reconstruction metrics by a significant margin. In camera pose estimation, our method performs on par with ViPE and outperforms all other baselines. Feedforward models such as VGGT and CUT3R struggle with the egocentric domain gap, where highly dynamic camera movements and significant motion blur lead to substantial errors in both reconstruction and pose estimation. EgoMono4D also performs poorly due to its reliance on only optical flow for correspondences. ViPE leverages low-level visuo-geometric priors including a video depth model, and uses classic loop closure to maintain accurate camera tracking, especially on texture-sparse Nymeria sequences. However, it produces overly smooth geometry, resulting in significantly worse 3D reconstruction metrics compared to EgoMo3R. Finally, EgoEgo's head-pose module performs poorly on our evaluation set. Its flow-based parameterization struggles with challenging head movements, particularly on the Nymeria sequences for which it was not trained. Overall, EgoMo3R demonstrates strong performance and robustness, achieving state-of-the-art results in both camera pose accuracy and 3D scene fidelity.

Table 2: **Camera Pose Estimation**: Comparison across metric and scale-invariant camera pose estimation. Lower values (↓) indicate better performance.

		Metric Scale			Scale Invariant		
	Method	ATE	RPE Rot	RPE Trans	ATE	RPE Rot	RPE Trans
EgoExo4D	EgoEgo [32]	0.421	5.023	0.089	0.275	5.023	0.067
	VGGT [71]	-	-	-	0.251	3.744	0.078
	CUT3R [74]	0.415	4.101	0.118	0.294	4.101	0.076
	ViPE [23]	0.426	**2.568**	0.137	**0.112**	**2.568**	**0.032**
	EgoMono4D [82]	1.351	4.445	0.347	0.345	4.445	0.077
	Ours	**0.226**	2.640	**0.065**	0.161	2.640	0.047
Nymeria	EgoEgo [32]	0.629	3.129	0.096	0.390	3.129	0.087
	VGGT [71]	-	-	-	0.275	2.956	0.107
	CUT3R [74]	0.471	2.801	0.119	0.341	2.801	0.093
	ViPE [23]	**0.276**	2.168	0.085	**0.122**	2.168	**0.039**
	EgoMono4D [82]	1.318	3.881	0.322	0.435	3.881	0.115
	Ours	0.325	**2.142**	**0.060**	0.239	**2.142**	0.065

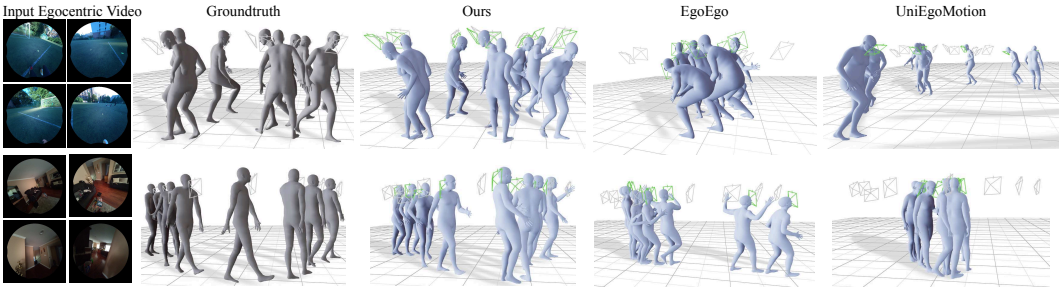

Figure 4: **Qualitative comparison of 3D human motion reconstruction.** EgoMo3R effectively reconstructs physically grounded human motion that globally aligns well with the groundtruth. In contrast, errors in EgoEgo's head trajectory estimation lead to significant global pose errors. While UniEgoMotion captures plausible local body poses, it similarly lacks the precise head tracking required for global accuracy. Please refer to our suppl. video for a better qualitative comparison.

4.3 Human Motion Reconstruction Results

Baselines. We evaluate human motion reconstruction of EgoMo3R against baseline methods that don't rely on a metric gravity-aligned camera trajectory. For EgoEgo [32], its head pose estimator is trained on optical flow features from the EgoExo4D dataset (see Sec. 4.2), while the gravity direction predictor is trained on AMASS [44]. For UniEgoMotion [49] and EgoAllo [80], we obtain a metric camera trajectory from CUT3R or ViPE and estimate the floor plane using their floor-segmented [56] pointmaps. Since floor estimation depends on reconstruction quality, CUT3R offers a balance of pose and geometry accuracy, whereas ViPE provides more accurate poses but weaker geometry. As an oracle upper bound, we also condition both methods on groundtruth trajectories. Given UniEgoMotion's support for flexible inputs, we also train its variant on EgoExo4D that directly reconstructs human motion from input video without requiring trajectory input. Both EgoEgo and UniEgoMotion baselines are trained on the EgoExo4D training set and evaluated on both EgoExo4D and Nymeria sequences to analyze their generalization capacity.

Metrics. We evaluate methods on their ability to track head motion using head rotation and head translation error [49, 32]. MPJPE calculates mean per-joint position error (in meters) after globally aligning predicted body joints with groundtruth body joints. MPJPE-PA is calculated after separately aligning each frame's motion and only measures the local human pose accuracy regardless of its global alignment. FC (Foot Contact) [49] quantifies the plausibility of foot-ground contact.

Results. Tab. 3 presents the human motion reconstruction results. EgoMo3R outperforms the baselines on both head tracking and global pose metrics. UEM-Img-only struggles to generate accurate global motion because it attempts to jointly estimate camera trajectories and body poses within a single model without explicit geometric constraints. As discussed in Sec. 4.2, EgoEgo's

Table 3: **Human Motion Reconstruction:** Lower values (↓) indicate better performance. Our method significantly outperforms baselines on head tracking and global motion metrics, while baselines retain local pose accuracy at the cost of significantly worse alignment between the motion and the scene.

	Method	Head Rot	Head Trans	MPJPE	MPJPE-PA	FC
EgoExo4D	EgoEgo [32]	0.574	0.449	0.441	0.143	0.046
	CUT3R + EgoAllo [80]	0.278	0.431	0.427	0.167	**0.035**
	CUT3R + UEM [49]	0.526	0.456	0.407	0.151	0.039
	ViPE + EgoAllo [80]	0.203	0.411	0.348	0.155	0.092
	ViPE + UEM [49]	0.531	0.396	0.337	0.145	0.088
	UEM-Img-only [49]	0.519	0.334	0.309	**0.073**	0.037
	Ours	**0.188**	**0.219**	**0.271**	0.127	0.053
	GT Traj. + EgoAllo	0.119	0.051	0.174	0.102	0.033
	GT Traj. + UEM	0.203	0.042	0.127	0.067	0.029
Nymeria	EgoEgo [32]	0.538	0.684	0.600	0.160	0.046
	CUT3R + EgoAllo [80]	0.232	0.488	0.437	0.178	**0.033**
	CUT3R + UEM [49]	0.448	0.503	0.442	0.156	0.035
	ViPE + EgoAllo [80]	0.222	0.344	0.341	0.158	0.084
	ViPE + UEM [49]	0.368	0.365	0.348	0.157	0.071
	UEM-Img-only [49]	0.806	0.769	0.696	**0.116**	0.038
	Ours	**0.180**	**0.318**	**0.319**	0.133	0.036
	GT Traj. + EgoAllo	0.115	0.070	0.149	0.108	0.031
	GT Traj. + UEM	0.170	0.068	0.140	0.103	0.030

Table 4: **Ablation Study**: Metrics are reported on EgoExo4D locomotion subset with metric scale alignment. Lower values (↓) are better. Rows 1–4 optimize only $(\boldsymbol{C}, \boldsymbol{S})$. Rows 5–6 add the motion prior, optimized directly or in an alternating manner (ours, gray). Row 7 replaces the motion prior with a camera-smoothness prior.

