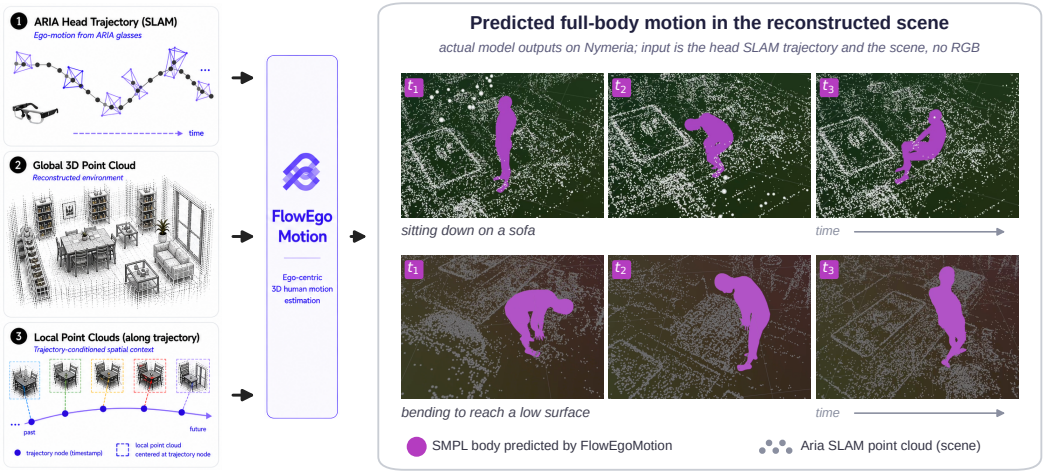
FLOWEGOMOTION: SCENE-CONDITIONED RECTIFIED FLOW FOR EGOCENTRIC FULL-BODY MOTION ESTIMATION

Anonymous authors
Paper under double-blind review



Head-mounted SLAM + scene → FlowEgoMotion (2-step rectified flow) → full-body motion in the scene

Figure 1: **Overview of FlowEgoMotion. Left (schematic):** the inputs from a head-mounted device, ① the SLAM head trajectory, ② the reconstructed 3D point cloud of the environment, and ③ local point clouds cropped along the trajectory; no RGB images are used. **Right (actual predictions):** FlowEgoMotion outputs on Nymeria, rendered in the reconstructed Aria SLAM point cloud. The generated body follows the head trajectory and adapts its articulation to the scene, e.g. when sitting down on a sofa or bending to a low surface.

ABSTRACT

Estimating full-body motion from a head-mounted egocentric device is severely underconstrained: visual–inertial SLAM gives reliable egomotion estimation, but most of the wearer's body is never directly visible, and prior diffusion-based approaches incur high inference cost while underusing scene context. We present **FlowEgoMotion**, a scene-conditioned rectified-flow model that combines the SLAM head trajectory with two complementary scene pathways – time-aligned local geometry fused into the egomotion stream, and global tokens encoding room-scale structure – to condition a Transformer that predicts a motion vector field, while global root motion is recovered analytically from the trajectory. On Nymeria, a controlled comparison with the same backbone shows that scene conditioning reduces MPJPE from 86.7 to 75.6 mm, Wrist-PE from 224.8 to 147.3 mm, and floor penetration from 13.5 to 0.5 mm, and the gains carry over to ADT without fine-tuning. FlowEgoMotion also improves on the closest published scene-aware system, HMD^2, on the same benchmark, reducing MPJPE from 83.6 to 75.6 mm, Wrist-PE from 166.4 to 147.3 mm, and floor penetration from 10.3 to 0.5 mm without using the RGB features that HMD^2 relies on. With only two flow steps, it needs about 12.5 ms of GPU compute per second of generated motion and is more accurate than a matched 30-step diffusion model at 15× lower cost. Code will be released upon acceptance.

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.

1 INTRODUCTION

Estimating full-body human motion from a wearable egocentric device is a fundamental capability for AR/VR, telepresence, assistive technologies, sports analysis, and human–robot interaction (Li et al., 2026; Guzov et al., 2025; Patel et al., 2025; Wang et al., 2025). Head-mounted systems already provide two highly informative signals for this task: metric egomotion estimation from visual-inertial SLAM and observations of the surrounding environment from onboard cameras or reconstructed scene geometry (Ma et al., 2024; Pan et al., 2023). Yet recovering body motion from these signals remains severely underconstrained: most of the body lies outside the field of view, hand visibility is intermittent, and similar head trajectories can correspond to substantially different articulated motions (Yi et al., 2025). Egocentric motion estimation is therefore better formulated as structured conditional generation under ambiguity than as deterministic pose regression.

Prior work has tackled this setting in two main directions: image-centric estimation that exploits visible body and scene evidence (Wang et al., 2023; Yang et al., 2024; Lee et al., 2025; Li et al., 2026), and sparse-sensing estimation from compact wearable cues such as head motion or SLAM (Li et al., 2023; Jiang et al., 2024; Yi et al., 2025; Cho & Joo, 2025). Despite strong progress, two challenges remain. First, leading egocentric motion systems are diffusion-based and often rely on guidance or post-hoc refinement, inflating inference cost (Li et al., 2023; Yi et al., 2025; Guzov et al., 2025; Patel et al., 2025; Cho & Joo, 2025). Second, many existing approaches are trained on synthetic or pseudo-ground-truth data, such as motion datasets with simulated device trajectories or automatically reconstructed annotations (Li et al., 2023; Yi et al., 2025; Patel et al., 2025). While effective for learning generic motion priors, these settings do not fully capture the noise, scene complexity, and body–environment interactions present in real wearable data, making it unclear which modeling choices transfer best in practice.

Recent datasets such as Nymeria (Ma et al., 2024), provide large-scale real-world egocentric motion with synchronized SLAM and scene observations, enabling direct evaluation of these design choices under realistic conditions. At the same time, scene-aware methods (Wang et al., 2023; Guzov et al., 2025; Patel et al., 2025) suggest that the information missing from head trajectories is strongly structured by the environment—layout, nearby surfaces, free space—motivating a model that treats scene context as a critical conditioning signal alongside SLAM dynamics rather than as a post-hoc correction.

We present **FlowEgoMotion**, a scene-conditioned rectified-flow model for egocentric full-body motion estimation with few-step efficient inference. The model learns a continuous-time transport from noise to articulated motion trajectories conditioned on three complementary cues: (i) an invariant representation of the SLAM trajectory, (ii) local scene features that provide time-aligned geometric evidence around the wearer, and (iii) global scene features that summarize room-scale structure and affordances. These cues serve as conditional inputs to Transformer blocks that predict a motion vector field. At the generative level, we adopt a conditional rectified flow (Lipman et al., 2023; Liu et al., 2023) framework, enabling accurate few-step sampling without any test-time optimization or post-processing guidance.

We evaluate FlowEgoMotion on the real-world egocentric benchmarks Nymeria (Ma et al., 2024) and ADT (Pan et al., 2023), and on the synthesized benchmarks AMASS (Mahmood et al., 2019) and RICH (Huang et al., 2022). On Nymeria, FlowEgoMotion improves on the published results of HMD^2 (Guzov et al., 2025), the closest published scene-aware system, both with local geometry alone and with full conditioning, without using images. Against a matched scene-unaware model with the same backbone, scene conditioning improves every metric on both real datasets, and a matched diffusion counterpart isolates the benefit of rectified flow.

