FORESIGHT TO MOTION: GOAL-CONDITIONED AUTOREGRESSIVE HUMAN MOTION PREDICTION FROM EGOCENTRIC VIDEO

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating plausible human motion from egocentric observations is important for human behavior understanding and embodied robotics. However, in first-person settings, most of the body remains outside the camera's field of view, making motion prediction from partial visual observations and egocentric camera motion challenging. Existing approaches rely on RGB-conditioned diffusion models, which are limited by sparse behavioral cues and fixed-horizon generation. To address these limitations, we introduce Foresight to Motion (FOM), a multimodal autoregressive framework for goal-driven, human motion prediction from egocentric video. Beyond visual observations, our model incorporates geometric and camera-motion cues together with explicit future interaction goals, providing guidance toward desired targets over extended temporal horizons. We further propose a three-stage training paradigm based on teacher forcing, self-forcing, and head-camera consistency, improving closed-loop robustness and reducing the train–inference discrepancy. Extensive experiments demonstrate strong performance on egocentric motion prediction benchmarks and improved cross-dataset generalization. Our goal-driven formulation also provides a flexible foundation for downstream human behavior modeling and video-conditioned motion planning for embodied systems.

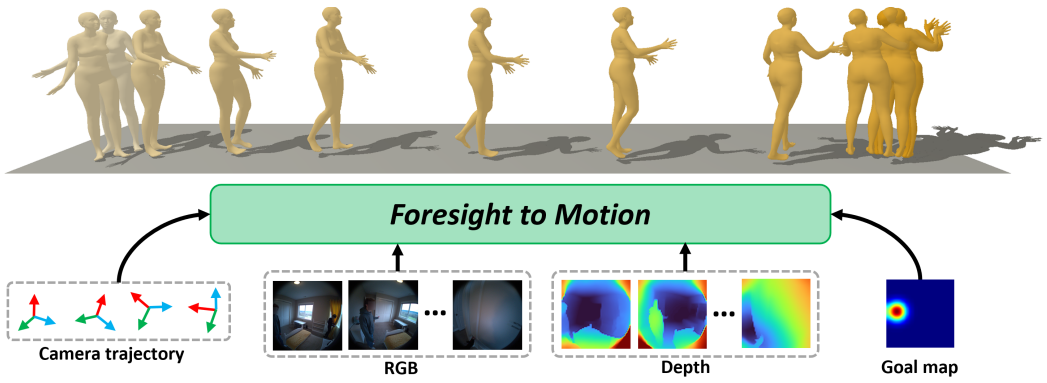

Figure 1: **Foresight to Motion** is a multimodal autoregressive model for long-horizon egocentric full-body motion prediction. It flexibly supports different combinations of camera trajectory, RGB, depth, and first-frame goal conditioning within a unified framework, and autoregressively generates temporally coherent motion.

1 INTRODUCTION

Human movement is tightly coupled to perception. A person turns toward an object, navigates around an obstacle, or reaches for a target based on visual evidence from the current viewpoint, while these actions in turn alter the first-person visual stream. This reciprocal relationship is central to augmented reality, avatar control, embodied simulation, and human-centered robotics. Yet, inferring full-body motion from an egocentric camera remains challenging because the camera observes the environment rather than the body that produces the motion. Most of the torso and lower body remain outside the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

field of view, visible hands are frequently occluded, and camera motion is itself coupled with head and body movement.

Recent egocentric datasets have enabled substantial progress in this direction. Ego4D (Grauman et al., 2022) established large-scale first-person video as a foundation for egocentric understanding. Nymeria (Ma et al., 2024) provides synchronized wearable observations and full-body motion capture in diverse real-world environments, while Aria Digital Twin (ADT) (Pan et al., 2023a) provides accurate camera trajectories, depth, object poses, and human-pose annotations. These datasets have supported recent advances in egocentric motion estimation, reconstruction, forecasting, and generation (Patel et al., 2025; Guzov et al., 2025; Yi et al., 2025; Escobar et al., 2025; Ge et al., 2026).

Despite this progress, long-horizon full-body motion prediction from egocentric video remains difficult. Existing approaches commonly rely on fixed-length sequence modeling or diffusion-based generation, which are less naturally suited to extending predictions over long temporal horizons. Moreover, egocentric observations alone may not sufficiently specify future intent: similar visual contexts can lead to substantially different motions depending on which object or spatial target the wearer intends to interact with. Explicit goal conditioning can therefore provide additional control over future behavior and, more broadly, support robotics applications in which a desired interaction target is specified by an upstream planner or agent.

We introduce FORESIGHT TO MOTION, a causal autoregressive framework for goal-driven, long-horizon full-body motion prediction from egocentric video. The model predicts motion sequentially from previous motion states and temporally aligned multimodal observations, including RGB, optional head trajectory, depth, and spatial goal conditions. The autoregressive formulation allows prediction to extend beyond a fixed temporal window, while goal conditioning enables the generated motion to be guided toward a desired interaction target. RGB, depth, and goal features are incorporated through causal cross-attention, while trajectory information is injected directly into the motion-token stream.

A key challenge of autoregressive prediction is the discrepancy between teacher-forced training and inference, where the model must condition on its own previous predictions. To improve long-horizon robustness, we introduce a three-stage training paradigm. We first train the model with teacher forcing to learn stable observation-to-motion mappings, then apply self-forcing post-training to expose the model to its own generated histories and reduce error accumulation. Finally, we introduce a head-camera consistency loss to better align the predicted body motion with the observed egocentric camera dynamics.

We evaluate FOM on both the Nymeria (Ma et al., 2024) and Aria Digital Twin (ADT) (Pan et al., 2023b) datasets, covering in-domain evaluation and cross-dataset generalization. Across quantitative and qualitative comparisons, our model consistently outperforms the diffusion-based UniEgoMotion Patel et al. (2025), particularly in global trajectory reconstruction and long-horizon motion prediction. Our contributions are threefold:

- We introduce FORESIGHT TO MOTION, a causal autoregressive framework for egocentric full-body motion prediction that supports long-horizon inference beyond a fixed window.
- We propose a goal-driven conditioning mechanism that allows motion prediction to be guided by a desired interaction target, enabling controllable generation and potential integration with embodied robotic planning systems.
- We develop a three-stage training paradigm consisting of teacher forcing, self-forcing post-training, and head-camera consistency optimization to improve robustness and motion quality during autoregressive rollout.

2 RELATED WORK

Egocentric full-body motion. Estimating full-body motion from a head-mounted camera is difficult because the body is largely unobserved and camera motion is coupled to the motion being estimated. EgoCast (Escobar et al., 2025) forecasts pose from egocentric video and proprioceptive observations. Ego4o (Wang et al., 2025) supports motion capture from different combinations of images, sparse inertial sensors, and language descriptions. EgoAllo (Yi et al., 2025) estimates allocentric body and hand motion from head-mounted images and SLAM trajectories, while EgoPoseFormer v2 (Li et al., 2026) improves egocentric motion estimation through temporally consistent transformer models and large-scale pseudo-labeling. Recent work also considers generative egocentric motion model-

ing. UniEgoMotion (Patel et al., 2025) unifies reconstruction, forecasting, and generation through conditional diffusion. HMD [2] (Guzov et al., 2025) incorporates head motion, scene geometry, and image features for online environment-aware generation. EgoPriMo (Ge et al., 2026) jointly models egocentric vision, body dynamics, and language for interactive humanoid control. EgoMotion (Hou et al., 2026) separates high-level vision-language reasoning from continuous motion generation. In contrast, our Foresight to Motion uses a causal autoregressive decoder whose visual conditioning can be captured from a wearable stream, partially observed, or synthesized by a video model.