	ATE	RPE Trans	Acc		Comp		MPJPE
			Mean	Median	Mean	Median	
$\mathcal{L}_{\text{flow}}$	0.604	0.222	0.426	0.352	0.207	0.122	–
$+\,\mathcal{L}_{\text{track}}$	0.361	0.111	0.418	0.342	0.213	0.139	–
$+\,\mathcal{L}_{\text{depth}}$	0.353	0.103	0.414	0.337	0.210	0.135	–
$+\,\mathcal{L}_{\text{scene}}$	0.371	0.120	0.398	0.319	0.211	0.131	–
$+\,\mathcal{L}_{\text{motion}} + \mathcal{L}_{\text{egocam}}$ (direct)	0.350	0.096	0.403	0.328	0.214	0.139	0.456
+ Alternating Optim. (ours)	0.318	0.095	0.396	0.320	0.209	0.137	0.320
Row 4 $+\,\mathcal{L}_{\text{smooth}}$	0.335	0.074	0.429	0.351	0.223	0.142	–

head pose estimation pipeline struggles on challenging in-the-wild egocentric sequences. Similarly, EgoAllo and UEM depend on imperfect camera trajectories provided by CUT3R/ViPE, causing errors to propagate into downstream motion estimation. With groundtruth trajectories, both EgoAllo and UEM improve substantially, indicating that a major portion of the motion error is induced by trajectory error. In contrast, EgoMo3R's joint optimization establishes a feedback loop that iteratively refines both camera poses and human motion. The resulting more accurate head trajectory translates into lower global motion error, even under loose $\mathrm{SE}(3)$-invariant conditioning. Interestingly, the baselines achieve competitive local pose accuracy and foot contact scores because their motion diffusion models strictly follow the potentially inaccurate input *canonical* trajectory to generate plausible human motion.

4.4 Ablation

Tab. 4 isolates each loss term and the role of the motion prior. Without the motion prior, there is no motion stage to alternate with, so rows 1–4 are the scene-only variants of our method. Optimizing $\mathcal{L}_{\text{flow}}$ alone leads to drift, which $\mathcal{L}_{\text{track}}$ largely corrects through long-range correspondences (ATE $0.604 \rightarrow 0.361$). $\mathcal{L}_{\text{depth}}$ further regularizes the depth scales $\boldsymbol{S}$. $\mathcal{L}_{\text{scene}}$ enforces floor coplanarity, which improves reconstruction accuracy but raises ATE. Adding the motion prior with direct joint optimization (row 5) improves the camera trajectory but often causes the motion to collapse (MPJPE 0.456). Alternating between the $(\boldsymbol{C}, \boldsymbol{S})$ and $\boldsymbol{M}$ stages stabilizes the optimization and gives the

Table 5: **Progression of Alternating Optimization**: Metrics are reported on EgoExo4D locomotion subset with metric scale alignment. Lower values (↓) are better. Compared to direct optimization, alternation prevents the motion $\boldsymbol{M}$ from collapsing to a suboptimal solution while also improving camera pose and scene reconstruction. We report results after two rounds throughout the paper (gray rows). A third round further improves pose, accuracy, and MPJPE.

			Acc		Comp		
	ATE	RPE Trans	Mean	Median	Mean	Median	MPJPE
Direct optimization	0.350	0.096	0.403	0.328	0.214	0.139	0.456
Alternating optimization							
Initialization	0.604	0.222	0.426	0.352	0.207	0.122	–
R1. Optimize $(\boldsymbol{C}, \boldsymbol{S})$	0.353	0.102	0.410	0.333	0.212	0.137	–
R1. Optimize $(\boldsymbol{M})$	–	–	–	–	–	–	0.347
R2. Optimize $(\boldsymbol{C}, \boldsymbol{S})$	0.318	0.095	0.396	0.320	0.209	0.137	–
R2. Optimize $(\boldsymbol{M})$	–	–	–	–	–	–	0.320
R3. Optimize $(\boldsymbol{C}, \boldsymbol{S})$	0.300	0.085	0.391	0.314	0.209	0.137	–
R3. Optimize $(\boldsymbol{M})$	–	–	–	–	–	–	0.315

lowest ATE, Acc Mean, and MPJPE (row 6). Replacing the motion prior with a camera-smoothness prior $\mathcal{L}_{\text{smooth}}$ (row 7) suppresses camera jumps and gives the lowest RPE Trans, but it is worse than our full model on all other metrics and produces no body motion.

Tab. 5 shows the steady progression of the alternating optimization and that the coupling works in both directions. *Motion→scene:* the first motion stage lowers ATE from 0.353 after the R1 scene stage to 0.318 after R2, as the prior anchors the trajectory in a gravity-aligned, grounded frame. *Scene→motion:* MPJPE decreases monotonically as $(\boldsymbol{C}, \boldsymbol{S})$ is refined across rounds ($0.347 \rightarrow 0.320 \rightarrow 0.315$), as a better trajectory gives better conditioning to the motion prior. A third round further improves pose, accuracy, and MPJPE.

5 Conclusion & Future Work

We present EgoMo3R, an optimization framework that jointly reconstructs egocentric human motion, camera trajectory, and metric, gravity-aligned scene geometry from casually captured monocular egocentric video. By combining low-level visuo-geometric cues with a diffusion-based motion prior conditioned on an $\mathrm{SE}(3)$-invariant representation of the camera trajectory, our approach exploits the inherent coupling between camera trajectory, scene geometry, and body pose in the egocentric setting. EgoMo3R outperforms specialized baselines on the EgoExo4D and Nymeria datasets, and our ablations show that a learned motion prior outperforms a simple smoothness penalty. This suggests that learned human motion distributions are a promising direction for egocentric camera and scene reconstruction. Looking ahead, we plan to explore motion priors that, unlike diffusion models, allow efficient likelihood estimation during optimization, such as VAE-based formulations [57]. Future work could also investigate feedforward models that jointly reconstruct scene and motion in a single pass, for which EgoMo3R reconstructions could serve as pseudo-labels. We believe that EgoMo3R provides a strong foundation for future research on joint modeling of egocentric motion and 3D scenes.