In summary, our contributions are:

- **Factorized scene conditioning for egocentric motion.** We show that local time-aligned geometry and room-scale global context are complementary rather than interchangeable conditioning signals, and that injecting them through different sub-networks—a gated residual into the SLAM stream for local geometry, and cross-attention plus AdaLN for global context—yields larger gains than either pathway alone (Sec. 4.2).
- **Few-step rectified flow for egocentric motion.** A conditional rectified flow trained with a logit-normal time schedule and a forward-kinematics auxiliary loss enables two-step ODE

inference without test-time guidance or refinement; a matched DDIM model with identical architecture and conditioning is less accurate even with 30 steps and $\sim 15\times$ more compute (Sec. 4.3).

- **Strong results on real wearable data.** On Nymeria, FlowEgoMotion improves on HMD^2 with matched local-geometry inputs and with full conditioning (MPJPE $83.6 \rightarrow 75.6$ mm, Floor Pen $10.3 \rightarrow 0.5$ mm) while needing only ≈ 12.5 ms of GPU compute per second of motion; scene conditioning improves all metrics over a matched scene-unaware model on Nymeria and, zero-shot, on ADT (Sec. 4).

2 Related Work

Egocentric pose and motion from visual observations. A major line of work estimates 3D pose, mesh, or motion directly from egocentric images or videos. SceneEgo introduced explicit scene reasoning for pose prediction (Wang et al., 2023). Subsequent image-centric systems extend this setting to stereo, truncated-view mesh recovery, whole-body fisheye capture, event cameras, and real-time inference, including EgoPoseFormer (Yang et al., 2024), egocentric stereo pose estimation (Akada et al., 2024), EgoHMR (Zhang et al., 2023), fisheye whole-body capture (Wang et al., 2024), EventEgo3D (Millerdurai et al., 2024), REWIND (Lee et al., 2025), EgoPoseFormer v2 (Li et al., 2026), and rear-camera egocentric pose estimation (Akada et al., 2025). These methods exploit direct visual evidence of the body and scene, whereas our setting focuses on SLAM-conditioned generative reconstruction when body visibility is sparse or absent.

Motion estimation from SLAM head trajectories and scene cues A complementary line reconstructs full-body motion from compact wearable cues such as head motion, SLAM trajectories, or sparse tracked devices (Li et al., 2023; Du et al., 2023; Dai et al., 2024; Jiang et al., 2024; Yi et al., 2025; Cho & Joo, 2025; Boscolo Camiletto et al., 2025). Scene-aware variants further show that environmental context reduces ambiguity (Wang et al., 2023; Guzov et al., 2025; Patel et al., 2025; Tang et al., 2024). Our work is closest to this scene-aware family. HMD^2 (Guzov et al., 2025) conditions an autoregressive diffusion model on a local SLAM point cloud and CLIP image features; we instead factorize scene conditioning into a time-aligned local stream and a room-scale global stream, use no images, and replace iterative diffusion with a two-step flow.

Real-world supervision for egocentric motion. Data availability has shaped this literature: EgoEgo decomposes the task to avoid paired video/motion supervision (Li et al., 2023), EgoAllo trains on synthesized device poses from AMASS (Yi et al., 2025), and broader resources include Ego-Humans (Khirodkar et al., 2023), ADT (Pan et al., 2023), Ego-Exo4D (Grauman et al., 2024), EE4D-Motion (Patel et al., 2025), and EgoWholeBody (Wang et al., 2024). Among these, Nymeria (Ma et al., 2024) is the most relevant: it provides real wearable streams, full-body ground truth, and metric world alignment. Orthogonally, Flow Matching and Rectified Flow (Lipman et al., 2023; Liu et al., 2023) provide continuous-time alternatives to diffusion, and recent motion-synthesis work (Hu et al., 2023; Cho et al., 2025) shows they can match diffusion quality in far fewer function evaluations. We bring this perspective to egocentric motion estimation, where existing scene-aware SLAM-conditioned systems remain largely diffusion-based.

3 Method

Our method is built around three principles: (i) the output should remain exactly consistent with the observed SLAM trajectory, so the model focuses on ambiguous articulation rather than re-estimating global motion; (ii) scene information should enter through two complementary pathways, a time-aligned local stream and a room-scale global stream; and (iii) the generative backbone should support low-latency inference, motivating a rectified-flow formulation with very few steps. An overview of the full pipeline is shown in Fig. 2.

3.1 Problem setup and conditioning representations

Given a temporal window of length T, we observe a device trajectory $\mathcal{T}_D := \{\mathbf{T}^W_{D,t}\}_{t=1}^T$ with $\mathbf{T}^W_{D,t} \in SE(3)$, where D and W denote the wearable device and world frames, together with a scene

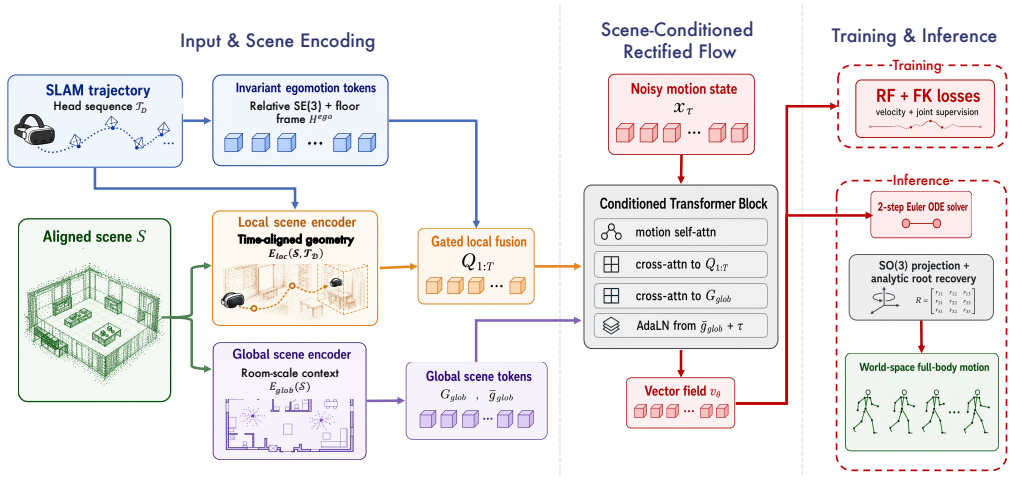


Figure 2: **Overview of FlowEgoMotion.** Invariant egomotion tokens derived from SLAM poses $\mathcal{T}_D$ are fused with a local, trajectory-aligned scene stream into $\mathbf{Q}_{1:T}$, while a global stream yields context tokens $\mathbf{G}^{\text{glob}}$ and a pooled summary $\bar{\mathbf{g}}^{\text{glob}}$. Stacked Transformer blocks (self-attention; cross-attention to $\mathbf{Q}_{1:T}$ and $\mathbf{G}^{\text{glob}}$; AdaLN driven by $\bar{\mathbf{g}}^{\text{glob}}$ and sample time τ) and a decoder predict the rectified-flow field $\mathbf{v}_\theta(\mathbf{x}_\tau)$. Trained with a rectified-flow velocity loss $\mathcal{L}_{\text{RF}}$ and a forward-kinematics joint-position loss $\mathcal{L}_{\text{FK}}$; inference uses a 2-step Euler ODE with analytic root recovery.

representation $\mathcal{S}$ (a reconstructed 3D scene or point cloud) aligned with the same frame. Our goal is to estimate a full-body motion sequence $\mathbf{y}_{1:T}$, $\mathbf{y}_t \in \mathbb{R}^{d_{\text{state}}}$, conditioned on $(\mathcal{T}_D, \mathcal{S})$. Each state $\mathbf{y}_t$ packs $d_{\text{state}} = 226$ values: a 16-dimensional body-shape vector $\boldsymbol{\beta}$ (shared across the sequence and replicated per frame for notational convenience), 21 body-joint rotation matrices in flattened 3×3 form, and 21 contact logits. To learn in Euclidean space, we concatenate all motion variables into a vector and project predicted rotations back to $SO(3)$ via polar projection at decoding. Contact logits are supervised through the rectified-flow term and used only as auxiliary outputs at inference; they do not enter the forward-kinematics loss.