Egocentric datasets and geometric context. Ego4D (Grauman et al., 2022) provides large-scale first-person video across diverse real-world environments. Nymeria (Ma et al., 2024) additionally provides synchronized multi-device egocentric recordings, full-body motion, and language annotations. NymeriaPlus (DeTone et al., 2026) extends this resource with improved motion representations and object- and scene-level annotations. ADT (Pan et al., 2023a) provides accurate camera poses, depth, object trajectories, and human motion in controlled environments. These resources also enable work on visuomotor prediction and interaction reasoning. Predictive visuomotor coordination jointly models head pose, gaze, and upper-body dynamics (Jia et al., 2025); 3D scanpath prediction models future visual attention (Ryan et al., 2026); and FICTION predicts future object interactions in 3D (Ashutosh et al., 2025). Our model uses trajectory, depth, and goal heatmaps as optional conditions for full-body generation while depth, and goal heatmaps as optional conditions for full-body generation while remaining operational when one or more modalities are unavailable.

Egocentric video generation and world models. Egocentric video generation models capture the relationship between visual dynamics and body motion. EgoControl (Pallotta et al., 2026) synthesizes future egocentric video conditioned on 3D body poses. Whole-body-conditioned egocentric video prediction (Bai et al., 2025) predicts video from body actions. EgoTwin (Xiu et al., 2025) jointly generates video and body motion with viewpoint-consistency constraints. These methods either generate visual observations from a specified body trajectory or jointly model both modalities. In contrast, our Foresight to Motion uses Wan 2.2 (Wan et al., 2025) as an external visual rollout model and learns the complementary mapping from first-person observations to full-body motion. This lets the motion decoder consume the same conditioning interface for captured and synthesized video.

3 METHODOLOGY

As illustrated in Figure 2, Foresight to Motion predicts full-body SMPL-X motion from multimodal egocentric observations. RGB, depth, goal, and camera-motion cues are encoded as causal conditioning signals for an Transformer and then decode the motion query to SMPL-X parameters in an autoregressive manner. We further adopt a three-stage training paradigm consisting of teacher forcing, self-forcing, and head-camera consistency post-training to boost the closed-loop rollout performance.

3.1 PROBLEM DEFINITION

Long-horizon video to motion. Given a sequence of multimodal observations with T frames, our goal is to recover the corresponding temporally coherent SMPL-X motion. We denote the predicted pose sequence and translation sequence as

$$\hat{\boldsymbol{\Theta}} = \left\{\hat{\boldsymbol{\theta}}_t\right\}_{t=1}^{T}, \qquad \hat{\mathbf{t}}_{1:T} = \left\{\hat{\mathbf{t}}_t\right\}_{t=1}^{T}. \tag{1}$$

We formulate the task as learning a general mapping from multimodal observations to an SMPL-X motion sequence:

$$\left(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\beta}}, \hat{\mathbf{t}}_{1:T}\right) = \mathcal{F}\left(\mathcal{X}^{\text{rgb}}, \mathcal{X}^{\text{depth}}, \mathcal{X}^{\text{goal}}, \mathcal{C}, \ldots\right). \tag{2}$$

where $\mathcal{X}^{\text{rgb}}$, $\mathcal{X}^{\text{depth}}$, and $\mathcal{X}^{\text{goal}}$ denote the RGB, depth, and goal-mask sequences, respectively, and $\mathcal{C}$ denote the estimated camera trajectory.

Human body parameterized model. We represent the 3D human body using the SMPL-X parametric model (Pavlakos et al., 2019), which provides a unified representation of the body, hands, and face. For each frame t, SMPL-X maps the pose parameters $\boldsymbol{\theta}_t \in \mathbb{R}^{53\times 3}$, represented in axis-angle form, the body shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$, and the global translation $\mathbf{t}_t \in \mathbb{R}^3$ to a 3D human mesh:

$$\mathbf{V}_t = \mathcal{M}_{\text{SMPL-X}}\left(\boldsymbol{\theta}_t, \boldsymbol{\beta}, \mathbf{t}_t\right) \in \mathbb{R}^{M\times 3}, \tag{3}$$

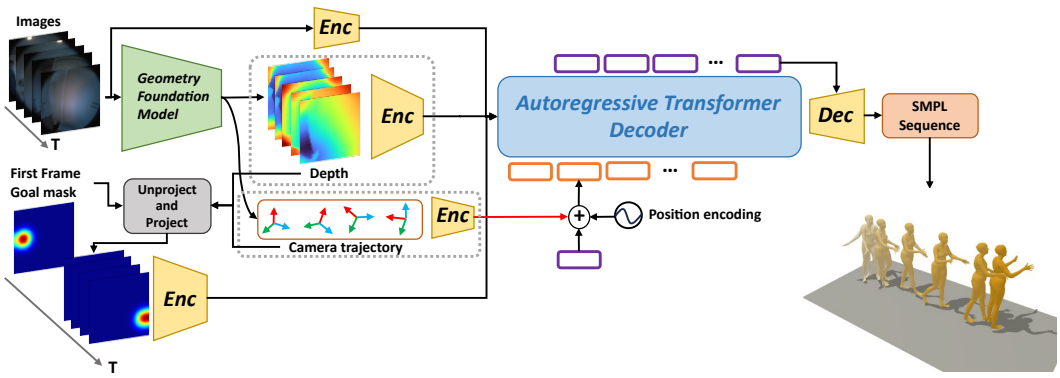

Figure 2: **Overall pipeline of Foresight to Motion** (FOM)**.** Given an egocentric image sequence, FOM extracts RGB features and uses a geometry foundation model to estimate depth and camera trajectory. An first-frame goal is propagated across the sequence to provide goal conditioning. An autoregressive Transformer conditioned on these modalities predicts full SMPL-X motion sequence.

where $M = 10{,}475$ denotes the number of mesh vertices. We assume that the body shape $\boldsymbol{\beta}$ remains constant throughout a sequence, while the pose and global translation vary over time.

Challenges. Egocentric motion prediction is difficult because most of the body lies outside the camera view. Existing RGB-based, fixed-window methods provide limited geometric and long-horizon cues, while similar observations may imply different motions depending on the interaction goal. These challenges motivate our multimodal, goal-conditioned autoregressive formulation.

3.2 Auto-regressive Motion Prediction

To facilitate long-horizon human motion prediction, we perform autoregressive sequence modeling for SMPL-X parameters prediction. Let $\mathbf{H}_{\text{motion}} = \{\mathbf{h}_t\}_{t=1}^{T}$ denote the sequence of latent motion tokens, where each $\mathbf{h}_t \in \mathbb{R}^d$ encodes the motion information at the corresponding time step. We further introduce a sequence-level latent shape feature y_{shape} that captures the body shape information shared across the entire sequence. The RGB frames, estimated depth maps, and goal masks are processed by modality-specific encoders to form the multimodal context tokens $\mathcal{X}$. In parallel, the camera trajectory estimated by the geometry foundation model is encoded into temporally aligned camera features $\mathcal{C} = \{\mathbf{c}_t\}_{t=1}^{T}$. We factorize the conditional distribution of the latent human motion representation as

$$p\left(h_{\text{shape}}, \mathbf{H}_{\text{motion}} \mid \mathcal{X}, \mathcal{C}\right) = p\left(h_{\text{shape}} \mid \mathcal{X}, \mathcal{C}\right) \times \prod_{t=1}^{T} p\left(\mathbf{h}_t \mid \mathbf{h}_{<t}, h_{\text{shape}}, \mathcal{X}, \mathbf{c}_t\right), \tag{4}$$

where $\mathbf{h}_{<t}$ denotes the preceding latent motion tokens. At each time step, we construct the autoregressive query by combining the preceding motion token with the temporally aligned camera feature and positional embedding:

$$\mathbf{q}_t = W_y \mathbf{h}_{t-1} + E_{\text{cam}}(\mathbf{c}_t) + \mathbf{p}_t. \tag{5}$$

The query sequence is then processed by a Transformer (Vaswani et al., 2017) decoder with causal self-attention and cross-attention to the multimodal context:

$$h_{\text{shape}}, \mathbf{H}_{\text{motion}} = \text{TransformerDecoder}_{\phi}\left(\mathbf{q}_0, \mathcal{X}, \mathcal{C}; \mathbf{M}_{\text{causal}}\right), \tag{6}$$

where $\mathbf{M}_{\text{causal}}$ is the causal mask which prevents each query from attending to future motion tokens. Finally, the generated latent representations are decoded into explicit SMPL-X parameters:

$$\hat{\boldsymbol{\beta}} = D_{\text{shape}}(h_{\text{shape}}), \tag{7}$$

$$\left(\hat{\boldsymbol{\theta}}_t, \hat{\mathbf{t}}_t\right) = D_{\text{motion}}\left(\mathbf{h}_t, h_{\text{shape}}\right). \tag{8}$$

where D_{shape} and D_{motion} are SMPL-X shape decoder and motion decoder.