Acknowledgment This research was conducted at Stanford University with support from Panasonic Holdings Corporation and partial funding from NIH grant R01AG089169.

References

[1] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski. Building rome in a day. *Communications of the ACM*, 54(10):105–112, 2011.

[2] O. Bar-Tal, L. Yariv, Y. Lipman, and T. Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113*, 2023.

[3] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European conference on computer vision*, pages 561–578. Springer, 2016.

[4] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on robotics*, 32(6):1309–1332, 2017.

[5] L. Carlone, R. Tron, K. Daniilidis, and F. Dellaert. Initialization techniques for 3d slam: A survey on rotation estimation and its use in pose graph optimization. In *2015 IEEE international conference on robotics and automation (ICRA)*, pages 4597–4604. IEEE, 2015.

[6] A. Castillo, M. Escobar, G. Jeanneret, A. Pumarola, P. Arbeláez, A. Thabet, and A. Sanakoyeu. Bodiffusion: Diffusing sparse observations for full-body human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4221–4231, 2023.

[7] S. Chen, H. Guo, S. Zhu, F. Zhang, Z. Huang, J. Feng, and B. Kang. Video depth anything: Consistent depth estimation for super-long videos. *arXiv:2501.12375*, 2025.

[8] Y. Chen, X. Chen, Y. Xue, A. Chen, Y. Xiu, and G. Pons-Moll. Human3r: Everyone everywhere all at once. *arXiv preprint arXiv:2510.06219*, 2025.

[9] K. Cho and H. Joo. Hand-aware egocentric motion reconstruction with sequence-level context. *arXiv preprint arXiv:2512.19283*, 2025.

[10] H. Chung, J. Kim, M. T. Mccann, M. L. Klasky, and J. C. Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

[11] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse. Monoslam: Real-time single camera slam. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1052–1067, 2007.

[12] D. DeTone, T. Malisiewicz, and A. Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018.

[13] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[14] C. Doersch, Y. Yang, M. Vecerik, D. Gokay, A. Gupta, Y. Aytar, J. Carreira, and A. Zisserman. Tapir: Tracking any point with per-frame initialization and temporal refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10061–10072, 2023.

[15] J. Engel, K. Somasundaram, M. Goesele, A. Sun, A. Gamino, A. Turner, A. Talattof, A. Yuan, B. Souti, B. Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023.

[16] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[17] J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

[18] K. Grauman, A. Westbury, L. Torresani, K. Kitani, J. Malik, T. Afouras, K. Ashutosh, V. Baiyya, S. Bansal, B. Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024.

[19] V. Guzov, Y. Jiang, F. Hong, G. Pons-Moll, R. Newcombe, C. K. Liu, Y. Ye, and L. Ma. Hmd 2: Environment-aware motion generation from single egocentric head-mounted device. In *2025 International Conference on 3D Vision (3DV)*, pages 1394–1405. IEEE, 2025.

[20] F. Hong, V. Guzov, H. J. Kim, Y. Ye, R. Newcombe, Z. Liu, and L. Ma. Egolm: Multi-modal language model of egocentric motions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5344–5354, 2025.

[21] M. Hu, W. Yin, C. Zhang, Z. Cai, X. Long, H. Chen, K. Wang, G. Yu, C. Shen, and S. Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10579–10596, 2024.

[22] V. T. Hu, D. W. Zhang, Y. M. Asano, G. J. Burghouts, and C. G. Snoek. Self-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18413–18422, 2023.

[23] J. Huang, Q. Zhou, H. Rabeti, A. Korovko, H. Ling, X. Ren, T. Shen, J. Gao, D. Slepichev, C.-H. Lin, et al. Vipe: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934*, 2025.

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

[25] N. Karaev, I. Rocco, B. Graham, N. Neverova, A. Vedaldi, and C. Rupprecht. Cotracker: It is better to track together. In *European conference on computer vision*, pages 18–35. Springer, 2024.

[26] N. Karaev, Y. Makarov, J. Wang, N. Neverova, A. Vedaldi, and C. Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6013–6022, 2025.

[27] K. Karunratanakul, K. Preechakul, E. Aksan, T. Beeler, S. Suwajanakorn, and S. Tang. Optimizing diffusion noise can serve as universal motion priors. In *arxiv:2312.11994*, 2023.

[28] K. Karunratanakul, K. Preechakul, S. Suwajanakorn, and S. Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2151–2162, 2023.

[29] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[30] V. Leroy, Y. Cabon, and J. Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pages 71–91. Springer, 2024.

[31] G. Li, Y. Chen, Y. Wu, K. Zhao, M. Pollefeys, and S. Tang. Egom2p: Egocentric multimodal multitask pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10830–10843, 2025.

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

[33] T. Li, D. Katabi, and K. He. Return of unconditional generation: A self-supervised representation generation method. *Advances in Neural Information Processing Systems*, 37:125441–125468, 2024.

[34] W. Li, Y. Xu, M. Yao, F. Liang, J. Sun, M. Wang, G. Zhang, L. Huang, and H. Li. Noperoomgs: Indoor 3d gaussian splatting optimization without camera pose input. *Advances in Neural Information Processing Systems*, 38:49124–49144, 2026.

[35] Z. Li, R. Tucker, F. Cole, Q. Wang, L. Jin, V. Ye, A. Kanazawa, A. Holynski, and N. Snavely. Megasam: Accurate, fast and robust structure and motion from casual dynamic videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10486–10496, 2025.

[36] G. Liu, H. Sun, J. Li, F. Yin, and Y. Yang. Accelerating diffusion models for inverse problems through shortcut sampling. *arXiv preprint arXiv:2305.16965*, 2023.

[37] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024.

[38] Z. Liu, J. Lin, W. Wu, and B. Zhou. Joint optimization for 4d human-scene reconstruction in the wild. *arXiv preprint arXiv:2501.02158*, 2025.

[39] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[40] J. Lu, J. Lin, H. Dou, A. Zeng, Y. Deng, Y. Zhang, and H. Wang. Dposer: Diffusion model as robust 3d human pose prior. *arXiv preprint arXiv:2312.05541*, 2023.

[41] Z. Luo, R. Hachiuma, Y. Yuan, and K. Kitani. Dynamics-regulated kinematic policy for egocentric pose estimation. *Advances in Neural Information Processing Systems*, 34:25019–25032, 2021.

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

[43] W.-C. Ma, A. J. Yang, S. Wang, R. Urtasun, and A. Torralba. Virtual correspondence: Humans as a cue for extreme-view geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15924–15934, 2022.