The model predicts *local* articulation rather than global root motion. Global placement is recovered analytically from the observed device trajectory through forward kinematics. Let R denote the body root frame and H the body head frame. Let $\mathbf{T}_H^D(\boldsymbol{\beta})$ denote the rigid transform from the body head to the device, and let $\mathbf{T}_{R,t}^H(\mathbf{y}_t, \boldsymbol{\beta})$ denote the head transform induced by forward kinematics under the predicted local articulation. The world-space root transform is

$$\mathbf{T}_{R,t}^W = \mathbf{T}_{D,t}^W \left(\mathbf{T}_H^D(\boldsymbol{\beta})\right)^{-1} \left(\mathbf{T}_{R,t}^H(\mathbf{y}_t, \boldsymbol{\beta})\right)^{-1}. \tag{1}$$

This factorization enforces exact consistency with the measured SLAM trajectory and lets the model focus on the genuinely ambiguous part of the problem: articulated motion.

Invariant egomotion representation. World-frame device poses carry nuisance variability (absolute position and heading) that does not predict articulation. We therefore adopt the invariant encoding of Yi et al. (2025) unchanged, so that downstream gains can be attributed to our generative formulation and scene conditioning. For each t we use the relative motion $\Delta\mathbf{T}_t = (\mathbf{T}_{D,t-1}^W)^{-1}\mathbf{T}_{D,t}^W$ and the device pose $\mathbf{T}_{D,t}^{\hat{D}_t}$ expressed in a canonical frame $\hat{D}_t$ obtained by projecting the device origin onto the floor (z=0, $+z$ up) and keeping only the device yaw, which removes absolute position and global yaw while preserving pitch, roll, and height:

$$\mathbf{u}_t := \text{pack}\big(\Delta\mathbf{T}_t,\ \mathbf{T}_{D,t}^{\hat{D}_t}\big) \in \mathbb{R}^{12}, \qquad \mathbf{c}_t := \Phi(\mathbf{u}_t), \tag{2}$$

where pack concatenates the $SE(3)$ logarithms (twist coordinates) of both transforms and Φ applies K=6 sinusoidal Fourier bands followed by a two-layer MLP to dimension d.

A token-wise linear encoder produces aligned egomotion tokens (temporal mixing happens later, inside the vector-field Transformer)

$$\mathbf{H}^{\text{ego}} := E_{\text{ego}}(\mathbf{c}_{1:T}), \qquad \bar{\mathbf{h}}^{\text{ego}} := \text{Pool}_T\big(\mathbf{H}^{\text{ego}}\big), \tag{3}$$

where $\mathbf{H}^{\text{ego}} \in \mathbb{R}^{T\times d}$, $\mathbf{H}_t^{\text{ego}} \in \mathbb{R}^d$ denotes the t-th egomotion token, and $\bar{\mathbf{h}}^{\text{ego}} \in \mathbb{R}^d$ is the pooled sequence summary. This representation preserves motion cues that are informative for articulation while discarding arbitrary world-frame dependence.

Dual-stream scene conditioning. The scene-conditioning mechanism is built around two complementary pathways. The first pathway injects *local*, time-aligned scene evidence directly into the egomotion stream. The second pathway supplies *global*, room-scale scene context to every Transformer block. The separation is deliberate: local evidence resolves near-body geometric ambiguity, while global context provides a higher-level prior over layout, affordances, and spatial structure.

Local scene-to-egomotion fusion. For each timestep, we extract a local scene descriptor aligned to the wearer trajectory,

$$\mathbf{l}_{1:T} := E_{\text{loc}}(\mathcal{S}, \mathcal{T}_D), \qquad \mathbf{l}_t \in \mathbb{R}^{d_{\text{loc}}}. \tag{4}$$

We then project egomotion and local scene features into a common latent space and fuse them through a gated residual:

$$\tilde{\mathbf{h}}_t^{\text{ego}} := \mathbf{P}_{\text{ego}}\,\mathbf{H}_t^{\text{ego}}, \qquad \tilde{\mathbf{l}}_t := \mathbf{P}_{\text{loc}}\,\mathbf{l}_t, \tag{5}$$

$$\boldsymbol{\delta}_t := \mathbf{W}_\delta\,\tilde{\mathbf{l}}_t, \qquad \boldsymbol{\alpha}_t := \sigma\big(\mathbf{W}_\alpha\,\tilde{\mathbf{l}}_t + \mathbf{b}_\alpha\big), \qquad \mathbf{q}_t := \tilde{\mathbf{h}}_t^{\text{ego}} + \boldsymbol{\alpha}_t \odot \boldsymbol{\delta}_t, \tag{6}$$

where $\boldsymbol{\alpha}_t \in \mathbb{R}^d$ is a learned gating vector. This treats local geometry as a trajectory-aligned refinement of the motion condition rather than as a separate side modality. For brevity, we denote the fused local-egomotion sequence by $\mathbf{Q}_{1:T} := (\mathbf{q}_t)_{t=1}^T$.

Global scene context. To capture room-scale layout and affordance structure, we encode the scene into a sequence of global context tokens

$$\mathbf{G}^{\text{glob}} := E_{\text{glob}}(\mathcal{S}) = \big(\mathbf{g}_m^{\text{glob}}\big)_{m=1}^M, \qquad \boldsymbol{\mu} \in \{0,1\}^M, \tag{7}$$

where $\mathbf{g}_m^{\text{glob}} \in \mathbb{R}^{d_{\text{glob}}}$ and $\boldsymbol{\mu}$ is a per-token validity mask that handles the variable number of available tokens across scenes (zero-padded up to $M = 512$). We also compute a pooled sequence-level summary

$$\bar{\mathbf{g}}^{\text{glob}} := \text{MaskPool}\big(\mathbf{G}^{\text{glob}}, \boldsymbol{\mu}\big) \in \mathbb{R}^{d_{\text{glob}}}. \tag{8}$$

The token sequence $\mathbf{G}^{\text{glob}}$ is injected into each Transformer block through dedicated cross-attention with key-padding mask $\boldsymbol{\mu}$, while the pooled scene summary $\bar{\mathbf{g}}^{\text{glob}}$ modulates the network through adaptive normalization. During training, we apply scene dropout so the model remains robust when scene cues are noisy or unavailable; exact initialization and dropout details are given in App. D.

3.2 Scene-conditioned rectified flow

We model motion generation as conditional transport from a Gaussian base distribution to $p(\mathbf{y}_{1:T} \mid \mathcal{T}_D, \mathcal{S})$. With a base sample $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ in packed motion space, a clean sample $\mathbf{y} \sim p_{\text{data}}(\cdot \mid \mathcal{T}_D, \mathcal{S})$, and transport time $\tau \in (0,1)$, we use the linear coupling and its target velocity

$$\mathbf{x}_\tau := (1-\tau)\mathbf{n} + \tau\,\mathbf{y}, \qquad \mathbf{v}^\star := \mathrm{d}\mathbf{x}_\tau/\mathrm{d}\tau = \mathbf{y} - \mathbf{n}. \tag{9}$$

We sample $\tau := \sigma(z)$ with $z \sim \mathcal{N}(0,1)$ (logit-normal (Esser et al., 2024)), which concentrates training near the middle of the path, where the interpolation is most ambiguous. The model learns a conditional vector field $\mathbf{v}_\theta(\mathbf{x}_\tau, \tau, \mathbf{Q}_{1:T}, \mathbf{G}^{\text{glob}}, \bar{\mathbf{g}}^{\text{glob}})$.