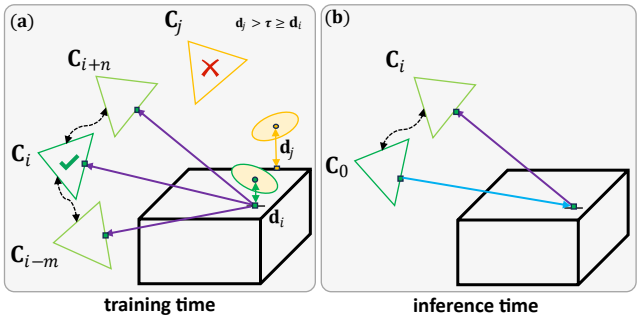

Figure 3: Goal guidance. C_i represent the camera view at frame i, d_i is the distance between sampled hand vertices and the nearest surface of the object bounding box. **(a)** The goal point is selected using ground truth hand and objects bounding box during training time. **(b)** In inference-time, the goal point of the first frame considered as an optional input, followed by 2D Gaussian encoding.

3.3 Goal-driven Multimodal-conditioning

Our model conditions motion generation on complementary visual, depth, goal, and camera-motion cues. To preserve temporal causality, both the autoregressive motion stream and the multimodal conditioning stream are causally masked, such that the prediction at time t cannot access any observation from future frames.

Visual and depth conditioning. Given an egocentric RGB sequence $\mathcal{I} = \{I_t\}_{t=1}^{T}$, we use ViPE (Huang et al., 2025) to estimate dense depth maps $\{D_t\}_{t=1}^{T}$ and camera poses $\{\mathcal{C}_t\}_{t=1}^{T}$:

$$\left(\{D_t\}_{t=1}^{T}, \{\mathcal{C}_t\}_{t=1}^{T}\right) = \Phi_{\text{ViPE}}(\mathcal{I}). \tag{9}$$

Each RGB frame is encoded by a frozen DINOv2 (Oquab et al., 2023) backbone, while the corresponding depth map and goal mask are processed by modality-specific encoders:

$$\mathbf{Z}_t^I = \phi_I(I_t), \qquad \mathbf{Z}_t^D = \phi_D(D_t), \qquad \mathbf{Z}_t^G = \phi_G(g_t). \tag{10}$$

The frame-wise multimodal context is constructed as $\mathcal{X}_t = \left[\mathbf{Z}_t^I; \mathbf{Z}_t^D; \mathbf{Z}_t^G\right]$,.

Causal multimodal attention. We apply a temporal causal mask to the cross-attention between motion queries and multimodal context. Specifically, a query at time t can only attend to condition tokens from frames $s \leq t$:

$$\mathbf{M}_{t,s}^{\text{cond}} = \begin{cases} 0, & s \leq t, \\ -\infty, & s > t. \end{cases} \tag{11}$$

This prevents future RGB, depth, or goal information from leaking into the current motion prediction.

Camera-trajectory conditioning. Camera motion provides a direct cue for the global motion of the wearer. We derive the camera trajectory from the ViPE camera poses and encode it as a temporally aligned feature $\boldsymbol{\tau}_t$. Instead of including it in the cross-attention context, we directly inject it into the autoregressive motion query using Equation 5.

Modality masking. To support different conditioning settings with a single model, we apply modality dropout during training. For each contextual modality $q \in \{I, D, G\}$, an unavailable condition is replaced by a learned modality-specific mask embedding:

$$\tilde{\mathbf{Z}}_t^q = \begin{cases} \mathbf{Z}_t^q, & \text{if } q \text{ is available,} \\ \mathbf{e}_{\text{mask}}^q, & \text{otherwise.} \end{cases} \tag{12}$$

Similarly, an unavailable camera trajectory is replaced by a learned trajectory-mask feature before being injected into the autoregressive query. During training, we randomly mask depth, goal, and camera-trajectory conditions, and additionally mask future RGB observations while retaining the initial frame. This enables the same architecture to operate under multiple inference settings. Additionally, future RGB slots can be filled by video rollout from a generated egocentric, e.g., from Wan 2.2 (Wan et al., 2025), without modifying the motion model.

Goal Conditioning. We explicitly condition the motion model on a spatial interaction goal. As illustrated in Figure 3, training and inference use different mechanisms to obtain the 3D goal, but share the same projection and encoding procedure. During training, we use ground-truth body motion and object annotations. At the first contact frame t_c, the world-space goal is defined as the closest point on the contacted object bounding-box surface to the sampled hand vertices:

$$\mathbf{g}^w = \arg\min_{\mathbf{p}\in\partial\mathcal{B}} \min_{\mathbf{v}\in\mathcal{H}_{t_c}} \|\mathbf{v}-\mathbf{p}\|_2, \tag{13}$$

where $\mathcal{B}$ denotes the contacted object bounding box and $\mathcal{H}_{t_c}$ denotes the selected hand vertices. At inference time, the user instead specifies a 2D goal $\mathbf{u}_0^g$ in the first egocentric frame. Using the estimated depth and camera intrinsics, we first recover its 3D location in the initial camera coordinate system and transform it into the world frame:

$$\mathbf{g}^{c_0} = D_0(\mathbf{u}_0^g)\mathbf{K}_0^{-1}\tilde{\mathbf{u}}_0^g, \qquad \mathbf{g}^w = \mathbf{T}_{c2w}^0\mathbf{g}^{c_0}. \tag{14}$$

where $D_0(\mathbf{u}_0^g)$ is the estimated depth at the selected goal pixel, $\mathbf{K}_0$ is the camera intrinsic matrix, $\tilde{\mathbf{u}}_0^g$ is the homogeneous image coordinate, and $\mathbf{T}_{c2w}^0$ denotes the camera-to-world transformation of the first frame. Once $\mathbf{g}^w$ is obtained, it remains fixed throughout the sequence. For each frame t, we transform it into the corresponding camera coordinate system and project it onto the image plane:

$$\mathbf{g}^{c_t} = \mathbf{T}_{w2c}^t\mathbf{g}^w, \qquad \mathbf{u}_t^g = \pi(\mathbf{K}_t\mathbf{g}^{c_t}). \tag{15}$$

where $\mathbf{T}_{w2c}^t$ denotes the world-to-camera transformation at frame t, $\mathbf{K}_t$ is the corresponding camera intrinsic matrix, and $\pi(\cdot)$ denotes perspective projection from homogeneous camera coordinates to the image plane. Each projected goal $\mathbf{u}_t^g$ is then encoded as a 2D Gaussian heatmap and used as a causally masked conditioning signal for motion generation.

3.4 Objectives and Multi-stage Training

We adopt a three-stage training curriculum to progressively bridge the gap between supervised autoregressive learning and closed-loop motion generation.