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

[45] M. Mardani, J. Song, J. Kautz, and A. Vahdat. A variational perspective on solving inverse problems with diffusion models. *arXiv preprint arXiv:2305.04391*, 2023.

[46] L. Müller, H. Choi, A. Zhang, B. Yi, J. Malik, and A. Kanazawa. Reconstructing people, places, and cameras. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21948–21958. IEEE, 2025.

[47] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos. Orb-slam: A versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

[48] L. Pan, D. Barath, M. Pollefeys, and J. L. Schönberger. Global Structure-from-Motion Revisited. In *European Conference on Computer Vision (ECCV)*, 2024.

[49] C. Patel, H. Nakamura, Y. Kyuragi, K. Kozuka, J. C. Niebles, and E. Adeli. Uniegomotion: A unified model for egocentric motion reconstruction, forecasting, and generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10318–10329, 2025.

[50] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019.

[51] G. Pavlakos, E. Weber, M. Tancik, and A. Kanazawa. The one where they reconstructed 3d humans and environments in tv shows. In *European Conference on Computer Vision*, pages 732–749. Springer, 2022.

[52] X. Peng, Z. Zheng, W. Dai, N. Xiao, C. Li, J. Zou, and H. Xiong. Improving diffusion models for inverse problems using optimal posterior covariance. *arXiv preprint arXiv:2402.02149*, 2024.

[53] L. Piccinelli, C. Sakaridis, Y.-H. Yang, M. Segu, S. Li, W. Abbeloos, and L. Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[54] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022.

[55] S. Raab, I. Leibovitch, P. Li, K. Aberman, O. Sorkine-Hornung, and D. Cohen-Or. Modi: Unconditional motion synthesis from diverse data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13873–13883, 2023.

[56] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[57] D. Rempe, T. Birdal, A. Hertzmann, J. Yang, S. Sridhar, and L. J. Guibas. Humor: 3d human motion model for robust pose estimation. In *International Conference on Computer Vision (ICCV)*, 2021.

[58] S. Rojas, M. Armando, B. Ghanem, P. Weinzaepfel, V. Leroy, and G. Rogez. Hamst3r: Human-aware multi-view stereo 3d reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5027–5037, 2025.

[59] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011.

[60] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020.

[61] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[62] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

[63] Y. Shafir, G. Tevet, R. Kapon, and A. H. Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023.

[64] C. Smith, D. Charatan, A. Tewari, and V. Sitzmann. Flowmap: High-quality camera poses, intrinsics, and depth via gradient descent. In *2025 International Conference on 3D Vision (3DV)*, pages 389–400. IEEE, 2025.

[65] K. Tateno, F. Tombari, I. Laina, and N. Navab. Cnn-slam: Real-time dense monocular slam with learned depth prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6243–6252, 2017.

[66] Z. Teed and J. Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020.

[67] Z. Teed and J. Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021.

[68] Z. Teed, L. Lipson, and J. Deng. Deep patch visual odometry. *Advances in Neural Information Processing Systems*, 36:39033–39051, 2023.

[69] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-or, and A. H. Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=SJ1kSyO2jwu`.

[70] J. Wang, N. Karaev, C. Rupprecht, and D. Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21686–21697, 2024.

[71] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025.

[72] J. Wang, R. Dabral, D. Luvizon, Z. Cao, L. Liu, T. Beeler, and C. Theobalt. Ego4o: Egocentric human motion capture and understanding from multi-modal input. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22668–22679, 2025.

[73] Q. Wang, V. Ye, H. Gao, W. Zeng, J. Austin, Z. Li, and A. Kanazawa. Shape of motion: 4d reconstruction from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9660–9672, 2025.

[74] Q. Wang, Y. Zhang, A. Holynski, A. A. Efros, and A. Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10510–10522, 2025.

[75] R. Wang, S. Xu, C. Dai, J. Xiang, Y. Deng, X. Tong, and J. Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5261–5271, 2025.

[76] R. Wang, S. Xu, Y. Dong, Y. Deng, J. Xiang, Z. Lv, G. Sun, X. Tong, and J. Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *arXiv preprint arXiv:2507.02546*, 2025.

[77] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20697–20709, 2024.

[78] Y. Xu, Y.-J. Li, X. Weng, and K. Kitani. Wide-baseline multi-camera calibration using person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13134–13143, 2021.

[79] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

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

[81] H. Yu, W. Liu, J. Bai, X. Gui, Y. Hou, Y. Ong, and Q. Zhang. Towards efficient and diverse generative model for unconditional human motion synthesis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2535–2544, 2024.

[82] C. Yuan, G. Chen, L. Yi, and Y. Gao. Self-supervised monocular 4d scene reconstruction for egocentric videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8863–8874, 2025.

[83] Y. Yuan and K. Kitani. 3d ego-pose estimation via imitation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 735–750, 2018.

[84] Y. Yuan and K. Kitani. Ego-pose estimation and forecasting as real-time pd control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10082–10092, 2019.

[85] D. Zhang, G. Li, J. Li, M. Bressieux, O. Hilliges, M. Pollefeys, L. Van Gool, and X. Wang. Egogaussian: Dynamic scene understanding from egocentric video with 3d gaussian splatting. In *2025 International Conference on 3D Vision (3DV)*, pages 1091–1102. IEEE, 2025.

[86] J. Zhang, C. Herrmann, J. Hur, V. Jampani, T. Darrell, F. Cole, D. Sun, and M.-H. Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024.

[87] W. Zhao, S. Liu, H. Guo, W. Wang, and Y.-J. Liu. Particlesfm: Exploiting dense point trajectories for localizing moving cameras in the wild. In *European Conference on Computer Vision*, pages 523–542. Springer, 2022.

[88] Y. Zhao, T. Y. Wang, B. Raj, M. Xu, J. Yang, and C.-H. P. Huang. Synergistic global-space camera and human reconstruction from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1216–1226, 2024.

[89] R. Zirvi, B. Tolooshams, and A. Anandkumar. Diffusion state-guided projected gradient for inverse problems. *arXiv preprint arXiv:2410.03463*, 2024.

A Extended Evaluation and Results

Given that large-scale EgoExo4D and Nymeria datasets consist of egocentric sequences ranging from several minutes to over an hour, we extract 8-second segments for evaluation (following UniEgoMotion [49]). To ensure a comprehensive analysis, our evaluation set contains two distinct subsets: (1) a representative subset sampled uniformly at random from the full dataset, and (2) a locomotion-heavy subset containing high camera translation and rotation. While the main paper reports results on the combined set containing 718 samples from EgoExo4D and 828 samples from Nymeria, we provide disaggregated results for each subset here. Additionally, we also evaluate a subset characterized by rotation-heavy camera movement which is challenging for visual SLAM methods [5]. The disaggregated results are provided in Table 6 and 7. We further evaluate per-frame depth accuracy of monocular and feedforward geometry models (Table 8), sensitivity to the choice of off-the-shelf priors (Table 9), and performance on $3\times$ longer sequences (Table 10). Please refer to the supplementary video for additional qualitative results and comparisons with the baselines.