The primary training objective is a masked, modality-aware velocity regression loss

$$\mathcal{L}_{\text{RF}} = \mathbb{E}_{\mathbf{n},\mathbf{y},\tau}\Big[\big\|\mathbf{v}_\theta(\mathbf{x}_\tau, \tau, \cdot) - (\mathbf{y} - \mathbf{n})\big\|_{\boldsymbol{\Lambda}}^2\Big], \tag{10}$$

where $\|\mathbf{a}\|_{\boldsymbol{\Lambda}}^2 := \mathbf{a}^\top \boldsymbol{\Lambda}\,\mathbf{a}$ and $\boldsymbol{\Lambda}$ is a block-diagonal weighting matrix over the body-shape, body-rotation, and contact components.

Joint-space auxiliary supervision. To better align training with the downstream task, we add an auxiliary forward-kinematics loss. Following the one-step target estimator commonly used in consistency- and flow-style training (Song et al., 2023), we form a differentiable estimate of the clean sample from the predicted velocity,

$$\hat{\mathbf{y}} := \mathbf{x}_\tau + (1-\tau)\,\mathbf{v}_\theta(\mathbf{x}_\tau, \tau, \cdot), \tag{11}$$

and supervise posed joint locations in world coordinates:

$$\mathcal{L}_{\mathrm{FK}} = \sum_{t,j} \omega_t \left\| \Pi_j\big(\hat{\mathbf{y}}_t;\, \mathbf{T}^W_{D,t}, \boldsymbol{\beta}\big) - \Pi_j\big(\mathbf{y}_t;\, \mathbf{T}^W_{D,t}, \boldsymbol{\beta}\big) \right\|_2^2, \tag{12}$$

where $\Pi_j \colon \mathbb{R}^{d_{\mathrm{state}}} \times SE(3) \times \mathbb{R}^{16} \to \mathbb{R}^3$ returns the world-space position of joint j via forward kinematics, and $\omega_t \in \{0,1\}$ is a frame-validity mask. Gradients of $\mathcal{L}_{\mathrm{FK}}$ flow through $\mathbf{v}_\theta$, the $SO(3)$ projection, and the forward-kinematics layer. The final training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{RF}} + \lambda_{\mathrm{FK}}\,\mathcal{L}_{\mathrm{FK}}. \tag{13}$$

Vector-field network. The conditional vector field is parameterized by a continuous-time Transformer operating on noisy motion tokens. Given the interpolated state $\mathbf{x}_\tau$, we first obtain a latent motion sequence

$$\mathbf{Z}^{(0)} := E_{\mathrm{traj}}(\mathbf{x}_\tau), \tag{14}$$

and embed transport time with $\mathbf{e}_\tau := \phi(\tau)$. Together with the pooled fused-motion summary and pooled global scene summary, this forms a sequence-level modulation vector

$$\boldsymbol{\Upsilon} := f\big(\mathrm{Pool}_T(\mathbf{Q}_{1:T}),\, \bar{\mathbf{g}}^{\mathrm{glob}},\, \mathbf{e}_\tau\big). \tag{15}$$

Each Transformer block applies self-attention over motion tokens, cross-attention to the fused ego-motion / local-scene tokens $\mathbf{Q}_{1:T}$, cross-attention to the global scene tokens $\mathbf{G}^{\mathrm{glob}}$ with key-padding mask $\boldsymbol{\mu}$, and a feed-forward network, with AdaLN and residual gating conditioned on $\boldsymbol{\Upsilon}$. Detailed block equations are given in App. D. After L blocks, a decoder maps the final latent sequence $\mathbf{Z}^{(L)}$ to the vector-field prediction:

$$\mathbf{v}_\theta := D\big(\mathbf{Z}^{(L)}\big). \tag{16}$$

Inference. At inference, we encode $(\mathcal{T}_D, \mathcal{S})$ into $(\mathbf{Q}_{1:T}, \mathbf{G}^{\mathrm{glob}}, \bar{\mathbf{g}}^{\mathrm{glob}})$, draw an initial sample $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and solve the conditional ODE

$$\frac{\mathrm{d}\mathbf{x}_\tau}{\mathrm{d}\tau} = \mathbf{v}_\theta\big(\mathbf{x}_\tau,\, \tau,\, \mathbf{Q}_{1:T},\, \mathbf{G}^{\mathrm{glob}},\, \bar{\mathbf{g}}^{\mathrm{glob}}\big), \qquad \tau \colon 0 \to 1, \tag{17}$$

to obtain the final motion sample $\mathbf{y} = \mathbf{x}_1$. In our experiments, we use a first-order Euler solver with only two transport steps. This low-step regime is central to the practical motivation of FlowEgoMotion. For sequences longer than the training window, inference is performed on overlapping temporal windows and the outputs are blended to form a coherent full-length trajectory. The only deterministic post-processing steps are projection of predicted rotations back to $SO(3)$ and analytic recovery of global root motion via Eq. (1).

4 Experiments

We focus on the real-egocentric benchmarks **Nymeria** and **ADT**; synthesized-egocentric results on **AMASS** and **RICH** are in App. B.

Evaluation protocol and metrics. All models are trained on temporal windows of length $T = 128$. At test time, we evaluate on *full sequences*, using the overlapping-window fusion scheme from Sec. 3.2; unless otherwise stated, inference uses only **2 flow steps**. Because prior work differs substantially in sensing inputs and reported metric suites (Yi et al., 2025; Guzov et al., 2025; Patel et al., 2025; Cho & Joo, 2025), we report the geometric metrics that are most comparable across methods, all in millimeters: **MPJPE**, the mean per-joint Euclidean error in the world frame; **PA-MPJPE**, its variant after a per-timestep Procrustes alignment (Yi et al., 2025); **Wrist-PE**, MPJPE restricted to the wrists (equivalent to Hand-PE in Guzov et al. (2025)); and **Floor Pen**, the mean penetration

Table 1: Real egocentric benchmarks on Nymeria and ADT (mm). Nymeria rows are grouped by input modality, so each group is a matched-input comparison. †Reported by HMD^2 (Guzov et al., 2025) on Nymeria (converted from cm); HMD^2 uses the same Aria SLAM trajectories and aggregated point clouds, the same root recovery from SLAM, and the same metric definitions, with full-sequence evaluation on unseen subjects (App. A.1). Its main setting h=180 uses 3 s of look-ahead within 4 s windows, comparable to our 4.3 s windows; h=10 is its 0.17 s low-latency variant. Wrist-PE equals HMD^2's Hand-PE. For all methods, $\pm$ is the standard deviation over repeated evaluation runs.