Motion objective. Let $\mathbf{z}_t \in \mathbb{R}^F$ denote the normalized per-frame motion representation, comprising joint rotations and positions in a canonical frame, the frame-to-frame global transform, hand pose, and shape parameters. The motion decoder predicts $\hat{\mathbf{z}}_t = D_{\text{motion}}(\mathbf{h}_t)$, from which the SMPL-X parameters $(\hat{\boldsymbol{\theta}}_t, \hat{\mathbf{t}}_t, \hat{\boldsymbol{\beta}})$ are recovered deterministically. All stages are supervised with a feature-weighted masked reconstruction loss:

$$\mathcal{L}_{\text{motion}} = \frac{\sum_{t=1}^{T}\sum_{k=1}^{F} m_t\, w_k \big(\hat{z}_{t,k} - z_{t,k}\big)^2}{\sum_{t=1}^{T}\sum_{k=1}^{F} m_t\, w_k}, \qquad w_k = \begin{cases} \lambda_{\text{glob}}, & k \in \mathcal{K}_{\text{glob}}, \\ 1, & \text{otherwise}, \end{cases} \tag{16}$$

where $m_t \in \{0,1\}$ indicates whether frame t is valid (non-padded), and $\mathcal{K}_{\text{glob}}$ indexes the global root-orientation and root-translation channels. Since errors in global motion accumulate during autoregressive rollout, we up-weight these channels with λ_{glob}. We use this objective function for both teacher-forcing and self-forcing stage.

Stage I: Teacher-forcing. We first train the model with teacher forcing, where the ground-truth preceding motion token is used as the autoregressive feedback. This allows all time steps to be processed in parallel, providing stable and efficient optimization. However, the model is only exposed to perfect motion histories, leading to a train–test discrepancy during closed-loop inference.

Stage II: Self-forcing. We then switch to self-forcing, where the model conditions on its own previous predictions. We unify the two training stages by defining the feedback token as

$$\tilde{\mathbf{z}}_{t-1} = \begin{cases} \mathbf{z}_{t-1}, & \text{teacher forcing}, \\ \text{sg}(\hat{\mathbf{z}}_{t-1}), & \text{self-forcing}, \end{cases} \tag{17}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operation. The autoregressive query is uniformly written as

$$\mathbf{q}_t = \mathbf{W}_z\tilde{\mathbf{z}}_{t-1} + \phi_\tau(\boldsymbol{\tau}_t) + \mathbf{p}_t. \tag{18}$$

Self-forcing exposes the model to its own accumulated prediction errors, thereby reducing exposure bias and better matching closed-loop inference.

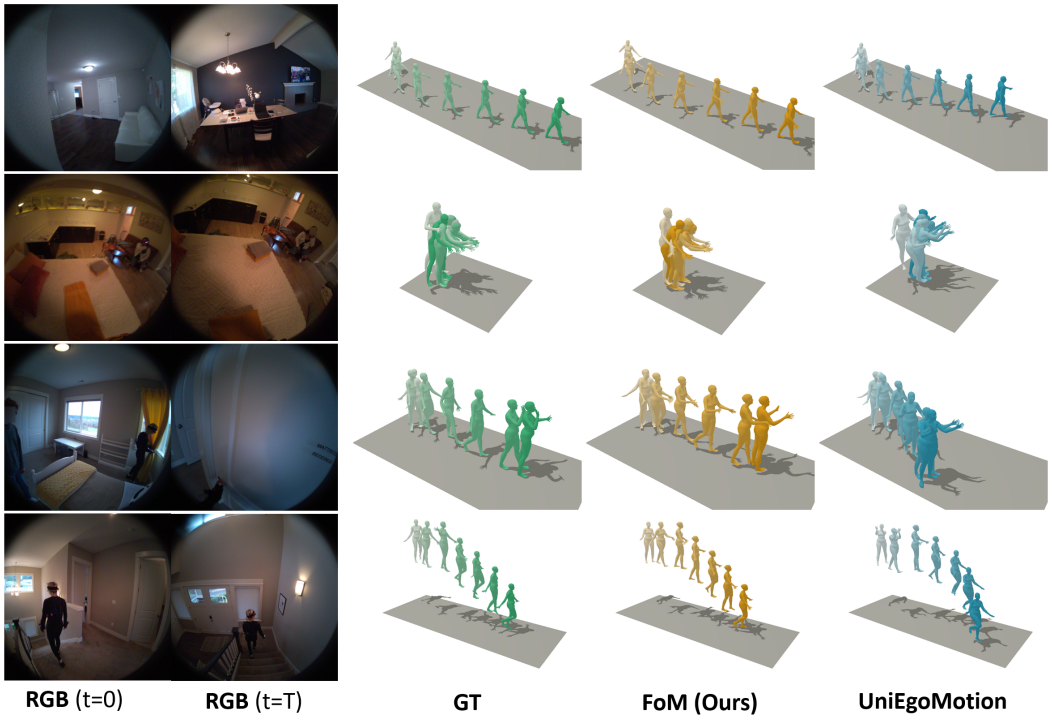

Figure 4: Qualitative comparison between our Foresight to Motion (FoM) and UniEgoMotion.

Table 1: Comparison of ours FoM with baseline on the ADT and Nymeria validation sets after using Nymeria Dataset as the training data. ADT data is being used for the cross dataset validation.

Testset	Method	MPJPE ↓	PA-MPJPE ↓	Body ↓	Hand ↓	Head Rot. ↓	Head Trans. ↓
ADT	UniEgoMotion	0.344	0.191	0.299	0.500	0.965	0.247
	FoM (Ours)	**0.254**	**0.186**	**0.189**	**0.477**	**0.861**	**0.126**
Nymeria	UniEgoMotion	0.304	0.118	0.258	0.427	0.506	0.236
	FoM (Ours)	**0.213**	**0.116**	**0.166**	**0.346**	**0.417**	**0.140**

Stage III: Head-camera consistency. Finally, we optionally fine-tune the model with an additional geometric consistency constraint between the predicted head position and the observed camera trajectory. Let $\hat{\mathbf{j}}_t^{\text{head}}$ denote the predicted head joint and $\mathbf{p}_t^{\text{cam}}$ the camera position. We define

$$\mathcal{L}_{\text{head}} = \frac{1}{T}\sum_{t=1}^{T}\left\|\hat{\mathbf{j}}_t^{\text{head}} - \mathbf{p}_t^{\text{cam}}\right\|_2^2, \tag{19}$$

and optimize

$$\mathcal{L} = \mathcal{L}_{\text{motion}} + \lambda_{\text{head}}\mathcal{L}_{\text{head}}. \tag{20}$$

This constraint is applied only when the camera trajectory is available, encouraging the predicted body motion to remain geometrically consistent with the observed egocentric camera motion.

4 Experiment

4.1 Setup

Datasets. We conduct experiments on the Nymeria (Nagymáté & Kiss, 2018) and Aria Digital Twin (ADT) (Pan et al., 2023a) datasets. For Nymeria, we construct a subject- and location-disjoint split containing 14,962 training sequences from 218 episodes, 57 people, and 13 locations, and 3,439 validation sequences from 29 episodes, 29 people, and 29 locations. We use 10-second clips sampled

Table 2: Ablation of conditioning modalities on the Nymeria validation set. All variants use the same model weights, with modalities removed only at inference. Best and second-best results are highlighted in red and yellow, respectively.

No.	Modalities				Nymeria Val.					
	RGB	Depth	Goal	Cam	MPJPE ↓	PA-MPJPE ↓	Body ↓	Hand ↓	Head Rot. ↓	Head Trans. ↓
(1)	✓	✓	✓	✓	**0.194**	**0.118**	**0.149**	**0.347**	**0.428**	**0.103**
(2)	✓	✓	✓		0.935	0.136	0.879	1.039	1.067	0.882
(3)	✓	✓		✓	0.577	0.159	0.525	0.725	0.815	0.489
(4)	✓	✓			1.065	0.152	0.997	1.188	1.203	1.001
(5)	✓			✓	0.660	0.149	0.606	0.796	0.885	0.580
(6)	✓				1.067	0.150	0.999	1.190	1.186	1.002
(7)		✓		✓	0.707	0.230	0.637	0.860	1.190	0.633
(8)		✓			1.189	0.234	1.096	1.327	1.467	1.135

Table 3: Ablation of training strategies on the Nymeria validation set. Best results are shown in **bold**.