B Implementation Details

We use RAFT [66] to compute both forward and backward optical flow, incorporating both into the optimization objective. Long-range correspondences are obtained using CoTracker3 [26] by seeding a 35×35 point grid at every 8-frame intervals and tracking them bi-directionally. Initial relative depth $\boldsymbol{D}^r$ and the metric scale prior $\boldsymbol{S}^m$ are provided by MoGe-2 models [76]. For the optimization objective, we set the loss weights to $1.0, 1.0, 5 \times 10^{-3}, 5 \times 10^{-3}$, and 5 for $\mathcal{L}_{\text{flow}}$, $\mathcal{L}_{\text{track}}$, $\mathcal{L}_{\text{scene}}$, $\mathcal{L}_{\text{depth}}$, and $\mathcal{L}_{\text{egocam}}$, respectively. For the motion prior $\mathcal{L}_{\text{motion}}$, we sample the noise level t in decreasing order following [40] and scale the loss weight according to the signal-to-noise ratio of the noised sample [54]. To handle outliers in $\mathcal{L}_{\text{flow}}$ and $\mathcal{L}_{\text{track}}$, we filter correspondences based on the Median Absolute Deviation (MAD) of the reprojection error. During each round of alternating optimization, we optimize $(\boldsymbol{C}, \boldsymbol{S})$ for 1000 steps, followed by 100 steps of optimization for $\boldsymbol{M}$. Note that when $(\boldsymbol{C}, \boldsymbol{S})$ is fixed, optimizing the motion for T steps is equivalent to DDIM sampling of $\boldsymbol{M}$ for T diffusion steps under the guidance of $\mathcal{L}_{\text{egocam}}$. We use the Adam optimizer with an initial learning rate of 0.01, decaying to 0.001 after 850 steps.

EgoMo3R takes approximately 8 min per 8-second clip on one NVIDIA L40S GPU: about 32 s to compute the visuo-geometric priors, 3 min and 4 min for scene optimization in the first and second rounds, and 5 s per round for motion optimization. This is slower than the feedforward baselines (a few seconds) and the optimization-based ViPE (about 1.5 min). However, after one round of optimization, at less than half the total cost, EgoMo3R already outperforms all baselines on metric 3D reconstruction on the EgoExo4D locomotion subset (Tabs. 5 and 6).

Our motion prior $p(\boldsymbol{M}|\mathcal{A}(\boldsymbol{C}))$ is implemented as a conditional denoising diffusion model $f_\phi(\boldsymbol{M}_t, \mathcal{A}(\boldsymbol{C}))$. Following [49], we adopt an egocentric body motion representation that implicitly defines a canonical reference frame with a ground plane and gravity direction. The denoiser is implemented as a 1D UNet [28], though we observed that transformer-based architecture [69] yields comparable performance. To evaluate the zero-shot motion generalization of EgoMo3R, we do not utilize motion annotations from EgoExo4D or Nymeria during training; and instead train exclusively on motion sequences from the AMASS dataset [44].

C More Discussion on Motion Prior

Our motion prior design (see Section 3.3) is motivated by the broader literature on inverse problems, guided sampling, and conditional generative models. Recall that our objective is to reconstruct body motion $\boldsymbol{M}$, camera trajectory $\boldsymbol{C}$, and depth scales $\boldsymbol{S}$ that are consistent with the observations provided by the egocentric video $\boldsymbol{I}$. Since the camera trajectory $\boldsymbol{C}$ is tightly coupled with the head trajectory $\mathcal{H}(\boldsymbol{M})$ derived from the body motion (enforced via $\mathcal{L}_{\text{egocam}}$), we regularize the optimization of $\boldsymbol{M}$ (and consequently $\boldsymbol{C}$) using a prior on $\boldsymbol{M}$. This formulation can be viewed as an inverse problem that aims to recover an unknown signal x from noisy and incomplete observations $y = \mathcal{O}(x) + \epsilon$, where a data term $\mathcal{L}_{\text{data}}(\mathcal{O}(x), y)$ is minimized alongside a prior term $\mathcal{L}_{\text{prior}}(x)$. In our context, we optimize $\boldsymbol{M}$ using data terms ($\mathcal{L}_{\text{flow}}$, $\mathcal{L}_{\text{track}}$, $\mathcal{L}_{\text{egocam}}$, etc.) regularized by the motion prior $p(\boldsymbol{M}|\cdot)$. Notably, the observations provide direct constraints only on a small subset of the motion $\boldsymbol{M}$ (i.e., the head trajectory), leaving the rest of the body pose highly dependent on the prior.

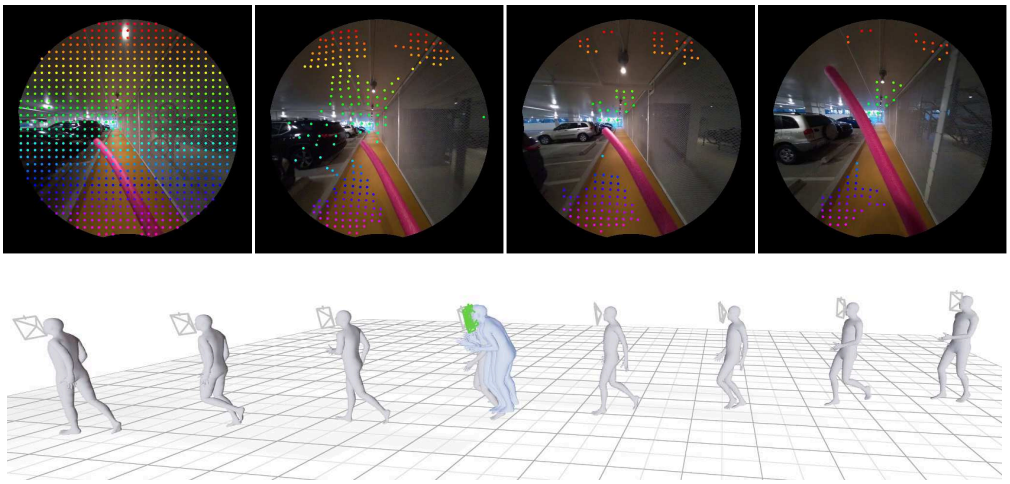

Figure 5: **Point-tracking Failure**: Inaccurate long-range correspondences from the point-tracking model cause EgoMo3R to predict stationary camera trajectory (green) and stationary motion (blue), whereas the groundtruth camera and motion trajectories (grey) move in a straight line.