Dataset	Method	Inputs	MPJPE↓	PA-MPJPE↓	Wrist-PE↓	Floor Pen↓
Nymeria	EgoEgo	SLAM	$172.4^{\pm 1.6}$	$154.5^{\pm 1.3}$	$356.8^{\pm 1.7}$	$22.3^{\pm 1.6}$
	Ours (motion only)	SLAM	$\mathbf{86.7}^{\pm \mathbf{0.2}}$	$\mathbf{77.9}^{\pm \mathbf{0.2}}$	$\mathbf{224.8}^{\pm \mathbf{3.5}}$	$\mathbf{13.5}^{\pm \mathbf{2.7}}$
	EgoAllo	SLAM + Image	$108.1^{\pm 1.9}$	$86.3^{\pm 1.7}$	$215.4^{\pm 1.8}$	$15.4^{\pm 1.2}$
	HMD^{2†} (PC only, w/o CLIP)	SLAM + Local scene	$89.7^{\pm 1.0}$	–	$203.8^{\pm 2.8}$	$9.9^{\pm 0.7}$
	Ours (local scene only)	SLAM + Local scene	$\mathbf{82.7}^{\pm \mathbf{0.2}}$	$\mathbf{72.6}^{\pm \mathbf{0.2}}$	$\mathbf{182.4}^{\pm \mathbf{3.6}}$	$\mathbf{4.3}^{\pm \mathbf{2.2}}$
	HMD^{2†} (h=10)	SLAM + Image + Local scene	$91.9^{\pm 0.5}$	–	$176.7^{\pm 2.1}$	$11.9^{\pm 0.4}$
	HMD^{2†} (h=180)	SLAM + Image + Local scene	$83.6^{\pm 0.8}$	–	$166.4^{\pm 2.1}$	$10.3^{\pm 0.6}$
	Ours (full)	SLAM + Local + Global scene	$\mathbf{75.6}^{\pm \mathbf{0.2}}$	$\mathbf{68.3}^{\pm \mathbf{0.2}}$	$\mathbf{147.3}^{\pm \mathbf{2.9}}$	$\mathbf{0.5}^{\pm \mathbf{0.1}}$
ADT	EgoEgo	SLAM	$182.6^{\pm 2.3}$	$153.9^{\pm 2.6}$	$328.5^{\pm 1.9}$	$18.2^{\pm 1.9}$
	EgoAllo	SLAM + Image	$155.1^{\pm 1.6}$	$129.3^{\pm 1.6}$	$287.2^{\pm 1.7}$	$12.7^{\pm 1.6}$
	Ours (motion only)	SLAM	$129.7^{\pm 1.7}$	$108.2^{\pm 1.7}$	$226.5^{\pm 1.9}$	$14.3^{\pm 1.7}$
	Ours (full)	SLAM + Local + Global scene	$\mathbf{112.2}^{\pm \mathbf{1.5}}$	$\mathbf{98.1}^{\pm \mathbf{1.4}}$	$\mathbf{193.9}^{\pm \mathbf{2.1}}$	$\mathbf{1.7}^{\pm \mathbf{1.1}}$

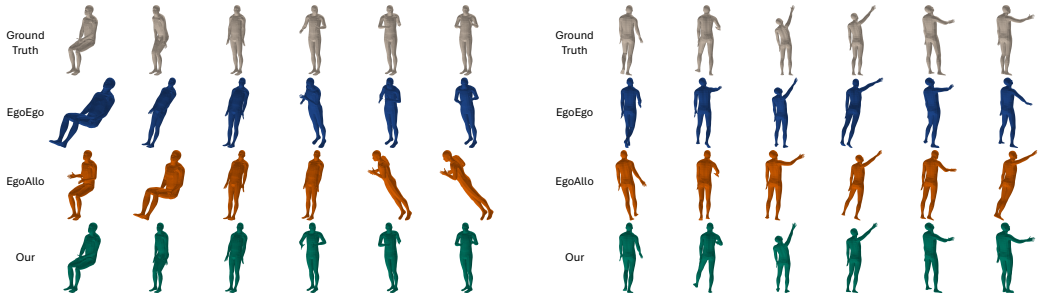


Figure 3: **Qualitative comparison on Nymeria.** Two representative test sequences showing SMPL reconstructions from EgoEgo (Li et al., 2023) (blue), EgoAllo (Yi et al., 2025) (orange), and FlowEgoMotion (Ours, green), against the ground truth (gray). EgoEgo, conditioned on SLAM alone, exhibits visible articulation drift in the limbs, while EgoAllo recovers a coarser posture but still shows foot–floor inconsistencies and inconsistent arm motion when the body leaves the field of view. FlowEgoMotion more faithfully reproduces the reference articulation and keeps the body grounded on the floor, reflecting the Floor-Pen reduction reported in Table 1.

depth below a per-frame floor proxy given by the lowest ground-truth joint over a short window (Guzov et al., 2025). We compare the full model against a matched *motion-only* rectified-flow baseline with the same backbone and training recipe, as well as **EgoEgo** and **EgoAllo** re-implemented in our evaluation stack. For **HMD^2**, whose code is not public, we use the numbers reported on Nymeria (see below). The local and global scene encoders are the lightweight point-cloud backbone **LitePT** (Yue et al., 2025) and the structured room-layout model **SceneScript** (Avetisyan et al., 2024).

4.1 Real egocentric benchmarks: Nymeria and ADT

Controlled effect of scene conditioning. Our rows in Table 1 share the backbone, training data, schedule, and inference procedure and differ only in the conditioning signal. Adding scene conditioning to the scene-unaware (motion-only) model improves every metric on Nymeria (MPJPE 86.7 → 75.6 mm, Wrist-PE 224.8 → 147.3 mm) and on ADT (MPJPE 129.7 → 112.2 mm), although ADT is never used for training and comes from an independent capture and reconstruction pipeline. Both of our models also outperform our re-implementations of EgoEgo and EgoAllo.

Table 2: Scene-pathway ablation on Nymeria. L-fuse denotes local LitePT features fused into the SLAM stream. G-xattn denotes global SceneScript cross-attention. G-AdaLN denotes pooled global-scene modulation. "Dropout" is the scene dropout probability p_{drop} (default 0.2). "wo AdaLN" / "wo xattn" isolate single global pathways; "wo $\mathcal{L}_{\text{FK}}$" removes the forward-kinematics auxiliary loss. The full configuration achieves the best result on every metric.

Variant	L-fuse	G-xattn	G-AdaLN	Dropout	MPJPE↓	PA-MPJPE↓	Wrist-PE↓	Floor Pen↓
No scene (motion only)	–	–	–	1.0	$86.7^{\pm 0.2}$	$77.9^{\pm 0.2}$	$224.8^{\pm 3.5}$	$13.5^{\pm 2.7}$
Local scene only	✓	–	–	0.2	$82.7^{\pm 0.2}$	$72.6^{\pm 0.2}$	$182.4^{\pm 3.6}$	$4.3^{\pm 2.2}$
Global scene only	–	✓	✓	0.2	$82.6^{\pm 0.2}$	$71.6^{\pm 0.2}$	$187.5^{\pm 3.2}$	$7.6^{\pm 2.3}$
Global scene w/o AdaLN	–	✓	–	0.2	$84.2^{\pm 0.2}$	$73.0^{\pm 0.2}$	$195.3^{\pm 2.8}$	$9.2^{\pm 2.4}$
Global scene w/o xattn	–	–	✓	0.2	$84.7^{\pm 0.2}$	$72.6^{\pm 0.2}$	$191.8^{\pm 3.1}$	$8.5^{\pm 2.1}$
No scene dropout	✓	✓	✓	0.0	$77.8^{\pm 0.2}$	$69.0^{\pm 0.2}$	$162.7^{\pm 3.1}$	$1.4^{\pm 0.3}$
High scene dropout	✓	✓	✓	0.3	$83.1^{\pm 0.2}$	$74.1^{\pm 0.2}$	$206.3^{\pm 3.3}$	$9.3^{\pm 2.1}$
FlowEgoMotion w/o $\mathcal{L}_{\text{FK}}$	✓	✓	✓	0.2	$80.3^{\pm 0.2}$	$71.6^{\pm 0.2}$	$176.6^{\pm 3.0}$	$2.3^{\pm 0.6}$
FlowEgoMotion (Full settings)	✓	✓	✓	0.2	$\mathbf{75.6^{\pm 0.2}}$	$\mathbf{68.3^{\pm 0.2}}$	$\mathbf{147.3^{\pm 2.9}}$	$\mathbf{0.5^{\pm 0.1}}$

Is the model actually using the scene? MPJPE averages over many joints and frames that the head trajectory already determines, so scene effects are clearest in contact-sensitive quantities. Relative to the motion-only model, the full model reduces floor penetration from 13.5 to 0.5 mm on Nymeria and from 14.3 to 1.7 mm on ADT, the fraction of ground-contact frames with a floating foot (>5 cm above the floor) from 17.1% to 3.6%, and the foot-skate ratio from 8.3% to 4.7% (definitions in App. A.4, following Han et al. (2025)). Conversely, corrupting *only the scene* at test time degrades accuracy (Table 3), so the predictions depend on the geometry rather than merely on extra capacity. Fig. 3 and App. B.2 show examples (sofa, stairs, bending) where the scene determines the support height or posture that scene-unaware baselines miss.