Training Strategy			Nymeria Val.					
Teacher	Self	HeadCam	MPJPE ↓	PA-MPJPE ↓	Body ↓	Hand ↓	H-Rot ↓	H-Trans ↓
✓			0.606	0.137	0.561	0.727	0.905	0.535
✓	✓		**0.194**	0.118	**0.149**	0.347	0.428	**0.103**
✓	✓	✓	0.213	**0.116**	0.166	**0.346**	**0.417**	0.140

at 15 FPS, resulting in 150 motion frames per sequence. For ADT, the constructed split contains 2,394 training sequences from 71 episodes and 299 validation sequences from 64 episodes. ADT is confined to a single indoor apartment environment with two annotated skeleton identities. Unless otherwise specified, all models are trained exclusively on the Nymeria training split. We evaluate on the Nymeria validation set for in-domain performance and on the ADT validation set for cross-dataset generalization; no ADT data are used for training.

Baseline. We compare against UniEgoMotion (Patel et al., 2025), a diffusion model that predicts fixed-length egocentric motion sequences. For a fair comparison, we retrain UniEgoMotion using the same Nymeria training split and evaluate both methods under the same protocol.

Metrics. We report MPJPE, MPJPE-PA, MPJPE-Body, MPJPE-H, Head Rotation Error, and Head Translation Error. MPJPE measures absolute 3D joint-position error, while MPJPE-PA applies Procrustes alignment to isolate pose accuracy from global placement. MPJPE-Body and MPJPE-H measure unaligned errors on the body and hand joints, respectively. Head Rotation Error (Head Rot.) measures the orientation error of the predicted head, while Head Translation Error (Head Trans.) measures its positional error. All metrics are reported in meters except Head Rotation Error.

4.2 Main Results

Quantitative results. As shown in Table 1, FOM consistently outperforms UniEgoMotion across both Nymeria and ADT. On Nymeria, FOM reduces MPJPE from 0.304 m to 0.213 m and Head Translation Error from 0.236 m to 0.140 m, corresponding to relative improvements of 29.9% and 40.7%, respectively. Similar gains are observed on ADT, where MPJPE decreases from 0.344 m to 0.254 m and Head Translation Error from 0.247 m to 0.126 m. Improvements in PA-MPJPE are comparatively smaller, suggesting that the main advantage of FOM lies in recovering global motion and long-horizon body placement rather than only improving locally aligned pose. The gains on ADT are obtained without training on ADT, demonstrating stronger cross-dataset generalization.

Qualitative results. Figure 4 further illustrates the difference between FOM and UniEgoMotion. Overall, FOM better preserves both local body motion and global trajectory over long rollouts. In the second row, FOM more closely reconstructs the forward-reaching hand and upper-body configuration,

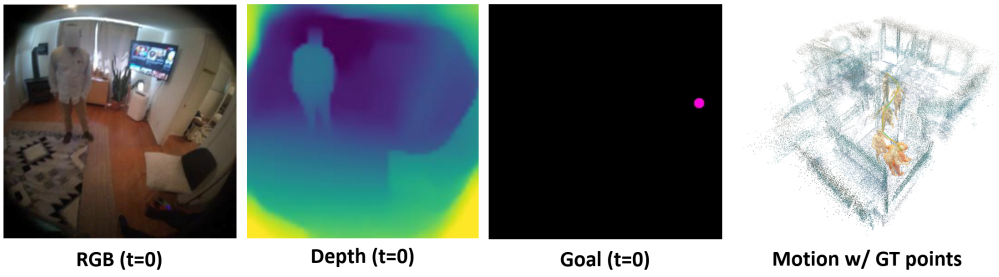

Figure 5: Visualization of predicted SMPL-X mesh sequence placed in Nymeria ground truth pointcloud.

while UniEgoMotion shows larger deviations in the arm and hand motion. In the fourth row, where the subject walks downstairs, FOM follows the ground-truth global trajectory more faithfully, whereas UniEgoMotion exhibits noticeable trajectory drift and displacement toward the later frames. Similar differences can be observed in the other examples, where FOM better maintains the overall direction and final body placement. These qualitative observations are consistent with the improvements in MPJPE, hand error, and head-translation error reported in Table 1.

4.3 Ablation and Analysis

Multimodal conditioning. Our model is designed to support flexible combinations of conditioning modalities within a single unified architecture. Using the modality masking strategy during training time, we are able to evaluate all variants in Table 2 use exactly the same model weights, with modalities removed only at inference time. Using all modalities consistently achieves the best performance across all metrics, demonstrating the complementary roles of RGB, depth, goal, and camera trajectory. Removing either the goal condition or camera trajectory leads to a substantial increase in global motion error: without the goal, MPJPE increases from 0.194 m to 0.577 m, while removing the camera trajectory increases it to 0.935 m. Depth provides additional geometric cues and further improves performance when combined with RGB and camera motion.

Training strategy. Table S2 shows the effect of our three-stage training paradigm. Teacher forcing alone yields an MPJPE of 0.606 m, whereas introducing self-forcing reduces it to 0.194 m, with similar improvements in body, hand, head-rotation, and head-translation errors. This confirms that training with the model's own predictions is critical for reducing exposure bias and maintaining stable closed-loop rollouts. Adding the head-camera stage further improves PA-MPJPE, hand error, and head-rotation error, although it slightly increases MPJPE and head-translation error, indicating a trade-off between local pose consistency and global translation accuracy.

Scene-aligned motion visualization. As shown in Figure 5, we further visualize the predicted human mesh sequence together with the ground-truth scene point cloud. The reconstructed body remains well aligned with the surrounding 3D geometry throughout the sequence, preserving consistent global position and motion relative to the scene. This qualitative alignment further demonstrates that our model recovers not only plausible body motion, but also spatially coherent motion within the scene.

5 Conclusion

We present Foresight to Motion, a multimodal autoregressive framework for long-horizon motion prediction from egocentric observations. Our approach integrates RGB, depth, camera trajectory, and goal guidance within a unified causal Transformer, while supporting flexible combinations of conditioning modalities. We further introduce explicit goal conditioning for target-directed motion and a three-stage training paradigm to improve closed-loop prediction. Experiments on Nymeria and ADT demonstrate consistent improvements over the diffusion-based baseline, with particularly strong gains in global motion and improved cross-dataset generalization. Qualitative results further show better alignment of predicted motion with both the global trajectory and surrounding 3D scene. Future work will extend the framework to generative video models and high-level planners, enabling applications in human behavior modeling and embodied motion planning.

AI USE STATEMENT

We have not used generative AI tools for tasks with required disclosure, and the remaining required disclosure tasks are not applicable to this work. We used generative AI tools to assist with software code editing and debugging, literature review, and grammar and readability improvements of the manuscript. All AI-assisted code was reviewed and tested by the authors, literature-related outputs were verified against the original sources, and all AI-assisted text was manually reviewed and revised. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our work studies egocentric full-body motion prediction using existing public datasets and does not involve any new human-subject data collection. For one of the datasets, identifiable facial regions were already anonymized through blurring before the data were made available to us. For the other dataset, the recordings were collected with participants' consent for public research release. We do not use identity information or attempt to recover anonymized personal information, and our analysis focuses on human motion and scene-level geometric cues. We use all datasets in accordance with their respective licenses and intended research purposes and adhere to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the model architecture, multimodal conditioning, goal construction, autoregressive formulation, and three-stage training paradigm in the main paper. Dataset construction, train/validation splits, evaluation metrics, optimization settings, and implementation details are also specified to facilitate reproduction of the reported results. The appendix provides additional experimental results, training-stage ablations, modality ablations, and qualitative visualizations. Together, these materials document the data processing, training, and evaluation procedures used in our experiments.