A straightforward approach to solving such inverse problems is to penalize the negative log-likelihood of x during optimization ($\mathcal{L}_{\text{prior}}(x) = -\log p(x)$). Early works utilized simple pose priors modeled via Gaussian Mixture Models (GMMs) for single-frame tasks, such as fitting a 3D body model to 2D images [3, 50]. While GMMs allow for exact likelihood calculation and efficient sampling, their modeling capacity is limited for high-dimensional, complex distributions like human motion. Recently, some works have used VAE-based [50, 57] or diffusion-based [40] priors for their superior expressive power. However, since these models do not provide an explicit likelihood, incorporating them into a first-order optimization is non-trivial. Consider a diffusion model that predicts $\mathbb{E}[x_0|x_t]$, where the subscript denotes the noise level. One way to solve an inverse problem with such a model is to apply guidance [13] during each denoising step:

$$\mathbb{E}[x_0|x_t, y] = \mathbb{E}[x_0|x_t] + \lambda \nabla_{x_t} \log p(y|x_t)$$

where the second term provides guidance toward the observation y. [13] trained a separate classifier $p(y|x_t)$ for all noise levels to provide exact guidance. However, in many inverse problems, defining the guidance likelihood $p(y|x_t)$ on noisy data x_t (rather than clean data x_0) is analytically intractable, as it requires marginalizing over x_0: $p(y|x_t) = \int p(y|x_0)p(x_0|x_t)dx_0$. While various approximations have been proposed [36, 10, 52, 89, 45] that work well for simple pose or motion fitting with full-body 2D observations [57, 3, 50, 40], these approximations often fail when observations y are weak and highly localized, such as controlling only a specific body part trajectory [63, 28, 27], leading to suboptimal or physically implausible results.

Instead, recent works [63, 28] directly model the *conditional* distribution $p(x|y)$. This not only circumvents the intractability of the guidance term but also simplifies the learning task, as conditional generation is generally easier to learn than unconditional generation [33]. For egocentric reconstruction, prior methods [80, 49, 32, 19] typically assume access to a gravity-aligned, metric camera trajectory $\boldsymbol{C}$ and model $p(\boldsymbol{M}|\boldsymbol{C})$ using conditional diffusion. This approach is highly effective because the head trajectory strongly constrains the distribution of the remaining body joints, especially torso and lower body.

In our setting, however, we are also optimizing for $\boldsymbol{C}$. One possible solution is to use an unconditional motion prior $p(\boldsymbol{M})$ and perform guided optimization using correspondence losses. But, as previously noted, this often leads to suboptimal results due to approximations. Instead, we use a conditional diffusion prior $p(\boldsymbol{M}|\mathcal{A}(\boldsymbol{C}))$, where the conditioning is based on the SE(3)-invariant trajectory $\mathcal{A}(\boldsymbol{C})$. This allows us to incorporate the motion prior early in the optimization process. Following [54], we do not backpropagate gradients through the prior model itself to maintain optimization stability.

While this design provides effective regularization for $\boldsymbol{C}$ and improves 3D reconstruction (see Table 4, row 5), direct optimization often causes the motion to collapse to a solution that satisfies the trajectory $\boldsymbol{C}$ but disregards global physical plausibility, leading to worse motion metrics. We hypothesize this

is due to simultaneous optimization, where gradients on $\boldsymbol{M}$ are often noisy relative to the data-term gradients on $\boldsymbol{C}$ during early optimization steps. While it is possible to mitigate this by carefully scheduling the loss weights as $\boldsymbol{C}$ approaches the optimal solution, such tuning is often brittle. Instead, we propose a simple alternating scheme that iterates between optimizing $(\boldsymbol{C}, \boldsymbol{S})$ and $\boldsymbol{M}$. As shown in Table 4 (row 6) in the paper and Table 5, this strategy significantly improves both 3D reconstruction and motion reconstruction metrics.

D Limitations & Future Work

While EgoMo3R achieves superior 3D scene reconstruction and accurate human motion estimation, it is not without limitations. We discuss several of them here and outline potential future directions.

The core design of EgoMo3R relies on the robustness and generalization of low-level visuo-geometric models on challenging egocentric data. We observe that while optical flow generally provides stable short-range correspondences, long-range point tracking remains a primary source of failure. One mode of point-tracking failure arises in regions with monotonic textures or around handheld objects. These erroneous tracks can derail camera trajectory optimization, leading to poor local minima, as illustrated in Fig. 5. A secondary source of failure involves frames with extreme motion blur, where both the relative depth and the metric scale prior $\boldsymbol{S}^m$ can become unreliable. This can result in sudden jitters or jumps in the estimated camera poses $\boldsymbol{C}$. Although these local inaccuracies typically do not compromise the global trajectory, they can degrade the qualitative consistency and temporal smoothness of the resulting 3D reconstruction and motion $\boldsymbol{M}$. Another minor failure mode is the lack of floor detection in *all* frames. While this does not catastrophically degrade performance, it weakens the alignment between motion and 3D reconstruction, leaving the system to rely primarily on the egocentric camera trajectory.

Note that EgoMo3R focuses on reconstructing scene and body pose from egocentric locomotion video input. This is motivated by the fact that torso and lower-body motion are highly correlated with head motion. However, wrist motion and hand manipulation are largely unrelated to the rest of the body motion. Some prior works in egocentric motion reconstruction, such as EgoAllo [80] and HaMoS [9], use off-the-shelf 3D hand predictors and guide motion generation with accurate hand trajectories to reconstruct whole-body motion. However, our focus is primarily on locomotion without egocentric trajectory input, where body dynamics are more strongly coupled. Future work could explore leveraging 3D scene context to reconstruct globally accurate hand motion alongside whole-body motion, which remains highly challenging without canonical trajectory.

Note that, like any SLAM method, dynamic objects violate our correspondence assumptions. We use Median Absolute Deviation (MAD) filtering to mitigate the effect of bad correspondences, which often arise from dynamic objects. However, when a large number of pixels in the input video belong to dynamic objects, MAD filtering may also remove valid correspondences. Many prior methods mask out dynamic objects using off-the-shelf estimators [23] or learned confidence masks [86, 35, 82]. While such dynamic-object masking did not provide significant benefits in our case, it can be easily incorporated into our pipeline by prompting the segmentation model used in our work with additional queries. Additionally, EgoMo3R inference is relatively slow due to its optimization-based framework compared to feedforward geometry models. However, it provides more accurate reconstruction and can serve as a source of pseudo-labels to improve feedforward methods. Further improvements may also be achieved through explicit loop closure, akin to traditional SLAM pipelines.

Although the canonical camera trajectory $\boldsymbol{C}$ used in EgoAllo and UniEgoMotion contains more information than the $\mathrm{SE}(3)$-invariant trajectory $\mathcal{A}(\boldsymbol{C})$, our motion reconstruction still outperforms these baselines. We hypothesize that because canonical camera trajectory is highly informative of lower-body motion, any error in canonicalizing the predicted camera trajectory can significantly affect motion generation in the absence of a feedback loop. In contrast, EgoMo3R conditions motion reconstruction on looser conditioning, thereby providing greater flexibility. However, removing canonicalization also discards crucial information about body motion, forcing the model to rely more on guidance from $\mathcal{L}_{\text{scene}}$ and $\mathcal{L}_{\text{egocam}}$ (both in the canonical frame), which is often brittle (see Sec. C). Future work could explore more effective ways to incorporate canonicalization information into motion generation.