Comparison with HMD^2. HMD^2 (Guzov et al., 2025) is evaluated on the same benchmark under closely matching conditions. Both methods use Nymeria's on-device Aria SLAM trajectory and the semi-dense point cloud aggregated over each recording (HMD^2 likewise assumes access to the full-sequence SLAM points, with scene and image features precomputed), both predict local articulation and take the root from the SLAM trajectory, and both are evaluated on complete test sequences of unseen subjects with the same Hand-PE and Floor-Pen definitions. HMD^2's main setting (h=180) uses 3 s of future context within 4 s windows, comparable to our 4.3 s windows. We therefore compare at matched input levels (Table 1). With SLAM and local geometry only, our local-scene variant improves on HMD^2's point-cloud-only model on all three metrics (89.7 → 82.7 mm MPJPE, 203.8 → 182.4 mm Wrist-PE, 9.9 → 4.3 mm Floor Pen). With full conditioning, FlowEgoMotion improves on HMD^2 (h=180) by 9.6% in MPJPE, 11.5% in Wrist-PE, and 95% in Floor Pen, even though HMD^2 additionally uses CLIP image features. Remaining differences are listed in App. A.1.

4.2 How much does each scene pathway matter?

We ablate the scene-conditioning design on Nymeria with the backbone, training schedule, optimizer, and inference procedure fixed; variants differ only in the enabled scene channels.

Table 2 reveals a consistent pattern. Both *local-only* and *global-only* conditioning improve over the motion-only baseline, but they help in different ways. The **global** pathway is stronger for overall pose accuracy, reducing MPJPE to 82.6 mm and PA-MPJPE to 71.6 mm, whereas the **local** pathway is especially effective for physically grounded details, reducing Floor Pen from 13.5 mm to 4.3 mm and Wrist-PE from 224.8 mm to 182.4 mm. The partial global variants further show that neither cross-attention nor AdaLN alone recovers the full gain. The complete model achieves the best result on every metric. Relative to the strongest single-branch variant for each metric, it improves MPJPE by **8.5%**, PA-MPJPE by **4.6%**, Wrist-PE by **19.2%**, and Floor Pen by **88.4%**. This shows that local geometry and room-scale context are complementary, and the full gain emerges only when both are integrated into the generative model.

Two further design choices contribute meaningfully. Scene dropout at $p_{\text{drop}} = 0.2$ is the best operating point: disabling it degrades MPJPE from 75.6 to 77.8 mm and Floor Pen from 0.5 to 1.4 mm, while raising it to 0.3 hurts every metric. The forward-kinematics auxiliary loss is similarly crit-

	RF	DDIM				
Steps	2	2	5	10	20	30
MPJPE↓	**75.6**	96.4	88.3	82.7	79.1	78.2
Time (ms)↓	**40**	42	103	203	405	608

Test-time input	MPJPE↓	PA-MPJPE↓
Clean (full model)	75.6	68.3
SLAM noise (2.5 cm, 0.5°)	82.4	71.2
50% scene points removed	85.9	74.1
Scene offset (20 cm, 10°)	91.7	77.4
Motion-only model	86.7	77.9

Table 3: **Left:** rectified flow (RF) vs. a DDIM counterpart with identical Transformer, scene pathways, data, optimizer, auxiliary losses, and evaluation on Nymeria; time is per 128-frame window on one A100. **Right:** robustness of the full model to test-time corruption of the Nymeria inputs, without retraining (protocol in App. A.5).

ical: removing it ($\mathcal{L}_{\text{FK}}$=0) raises MPJPE by 4.7 mm and Wrist-PE by 29.3 mm, confirming that joint-position supervision is important in the few-step flow regime.

4.3 Rectified flow, runtime, sampling, and robustness

Rectified flow vs. diffusion. A DDIM (Song et al., 2021) counterpart differing only in objective and sampler (Table 3, left) reaches 96.4 mm MPJPE at two steps versus 75.6 mm for rectified flow, and even with 30 steps (78.2 mm) it remains less accurate while costing 15× more.

Runtime. We report throughput as GPU compute per second of generated motion, which is independent of frame rate. Under the same conditions as HMD^2 (one A100, scene features precomputed), the network generates a 128-frame window in 40 ms; with stride-96 windows at 30 Hz this is ≈12.5 ms per second of motion. From its reported throughput (>1350 and >70 frames/s at 60 Hz), HMD^2 needs under 44 ms (h=180) and under 860 ms (h=10). Adding our per-frame LitePT descriptors gives ≈18.5 ms, and SceneScript tokens add 2–3 s once per environment (App. A.2).

Stochastic sampling. With K=10 noise seeds per condition, the mean pairwise joint diversity is 26.8 mm, the mean-of-10 MPJPE is 75.9 mm (single sample: 75.6), and the best-of-10 MPJPE is 72.8 mm, so samples differ while staying accurate (App. A.3).

Robustness to imperfect inputs. Moderate SLAM noise and removing half of the scene points degrade accuracy gradually (Table 3, right). A large 20 cm/10° scene–trajectory misalignment is the most harmful, which confirms that the predictions rely on scene geometry. PA-MPJPE degrades less than MPJPE, consistent with the root being taken directly from SLAM.

5 Conclusion

We presented FlowEgoMotion, a scene-conditioned rectified-flow model that estimates full-body motion from a head-mounted SLAM trajectory and a reconstructed scene. Three design choices underpin it: global root motion is recovered analytically from the SLAM trajectory, so the generator models only the ambiguous articulation; scene information enters through two complementary pathways, time-aligned local geometry fused into the egomotion stream and room-scale tokens injected through cross-attention and AdaLN; and a rectified-flow objective enables accurate two-step sampling. On Nymeria, the full model improves on HMD^2 (MPJPE 83.6 → 75.6 mm, Floor Pen 10.3 → 0.5 mm) without using RGB. Compared with a matched scene-unaware model, the largest gains are in grounding and contact, with floating-foot frames dropping from 17.1% to 3.6%, and they carry over to ADT without fine-tuning. Two-step rectified flow is also more accurate than a 30-step diffusion counterpart at 15× lower cost. These results show that factorized scene conditioning resolves much of the ambiguity of egocentric motion, and that few-step flows make it practical.

Limitations and future work. Like HMD^2, FlowEgoMotion assumes an aligned SLAM map and several seconds of temporal context; a streaming variant, incrementally built maps, and confidence-aware scene weighting under large misalignment are natural extensions. Adding RGB for visible hands (App. A.6) and finger articulation is left to future work.

REPRODUCIBILITY STATEMENT

Architecture, hyperparameters, losses, and training and inference settings are given in App. D; dataset splits and preprocessing in App. C; metric definitions in Sec. 4 and App. A.4; and the runtime and robustness protocols in App. A. We will release code, checkpoints, cached scene descriptors, and evaluation scripts upon acceptance.