REFERENCES

Kumar Ashutosh, Georgios Pavlakos, and Kristen Grauman. FIction: 4d future interaction prediction from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Yutong Bai, Danny Tran, Amir Bar, Yann LeCun, Trevor Darrell, and Jitendra Malik. Whole-body conditioned egocentric video prediction, 2025. URL https://arxiv.org/abs/2506.21552.

Daniel DeTone, Federica Bogo, Eric-Tuan Le, Duncan Frost, Julian Straub, Yawar Siddiqui, Yuting Ye, Jakob Engel, Richard Newcombe, and Lingni Ma. Nymeriaplus: Enriching nymeria dataset with additional annotations and data. *arXiv preprint arXiv:2603.18496*, 2026.

Maria Escobar, Juanita Puentes, Cristhian Forigua, Jordi Pont-Tuset, Kevis-Kokitsi Maninis, and Pablo Arbeláez. Egocast: Forecasting egocentric human pose in the wild. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pp. 5831–5841, 2025.

Haoyang Ge, Peng Ren, Yukun Shi, Cong Huang, Kun Li, and Kai Chen. Egoprimo: Egocentric motion generation for interactive humanoid control. *arXiv preprint arXiv:2606.08495*, 2026.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18995–19012, 2022.

Vladimir Guzov, Yifeng Jiang, Fangzhou Hong, Gerard Pons-Moll, Richard Newcombe, C. Karen Liu, Yuting Ye, and Lingni Ma. Hmd^2: Environment-aware motion generation from single egocentric head-mounted device. In *International Conference on 3D Vision (3DV)*, March 2025.

Ruibing Hou, Mingyue Zhou, Yuwei Gui, Mingshuang Luo, Bingpeng Ma, Hong Chang, Shiguang Shan, and Xilin Chen. Egomotion: Hierarchical reasoning and diffusion for egocentric vision-language motion generation. *arXiv preprint arXiv:2604.19105*, 2026.

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, et al. Vipe: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934*, 2025.

Wenqi Jia, Bolin Lai, Miao Liu, Danfei Xu, and James M. Rehg. Learning predictive visuomotor coordination. *arXiv preprint arXiv:2503.23300*, 2025.

Zhenyu Li, Sai Kumar Dwivedi, Filip Maric, Carlos Chacón, Nadine Bertsch, Filippo Arcadu, Tomas Hodan, Michael Ramamonjisoa, Peter Wonka, Amy Zhao, Robin Kips, Cem Keskin, Anastasia Tkach, and Chenhongyi Yang. Egoposeformer v2: Accurate egocentric human motion estimation for AR/VR. *arXiv preprint arXiv:2603.04090*, 2026.

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

Gergely Nagymáté and Rita M. Kiss. Application of OptiTrack motion capture systems in human movement analysis: A systematic literature review. *Recent Innovations in Mechatronics (RIiM)*, 5 (1):1–9, 2018. doi: 10.17667/riim.2018.1/13.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Enrico Pallotta, Sina Mokhtarzadeh Azar, Lars Doorenbos, Serdar Ozsoy, Umar Iqbal, and Juergen Gall. Egocontrol: Controllable egocentric video generation via 3D full-body poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Carl Yuheng Ren. Aria digital twin: A new benchmark dataset for egocentric 3D machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20076–20086, 2023a.

Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20076–20086. IEEE, 2023b.

Chaitanya Patel, Hiroki Nakamura, Yuta Kyuragi, Kazuki Kozuka, Juan Carlos Niebles, and Ehsan Adeli. Uniegomotion: A unified model for egocentric motion reconstruction, forecasting, and generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10318–10329, 2025.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael 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 (CVPR)*, pp. 10975–10985, 2019.

Fiona Ryan, Ishwarya Ananthabhotla, Yijun Qian, Judy Hoffman, James M. Rehg, Vamsi Krishna Ithapu, and Calvin Murdock. Forecasting 3d scanpaths in egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Team Wan, Ang Wang, Baole Ai, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jian Wang, Rishabh Dabral, Diogo Luvizon, Zhe Cao, Lingjie Liu, Thabo Beeler, and Christian Theobalt. Ego4o: Egocentric human motion capture and understanding from multi-modal input. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22668–22679, 2025.

Jingqiao Xiu, Fangzhou Hong, Yicong Li, Mengze Li, Wentao Wang, Sirui Han, Liang Pan, and Ziwei Liu. Egotwin: Dreaming body and view in first person. *arXiv preprint arXiv:2508.13013*, 2025.

Brent Yi, Vickie Ye, Maya Zheng, Yunqi Li, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. Estimating body and hand motion in an ego-sensed world. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7072–7084, 2025.

A APPENDIX

A.1 OVERVIEW

In Section A.2, we will introduce the implementation details and training configurations. In Section A.3, we will provide more quantitative results and analysis. In Section A.4, we will provide additional qualitative results. In Section A.5, we will explain all the metrics we used in our experiments. Finally, we will demonstrate the limitation of our approach in Section A.6

A.2 IMPLEMENTATION DETAILS

We train all models on one NVIDIA H200 GPUs using AdamW with a constant learning rate of 3×10^{-5} and a weight decay of 0.01. Gradients are clipped to a maximum norm of 1.0, and we maintain an exponential moving average (EMA) of the model weights with a decay of 0.999 for evaluation. Unless otherwise specified, motion sequences contain 150 frames sampled at 15 FPS (10 seconds). Training follows the three-stage paradigm described above: we first train with teacher forcing for 300 epochs using a batch size of 64, followed by 100 epochs of self-forcing with a batch size of 16, and finally 100 epochs of head-camera consistency fine-tuning with the same batch size of 16. The dataset statistics of Nymeria and ADT, are shown in Table S3 and Table S4.

A.3 ADDITIONAL QUANTITATIVE RESULTS

Table S1: Ablation of training schedule on the Nymeria validation set. Teacher-forcing is evaluated at epochs 300, 400, and 500; self-forcing and self-forcing + HeadCam report the best checkpoint over 100 training epochs. Best results are shown in **bold**.

Training Strategy		**Nymeria Val.**					
Method	Epoch	MPJPE ↓	PA-MPJPE ↓	Body ↓	Hand ↓	H-Rot ↓	H-Trans ↓
Teacher	300	0.606	0.137	0.561	0.727	0.905	0.535
Teacher	400	0.661	0.133	0.616	0.766	0.861	0.601
Teacher	500	0.738	0.133	0.685	0.852	0.988	0.676
Self	100†	**0.194**	0.118	**0.149**	0.347	0.428	**0.103**
Self + HeadCam	100†	0.213	**0.116**	0.166	**0.346**	**0.417**	0.140

† Best of 100 epochs (self-forcing warm-started from teacher epoch 299 and self-forcing + HeadCam started from the self-forcing checkpoint).

Table S2: FOM Testing on ADT after the complete 3 step training on ADT Dataset

Method	Epoch	MPJPE↓	PA-MPJPE↓	Body↓	Hand↓	H-Rot↓	H-Trans↓
Teacher	300	1.033	0.115	1.028	1.066	0.864	1.004
Self	100	1.406	**0.059**	1.303	1.436	1.014	1.437
Self + HeadCam	100	**0.562**	0.064	**0.720**	**0.639**	**0.810**	**0.366**

Three-stage quantitative results on ADT. Table S2 further evaluates the three-stage training paradigm when training and testing on ADT. Teacher forcing provides a stable initial model, while self-forcing substantially improves PA-MPJPE from 0.115 m to 0.059 m, indicating better locally aligned pose prediction. However, its absolute MPJPE and head-translation error increase, suggesting increased drift in global motion during closed-loop rollout. Adding head-camera consistency largely corrects this drift, reducing MPJPE to 0.562 m and Head Translation Error to 0.366 m, while also achieving the best body, hand, and head-rotation errors. These results highlight the complementary roles of self-forcing for pose refinement and head-camera consistency for maintaining accurate global motion.