Table 6: **3D Reconstruction**: Comparison across metric and scale-invariant 3D reconstruction on three subsets (random, locomotion-heavy, rotation-heavy) of each dataset. Lower values (↓) indicate better performance.

			Metric Scale				Scale Invariant			
			Acc		Comp		Acc		Comp	
		Method	Mean	Median	Mean	Median	Mean	Median	Mean	Median
EgoExo4D	Random	VGGT [71]	–	–	–	–	0.432	0.382	0.425	0.231
		CUT3R [74]	0.402	0.372	0.613	0.480	0.385	0.332	0.314	0.214
		ViPE [23]	0.682	0.545	0.829	0.607	0.384	0.329	0.320	0.173
		EgoMono4D [82]	0.638	0.565	0.637	0.510	0.431	0.390	0.533	0.384
		Ours	**0.316**	**0.254**	**0.163**	**0.110**	**0.267**	**0.211**	**0.135**	**0.081**
	Locomotion	VGGT [71]	–	–	–	–	0.507	0.453	0.454	0.220
		CUT3R [74]	0.463	0.421	0.713	0.531	0.433	0.367	0.367	0.237
		ViPE [23]	1.120	0.765	0.808	0.540	0.464	0.402	0.382	0.214
		EgoMono4D [82]	0.599	0.524	0.773	0.580	0.506	0.455	0.651	0.440
		Ours	**0.396**	**0.320**	**0.209**	**0.137**	**0.349**	**0.278**	**0.180**	**0.115**
	Rotation	VGGT [71]	–	–	–	–	0.667	0.594	0.604	0.330
		CUT3R [74]	0.680	0.635	1.036	0.755	0.590	0.518	0.585	0.375
		ViPE [23]	0.708	0.648	0.922	0.628	0.626	0.557	0.608	0.388
		EgoMono4D [82]	0.790	0.738	1.224	0.916	0.705	0.653	1.015	0.687
		Ours	**0.638**	**0.552**	**0.377**	**0.261**	**0.566**	**0.485**	**0.359**	**0.247**
Nymeria	Random	VGGT [71]	–	–	–	–	0.320	0.276	0.303	0.155
		CUT3R [74]	0.337	0.289	0.369	0.242	0.321	0.270	0.293	0.183
		ViPE [23]	0.332	0.298	0.492	0.307	0.319	0.274	0.250	0.128
		EgoMono4D [82]	0.507	0.427	0.534	0.379	0.369	0.327	0.516	0.357
		Ours	**0.227**	**0.174**	**0.117**	**0.067**	**0.201**	**0.153**	**0.101**	**0.055**
	Locomotion	VGGT [71]	–	–	–	–	0.354	0.305	0.318	0.148
		CUT3R [74]	0.357	0.309	0.356	0.217	0.357	0.303	0.296	0.181
		ViPE [23]	3.118	2.712	1.028	0.820	0.373	0.323	0.296	0.155
		EgoMono4D [82]	0.476	0.410	0.587	0.408	0.423	0.386	0.578	0.397
		Ours	**0.269**	**0.206**	**0.146**	**0.079**	**0.248**	**0.191**	**0.135**	**0.075**
	Rotation	VGGT [71]	–	–	–	–	0.383	0.338	0.331	0.164
		CUT3R [74]	0.336	0.287	0.316	0.193	0.341	0.287	0.261	0.158
		ViPE [23]	0.386	0.332	0.474	0.269	0.371	0.325	0.375	0.227
		EgoMono4D [82]	0.449	0.388	0.516	0.335	0.408	0.372	0.524	0.334
		Ours	**0.305**	**0.236**	**0.156**	**0.088**	**0.288**	**0.227**	**0.164**	**0.097**

Table 7: **Camera Pose Estimation**: Comparison across metric and scale-invariant camera pose estimation on three subsets (random, locomotion-heavy, rotation-heavy) of each dataset. Lower values (↓) indicate better performance.

			Metric Scale			Scale Invariant		
		Method	ATE	RPE Rot	RPE Trans	ATE	RPE Rot	RPE Trans
EgoExo4D	Random	EgoEgo [32]	0.246	2.652	0.051	0.171	2.652	0.053
		VGGT [71]	–	–	–	0.156	2.510	0.049
		CUT3R [74]	0.269	2.341	0.075	0.174	2.341	0.045
		ViPE [23]	0.308	1.378	0.082	**0.053**	1.378	**0.020**
		EgoMono4D [82]	0.971	2.829	0.252	0.204	2.829	0.049
		Ours	**0.131**	**1.240**	**0.034**	0.082	**1.240**	0.024
	Locomotion	EgoEgo [32]	0.589	7.876	0.126	0.512	7.876	0.111
		VGGT [71]	–	–	–	0.343	4.930	0.106
		CUT3R [74]	0.556	5.796	0.160	0.410	5.796	0.105
		ViPE [23]	0.540	**3.711**	0.190	**0.169**	**3.711**	**0.044**
		EgoMono4D [82]	1.717	6.002	0.439	0.481	6.002	0.103
		Ours	**0.318**	3.988	**0.095**	0.237	3.988	0.070
	Rotation	EgoEgo [32]	1.015	14.620	0.219	0.831	14.620	0.159
		VGGT [71]	–	–	–	0.714	13.338	0.204
		CUT3R [74]	1.144	14.969	0.380	0.737	14.969	0.192
		ViPE [23]	1.416	10.668	0.580	**0.404**	10.668	**0.106**
		EgoMono4D [82]	2.286	10.967	0.620	0.757	10.967	0.164
		Ours	**0.579**	**10.199**	**0.209**	0.417	**10.199**	0.140
Nymeria	Random	EgoEgo [32]	0.352	3.057	0.053	0.261	3.057	0.073
		VGGT [71]	–	–	–	0.157	1.893	0.056
		CUT3R [74]	0.323	1.831	0.071	0.217	1.831	0.051
		ViPE [23]	**0.173**	1.235	0.042	**0.062**	1.235	**0.020**
		EgoMono4D [82]	0.966	2.667	0.238	0.254	2.667	0.065
		Ours	0.190	**1.104**	**0.032**	0.127	**1.104**	0.032
	Locomotion	EgoEgo [32]	0.901	6.774	0.138	0.772	6.774	0.208
		VGGT [71]	–	–	–	0.393	4.043	0.159
		CUT3R [74]	0.621	3.788	0.167	0.465	3.788	0.136
		ViPE [23]	**0.388**	**3.202**	0.133	**0.185**	**3.202**	**0.060**
		EgoMono4D [82]	1.666	5.131	0.408	0.618	5.131	0.164
		Ours	0.463	3.208	**0.090**	0.352	3.208	0.098
	Rotation	EgoEgo [32]	0.612	9.803	0.108	0.554	9.803	0.131
		VGGT [71]	–	–	–	0.347	**5.445**	0.153
		CUT3R [74]	0.550	6.132	0.211	0.396	6.132	0.134
		ViPE [23]	0.635	6.397	0.272	**0.248**	6.397	**0.078**
		EgoMono4D [82]	1.738	7.816	0.526	0.516	7.816	0.144
		Ours	**0.467**	**6.056**	**0.122**	0.404	6.056	0.115