AI USE STATEMENT

Generative AI tools were used to assist with drafting and editing the manuscript, format LaTeX content, and support the design of scientific figures. All AI-assisted suggestions and outputs were reviewed by the authors for correctness, relevance, and originality. The authors take full responsibility for the final methodology, experiments, claims, and presentation.

REFERENCES

Hiroyasu Akada, Jian Wang, Vladislav Golyanik, and Christian Theobalt. 3d human pose perception from egocentric stereo videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 767–776, 2024.

Hiroyasu Akada, Jian Wang, Vladislav Golyanik, and Christian Theobalt. Bring your rear cameras for egocentric 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Armen Avetisyan, Christopher Xie, Henry Howard-Jenkins, Tsun-Yi Yang, Samir Aroudj, Suvam Patra, Fuyang Zhang, Duncan P. Frost, Luke Holland, Campbell Orme, Jakob J. Engel, Edward Miller, Richard A. Newcombe, and Vasileios Balntas. Scenescript: Reconstructing scenes with an autoregressive structured language model. In *European Conference on Computer Vision (ECCV)*, pp. 247–263, 2024.

Andrea Boscolo Camiletto, Jian Wang, Eduardo Alvarado, Rishabh Dabral, Thabo Beeler, Marc Habermann, and Christian Theobalt. Frame: Floor-aligned representation for avatar motion from egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17497–17507, 2025.

Jungbin Cho, Junwan Kim, Jisoo Kim, Minseo Kim, Mingu Kang, Sungeun Hong, Tae-Hyun Oh, and Youngjae Yu. DisCoRD: Discrete tokens to continuous motion via rectified flow decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14602–14612, 2025.

Kyungwon Cho and Hanbyul Joo. Hand-aware egocentric motion reconstruction with sequence-level context. *arXiv preprint arXiv:2512.19283*, 2025.

Peng Dai, Yang Zhang, Tao Liu, Zhen Fan, Tianyuan Du, Zhuo Su, Xiaozheng Zheng, and Zeming Li. Hmd-poser: On-device real-time human motion tracking from scalable sparse observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 874–884, 2024.

Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, and Artsiom Sanakoyeu. Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 481–490, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Kristen Grauman 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 (CVPR)*, pp. 19383–19400, 2024.

Vladimir Guzov, Yifeng Jiang, Fangzhou Hong, Gerard Pons-Moll, Richard A. Newcombe, C. Karen Liu, Yuting Ye, and Lingni Ma. HMD^2: Environment-aware motion generation from single egocentric head-mounted device. In *2025 International Conference on 3D Vision (3DV)*, pp. 1394–1405, 2025.

Gaoge Han, Mingjiang Liang, Jinglei Tang, Yongkang Cheng, Wei Liu, and Shaoli Huang. ReinDiffuse: Crafting physically plausible motions with reinforced diffusion model. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2218–2227, 2025.

Vincent Tao Hu, Wenzhe Yin, Pingchuan Ma, Yunlu Chen, Basura Fernando, Yuki M. Asano, Efstratios Gavves, Pascal Mettes, Bjorn Ommer, and Cees G. M. Snoek. Motion flow matching for human motion synthesis and editing. *arXiv preprint arXiv:2312.08895*, 2023.

Chun-Hao P Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J Black. Capturing and inferring dense full-body human-scene contact. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13274–13285, 2022.

Jiaxi Jiang, Paul Streli, Manuel Meier, and Christian Holz. EgoPoser: Robust real-time egocentric pose estimation from sparse and intermittent observations everywhere. In *European Conference on Computer Vision (ECCV)*, pp. 277–294, 2024.

Rawal Khirodkar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, and Kris Kitani. Egohumans: An ego-centric 3d multi-human benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19807–19819, 2023.

Jihyun Lee, Weipeng Xu, Alexander Richard, Shih-En Wei, Shunsuke Saito, Shaojie Bai, Te-Li Wang, Minhyuk Sung, Tae-Kyun Kim, and Jason M. Saragih. REWIND: Real-time egocentric whole-body motion diffusion with exemplar-based identity conditioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7095–7104, 2025.

Jiaman Li, C. Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17142–17151, 2023.

Zhenyu Li, Sai Kumar Dwivedi, Filip Maric, Carlos Chacon, 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.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

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 J. Engel, Renzo De Nardi, and Richard A. Newcombe. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *European Conference on Computer Vision (ECCV)*, pp. 445–465, 2024.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Christen Millerdurai, Hiroyasu Akada, Jian Wang, Diogo Luvizon, Christian Theobalt, and Vladislav Golyanik. Eventego3d: 3d human motion capture from egocentric event streams. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1186–1195, 2024.

Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar M. Parkhi, Richard A. Newcombe, and 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. 20133–20143, 2023.

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.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Jiangnan Tang, Jingya Wang, Kaiyang Ji, Lan Xu, Jingyi Yu, and Ye Shi. A unified diffusion framework for scene-aware human motion estimation from sparse signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21251–21262, 2024.

Jian Wang, Diogo C. Luvizon, Weipeng Xu, Lingjie Liu, Kripasindhu Sarkar, and Christian Theobalt. Scene-aware egocentric 3D human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13031–13040, 2023.

Jian Wang, Zhe Cao, Diogo Luvizon, Lingjie Liu, Kripasindhu Sarkar, Danhang Tang, Thabo Beeler, and Christian Theobalt. Egocentric whole-body motion capture with fisheyevit and diffusion-based motion refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 777–787, 2024.

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.

Chenhongyi Yang, Anastasia Tkach, Shreyas Hampali, Linguang Zhang, Elliot J. Crowley, and Cem Keskin. EgoPoseFormer: A simple baseline for stereo egocentric 3D human pose estimation. In *European Conference on Computer Vision (ECCV)*, pp. 401–417, 2024.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7072–7084, 2025.

Yuanwen Yue, Damien Robert, Jianyuan Wang, Sunghwan Hong, Jan Dirk Wegner, Christian Rupprecht, and Konrad Schindler. Litept: Lighter yet stronger point transformer. *arXiv preprint arXiv:2512.13689*, 2025.

Siwei Zhang, Qianli Ma, Yan Zhang, Sadegh Aliakbarian, Darren Cosker, and Siyu Tang. Probabilistic human mesh recovery in 3d scenes from egocentric views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7989–8000, 2023.

A Additional Analyses

A.1 Comparison protocol with HMD^2

HMD^2 (Guzov et al., 2025) is the closest published scene-aware system evaluated on Nymeria. Its weights, inference code, and per-sequence predictions are not public, so all HMD^2 numbers in Table 1 are taken from its paper (converted from cm to mm): the main table for h=180 and h=10, and its input ablation for the point-cloud-only (w/o CLIP) variant. Table 4 lists the evaluation conditions. The two protocols agree on the benchmark, the sensing pipeline, the use of a precomputed full-recording SLAM point cloud, the root-recovery paradigm, full-sequence evaluation on unseen subjects, the metric definitions, and the timing assumptions. The remaining differences are the extra modality each method uses (CLIP image features for HMD^2; room-level SceneScript tokens computed from the same point cloud for ours), the skeleton convention, and the exact split lists. The matched-input groups in Table 1 control for the first difference.

Table 4: Evaluation conditions of HMD^2 (from its paper and supplementary material) and FlowEgoMotion.