Training schedule ablation. To ensure a fair comparison between different training stages, we control for the total number of training epochs. The teacher-forcing model at epoch 300 serves as

Table S3: Train/Validation split constructed from the Nymeria dataset. We partition by *location* and *people* so that no recording site appears in both splits, yielding disjoint sets of environments and participants.

	Training	**Validation**
Sequences	14,962	3,439
Episodes	218	29
People	57	29
Locations	13 (Loc_04–17)	29 (Loc_18–47)

Table S4: Train/Validation split constructed from the Aria Digital Twin (ADT) dataset. We use apartment sequences with body-pose annotations and extract contact-centered motion clips. Unlike Nymeria, ADT is confined to a single indoor apartment scene and two skeleton identities, including dual-person recordings.

	Training	Validation
Sequences	2,394	299
Episodes	71	64
People	2	2
Locations	1 (Apartment)	1 (Apartment)

the initialization for self-forcing. We therefore compare the self-forcing model after 100 additional epochs against the teacher-forcing model at epoch 400, such that both receive the same total training budget. Similarly, the self-forcing + HeadCam model is compared against the teacher-forcing model at epoch 500, matching the additional 100 epochs of HeadCam fine-tuning. As shown in Table S1, simply extending teacher-forcing training does not improve closed-loop performance; instead, MPJPE increases from 0.606 m to 0.661 m and 0.738 m. In contrast, self-forcing substantially reduces MPJPE to 0.194 m, confirming that the gain comes from reducing the train–inference mismatch rather than from additional training epochs. Head-camera fine-tuning further improves PA-MPJPE, hand error, and head-rotation error, while introducing a small trade-off in global translation accuracy.

A.4 Additional qualitative results

Additional qualitative comparison between ours FOM and UniEgoMotion. Figure S1 presents additional qualitative comparisons between FOM and UniEgoMotion across diverse motion sequences. FOM more closely follows the ground-truth temporal evolution of both body pose and global trajectory, particularly in sequences involving changes in walking direction, reaching motions, and interaction-related upper-body movement. UniEgoMotion exhibits larger deviations in several examples, including drift in global placement and less accurate arm or hand configurations toward later frames. These additional cases further support the quantitative improvements reported in the main paper.

Additional samples for human motion placed in ground truth pointcloud. Figure S2 presents additional examples across different indoor layouts and interaction targets. The predicted motion sequences remain spatially compatible with the reconstructed scene geometry while adapting to different goal locations and camera viewpoints. These examples further show that the model can generate goal-directed trajectories that respect the surrounding environment across diverse scenes, rather than overfitting to a specific room layout or motion pattern.

Comparison between prediction w/ goal mask and w/o goal mask. Figure S3 compares predictions from the same trained checkpoint with and without the goal mask at inference time, isolating the effect of goal conditioning from differences in model parameters. In the first row, providing the goal substantially changes the predicted global trajectory and body placement, guiding the motion toward the intended interaction target and producing a trajectory closer to the ground truth. In the second row, the goal condition also affects the local upper-body motion, particularly the arm and hand configuration near the interaction. These examples show that goal conditioning influences not only where the body moves globally, but also how the model executes target-directed interaction motions.

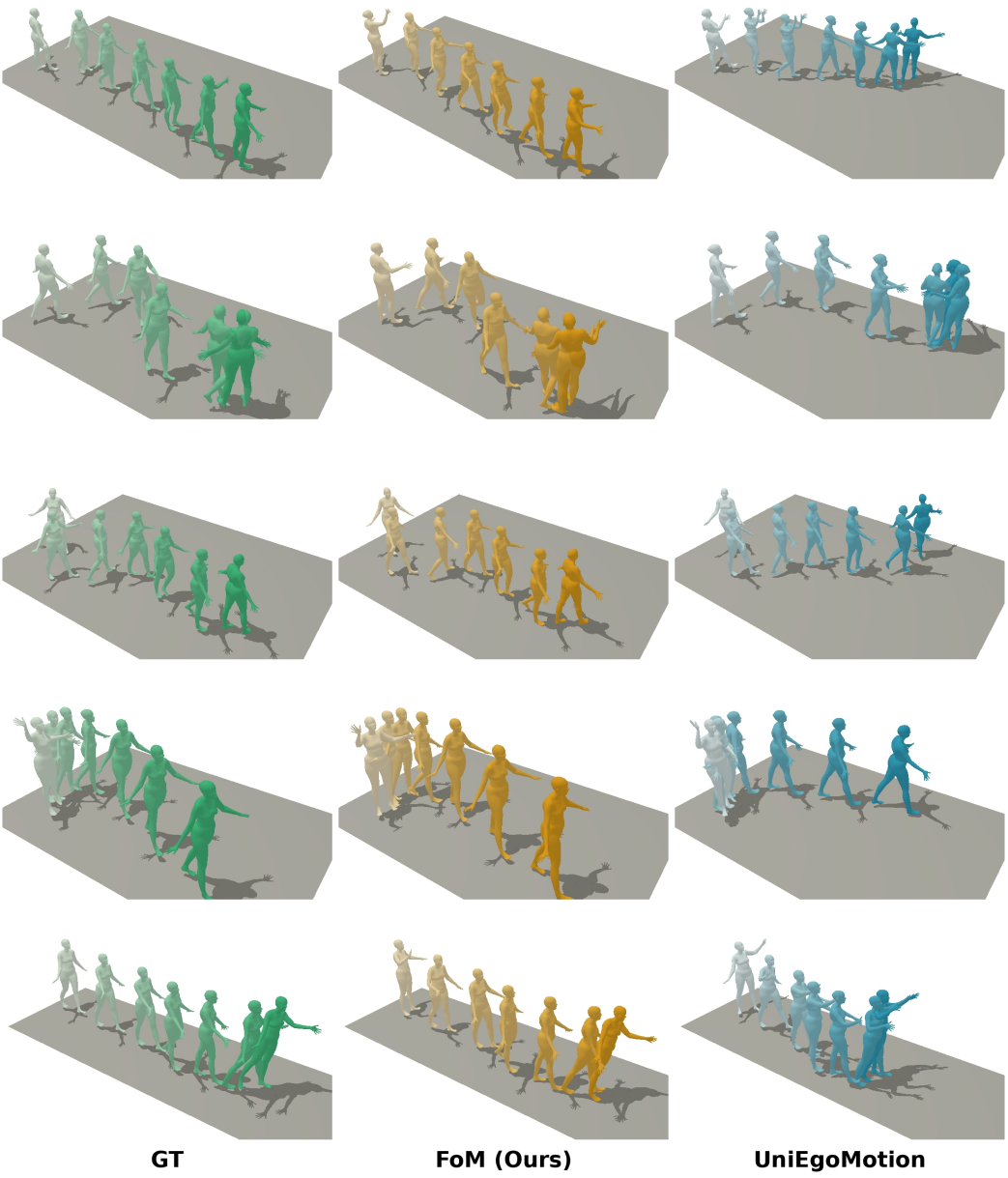

Figure S1: Additional qualitative comparison between ours FOM and UniEgoMotion.

A.5 METRICS

Positions are in metres. A sequence of length T is scored by averaging a per-frame residual; the number in the paper is the mean of those sequence scores over validation clips.