Table 8: **Per-Frame Depth Accuracy**: Depth predicted by multi-view feedforward geometry models (top) and monocular depth models (bottom) on all frames of the EgoExo4D locomotion subset, using the full rectified frames (512×512) and their central 384×384 crop. Monocular depth models are more accurate than feedforward geometry models on every metric in both settings, which motivates their use as our geometry basis.

	Full frame				Center crop			
	Metric		Scale Invariant		Metric		Scale Invariant	
Method	AbsRel↓	$\delta_{1.25}$ ↑	AbsRel↓	$\delta_{1.25}$ ↑	AbsRel↓	$\delta_{1.25}$ ↑	AbsRel↓	$\delta_{1.25}$ ↑
VGGT [71]	–	–	0.543	0.503	–	–	0.353	0.576
CUT3R [74]	0.456	0.372	0.175	0.772	0.445	0.418	0.292	0.658
MoGe-2-L [76] (Used in EgoMo3R)	0.313	0.551	0.158	0.809	0.251	0.651	0.130	0.853
UniDepth-v2-L [53] (Used in ViPE)	0.262	0.656	0.153	0.824	0.285	0.619	0.138	0.843
Metric3D-v2-L [21]	0.228	0.632	0.146	0.828	0.226	0.618	0.139	0.841

Table 9: **Sensitivity to Off-the-Shelf Priors**: We replace the point tracker or the depth model in the scene-only setting (row 3 of Tab. 4). Lower values (↓) indicate better performance. Weaker priors degrade the results gradually, without catastrophic failure. The weaker tracker has the larger effect, mainly on ATE and RPE Trans, since less reliable long-range correspondences lead to more drift.

	ATE	RPE Trans	Acc		Comp	
Tracker + Depth			Mean	Median	Mean	Median
CoTracker3 + MoGe-2 (ours)	**0.353**	**0.103**	**0.414**	**0.337**	**0.210**	**0.135**
CoTracker3 + MoGe	0.389	0.123	0.415	0.345	0.213	0.141
CoTracker + MoGe-2	0.505	0.171	0.454	0.376	0.243	0.147

Table 10: **Long Sequences**: Results on 24 s EgoExo4D locomotion sequences, $3\times$ the length of the main evaluation clips, with metric scale alignment. Lower values (↓) indicate better performance. Our motion prior, trained on 8 s windows, is applied over overlapping windows with MultiDiffusion [2]; the rest of the optimization is unchanged. VGGT runs out of memory on a 48 GB NVIDIA L40S GPU. Even without explicit loop closure, EgoMo3R remains best on all reconstruction metrics and on ATE, which we attribute to the long-range tracks from *offline* CoTracker3 [26].

	Acc		Comp		ATE	RPE Rot	RPE Trans
Method	Mean	Median	Mean	Median			
VGGT [71]	out of memory						
CUT3R [74]	0.614	0.506	0.621	0.401	1.226	10.879	**0.251**
ViPE [23]	0.735	0.633	0.732	0.474	1.668	**8.308**	0.416
EgoMono4D [82]	0.786	0.699	0.836	0.554	2.411	8.851	0.415
Ours	**0.505**	**0.398**	**0.212**	**0.105**	**0.913**	9.605	0.343

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 abstract and introduction clearly state the contributions in joint optimization of camera poses, scene geometry, and human motion from egocentric locomotion video, supported by experimental results on EgoExo4D and Nymeria sequences (Tables 1-3) and ablation studies (Table 4).

   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 Appendix discusses the main limitations of EgoMo3R, including its reliance on visuo-geometric priors, long optimization time, and lack of hand motion optimization.

   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: NA.

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 implementation details in the Appendix, including the specific off-the-shelf models used, loss weights, optimizer settings, learning rate schedule, and the number of optimization steps per alternating round. The training data (AMASS) and evaluation datasets (EgoExo4D, Nymeria) are publicly available, and we commit to releasing code.

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 evaluation datasets (EgoExo4D, Nymeria) and the training dataset (AMASS) used for the motion prior are publicly available. We mention in the Appendix that the code will be released.

   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 Appendix provides comprehensive implementation details, including the loss weights, learning rate schedule, etc. The data splits and evaluation protocol are also described, and will be released with our code for reproducibility.

   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: [Yes]

Justification: Due to the computational cost of evaluation, prior works in this area do not report error bars, and we follow this convention. Instead, we report both mean and median metrics for 3D reconstruction (following prior works) and provide disaggregated results across three subsets (random, locomotion-heavy, rotation-heavy) in the Appendix to characterize performance variation.

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 Appendix reports that two rounds of alternating optimization take approximately 8 minutes per sample on NVIDIA L40S 48G GPU, including the one-time computation of visuo-geometric priors.

   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: The research conforms with the NeurIPS Code of Ethics. We use only publicly available datasets (EgoExo4D, Nymeria, AMASS) collected with appropriate consent by their original authors, and the work poses no foreseeable ethical concerns.

   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: [Yes]

    Justification: The paper's introduction highlights positive societal impacts in AR/VR, assistive technologies, healthcare, and robot learning. As with any egocentric perception system, potential negative impacts include privacy concerns associated with capture and reconstruction of personal environments; these concerns are widely explored in the egocentric vision community, and our work does not introduce any new concerns beyond those already discussed.

    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: The paper does not release any new datasets or pre-trained models with a high risk for misuse. The proposed method is an optimization framework that builds on existing publicly available models and datasets.

    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: All existing assets used in the paper, including datasets (EgoExo4D, Nymeria, AMASS) and pre-trained models (RAFT, CoTracker3, MoGe-2, SAM 2), are properly cited with references to their original publications. We use these assets in accordance with their respective licenses and terms of use.

    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: [N/A]

    Justification: The paper does not release any new datasets or models at submission time. Code will be released with appropriate documentation.

    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: The paper does not involve crowdsourcing or research with human subjects. All experiments are conducted on existing publicly available datasets.

    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: The paper does not involve research with human subjects. All experiments use existing publicly available datasets that were collected by their original authors with appropriate approvals.

    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: LLMs are not used as any important, original, or non-standard component of the core methodology in this research.

    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.