A.5.1 HEAD ROTATION (H-ROT)

Let R_t be the global rotation of the head joint (SMPL-X joint 15), obtained by composing local pose along the kinematic tree. The per-frame residual is

$$e_t^{\text{rot}} = \left\| {R_t^{\text{pred}}}^{-1} R_t^{\text{gt}} - I \right\|_F. \tag{S1}$$

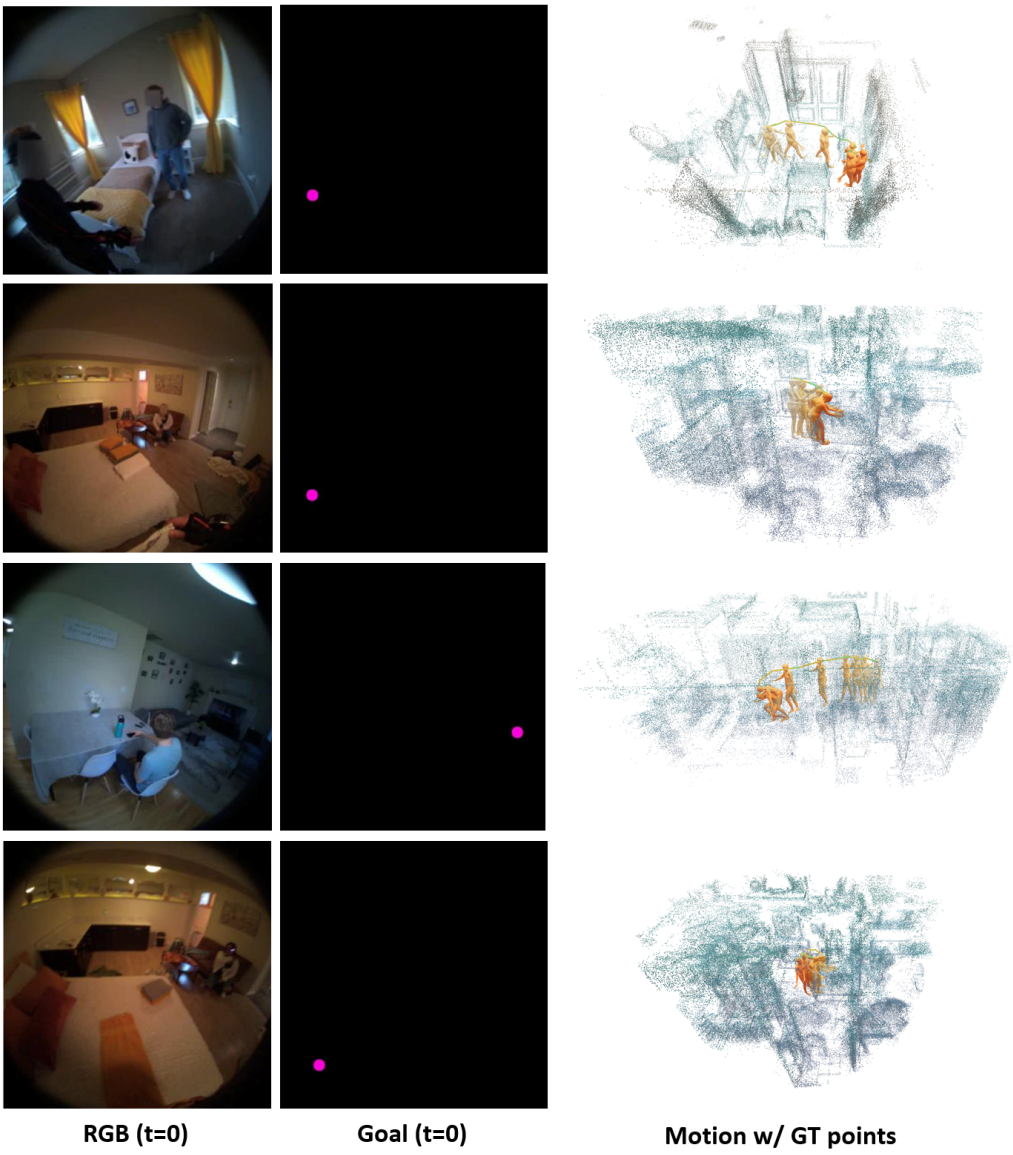

Figure S2: Additional samples for human motion placed in ground truth pointcloud.

If that relative rotation has angle θ_t, then $e_t^{\text{rot}} = 2\sqrt{2}\,|\sin(\theta_t/2)|$. The sequence score is $\frac{1}{T}\sum_t e_t^{\text{rot}}$.

A.5.2 Head translation (H-Trans)

The head point is the left-eye joint (index 23):

$$e^{\text{trans}} = \frac{1}{T}\sum_{t=1}^{T} \left\| x_t^{\text{pred}} - x_t^{\text{gt}} \right\|_2. \tag{S2}$$

A.5.3 MPJPE

For a chosen joint set $\mathcal{J}$,

$$\text{MPJPE}(\mathcal{J}) = \frac{1}{T|\mathcal{J}|}\sum_{t=1}^{T}\sum_{j\in\mathcal{J}} \left\| j_{t,j}^{\text{pred}} - j_{t,j}^{\text{gt}} \right\|_2. \tag{S3}$$

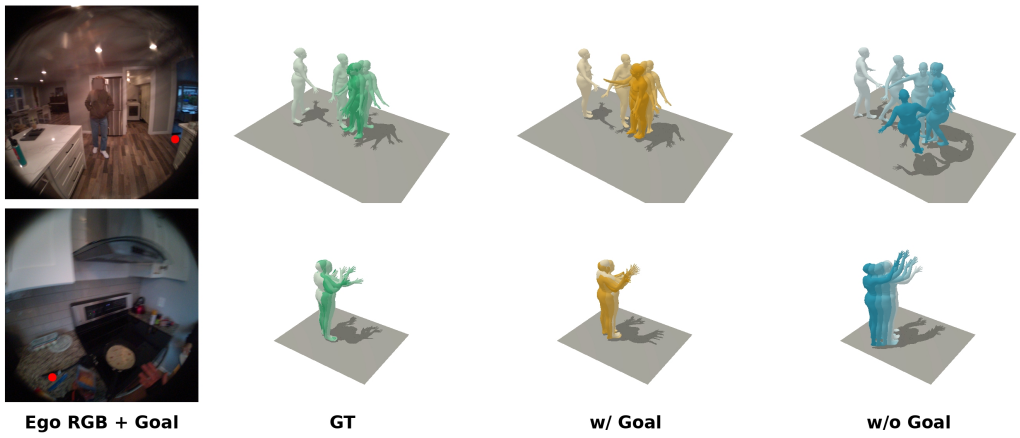

Figure S3: Comparison between prediction w/ goal mask and w/o goal mask.

A.5.4 MPJPE-PA

The same average after a per-frame similarity transform. At each t, scale s_t, rotation R_t, and translation t_t are chosen so $s_t R_t J_t^{\text{pred}} + t_t$ is as close as possible to J_t^{gt} in the least-squares sense. Joint error is then computed on that aligned pose.

Table S5: MPJPE variants. Slices are 0-based SMPL-X indices as in the eval code.

Name	Joints $\mathcal{J}$	Alignment
MPJPE	all SMPL-X joints	none
MPJPE-PA	all SMPL-X joints	per-frame Procrustes
MPJPE-Body	0:22 (body, incl. root)	none
MPJPE-H	25:55 (both hands)	none

The main table uses MPJPE, MPJPE-PA, MPJPE-Body, MPJPE-H, H-ROT and H-TRANS.

A.6 Limitations

Our current formulation does not explicitly model physical contact, either in the network design or through contact-aware supervision. As a result, predicted motions may exhibit body–object interpenetration during interaction. We also do not impose explicit constraints between the feet and the ground, which may lead to foot sliding or imperfect ground contact. In addition, the datasets used in this work primarily focus on indoor, interaction-centered scenarios. Extending the framework to more diverse environments, outdoor settings, and a broader range of human activities remains an important direction for future work.