	HMD^2	FlowEgoMotion
Shared conditions		
Benchmark	Nymeria, real Aria recordings	Nymeria, real Aria recordings
Trajectory	On-device Aria (MPS) SLAM	On-device Aria (MPS) SLAM
Scene geometry	Semi-dense SLAM points of the full ≈15-min recording, precomputed	Semi-dense point cloud aggregated over the recording, precomputed
Root motion	Local rotations predicted; root stitched to the SLAM head	Local rotations predicted; root recovered from SLAM (Eq. (1))
Evaluation	Complete test sequences, unseen subjects	Complete test sequences, unseen subjects
Metrics	MPJPE, Hand-PE (wrists), Floor Pen (lowest-GT-joint proxy)	MPJPE, Wrist-PE (= Hand-PE), Floor Pen (same proxy)
Timing	One A100; point-cloud and CLIP features precomputed	One A100; scene features precomputed
Differences		
Extra modality	CLIP image embeddings	SceneScript room tokens (same point cloud)
Local encoder	2 m cube, voxel autoencoder	1.5 m sphere, frozen LitePT
Temporal context	4 s windows (60 Hz); 3 s look-ahead (h=180) or 0.17 s (h=10)	4.3 s windows (30 Hz), stride 96
Generator	Diffusion, 20 strided steps	Rectified flow, 2 steps
Skeleton	Xsens (23 joints)	SMPL-H body
Test split	224 sequences (56 h)	253 sequences (63 h)

A.2 Runtime breakdown

Table 5 decomposes the inference cost on a single NVIDIA A100 (40 GB). As in HMD^2, whose reported throughput assumes precomputed point-cloud and CLIP features, the generative stage is timed with scene features available. With stride-96 windows at 30 Hz, the network produces ≈2400 frames/s, i.e. ≈12.5 ms of compute per second of motion; adding LitePT descriptors for the 96 new frames of each window (19.2 ms) gives ≈1600 frames/s, or ≈18.5 ms per second of motion. SceneScript tokens are computed once per static environment (2–3 s, versus 900 s of motion in a 15-minute Nymeria recording). For reference, HMD^2 reports >1350 frames/s at 60 Hz for h=180 and >70 frames/s for h=10 on the same GPU type (under 44 and 860 ms per second of motion), and about 61 frames/s for h=10 when its features are computed on the same device. The global representation stays fixed while the environment is static, whereas the local descriptor is updated as the wearer moves.

Table 5: Inference cost by component. Scene reconstruction, data loading, window blending, and mesh rendering are not included.

Component	Frequency	Time
SceneScript global tokens	once per static environment	2–3 s
LitePT local descriptor (256 points)	per frame	$\approx$0.2 ms (25.6 ms per window)
Vector-field network, 2 Euler steps	per 128-frame window	40 ms

A.3 Multiple samples

The ODE solver is deterministic once the initial noise $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is fixed, so different seeds produce different articulated motions for the same trajectory and scene, while global placement stays constrained by the observed head trajectory. Drawing $K{=}10$ samples per condition on Nymeria gives an average pairwise body-local joint diversity of 26.8 ± 0.42 mm, a mean-of-10 MPJPE of 75.9 ± 0.16 mm, and a best-of-10 MPJPE of 72.8 ± 0.18 mm, compared with 75.6 mm for a single sample (Table 1); here $\pm$ is the standard error over test sequences. The lower best-of-10 error indicates that the sample set contains motions closer to the ground truth. These numbers show sample diversity; they do not establish calibrated uncertainty or failure detection.

A.4 Physical-plausibility metrics

Floor Pen measures only below-floor penetration. Following prior evaluations of floating and skating artifacts (Han et al., 2025), we additionally report (i) the *floating-foot ratio*, the percentage of ground-truth contact frames in which the predicted lowest foot joint lies more than 5 cm above the floor, and (ii) the *skate ratio*, the percentage of frames in which either foot slides more than 2.5 cm while in contact with the ground. Table 6 compares the matched motion-only model and the full model on Nymeria. Together with GND (Table 7), these metrics cover penetration, floating, and skating; they do not measure contact timing or interpenetration with non-floor geometry.

Table 6: Grounding and physical plausibility on Nymeria (lower is better).

Model	Floor Pen (mm)	Floating-foot ratio (%)	Skate ratio (%)
Motion only (no scene)	13.5	17.1	8.3
Full model	**0.5**	**3.6**	**4.7**

A.5 Robustness protocol

All perturbations are applied to the Nymeria test inputs of the trained full model, without retraining. *SLAM noise:* zero-mean Gaussian noise is added to every device pose, with standard deviation 2.5 cm in translation and 0.5° in rotation. *Partial scene:* 50% of the scene points are removed uniformly at random. *Misalignment:* the scene is rigidly offset from the trajectory by 20 cm in translation and 10° in yaw. Results are in Table 3 (right). The evaluation already uses real on-device Project Aria SLAM and reconstructions rather than idealized trajectories, and ADT (Table 1) provides an independent sensing and reconstruction pipeline on which the model is evaluated without fine-tuning. Training-time scene dropout also reduces over-reliance on scene features (Table 2). Under large misalignment the full model is worse than the motion-only model, which motivates confidence-aware scene weighting or a fallback to motion-only conditioning.

A.6 Geometry versus RGB

FlowEgoMotion does not use RGB images; the egocentric images in Fig. 4 are shown for visualization only. We view the two modalities as complementary: images directly observe visible body parts and manipulated objects, whereas geometry constrains motion through nearby surfaces, support height, free space, and layout. This is consistent with prior scene-aware work (Wang et al., 2023; Tang et al., 2024), and HMD^2 reports that point-cloud conditioning mainly improves scene alignment and sitting, while image features mainly help visible-hand and upper-body localization (Guzov

Table 7: Synthesized-egocentric benchmarks under SLAM-only conditioning. AMASS and RICH do not contain native egocentric data; device trajectories are synthesized following the prior SLAM-conditioned protocol (Yi et al., 2025). Baseline values for EgoEgo and EgoAllo follow the published benchmark. **GND** is a per-trajectory binary grounding indicator from Yi et al. (2025) that flags whether the feet ever touch the ground plane, capturing the "floating" failure mode of ego-sensed motion.

Dataset	Method	Inputs	MPJPE↓	PA-MPJPE↓	GND↑
AMASS	EgoEgo	SLAM	167.4 ± 2.1	145.8 ± 2.0	0.92 ± 0.01
	EgoAllo	SLAM	119.7 ± 1.3	101.1 ± 1.3	1.00 ± 0.00
	Ours	SLAM	$\mathbf{104.5 \pm 1.3}$	$\mathbf{80.6 \pm 1.2}$	1.00 ± 0.00
RICH	EgoEgo	SLAM	207.8 ± 6.9	187.8 ± 6.8	0.88 ± 0.03
	EgoAllo	SLAM	176.2 ± 5.6	160.1 ± 5.9	0.96 ± 0.02
	Ours	SLAM	$\mathbf{142.4 \pm 2.5}$	$\mathbf{129.5 \pm 2.9}$	$\mathbf{0.97 \pm 0.01}$

et al., 2025). Even without RGB, geometry reduces Wrist-PE by 34.5% relative to the motion-only model ($224.8 \rightarrow 147.3$ mm). A breakdown by wrist visibility, and fusing RGB with our scene pathways, are left to future work.

B Qualitative Results and Additional Experiments

B.1 Synthesized-egocentric benchmarks

While the main paper evaluates the full scene-conditioned model on Nymeria and ADT, we additionally evaluate the SLAM-only variant of FlowEgoMotion on AMASS and RICH. These datasets do not contain native egocentric sensing; device trajectories are synthesized following the SLAM-conditioned protocol of Yi et al. (2025). We therefore use them as supporting evidence for the underlying motion prior rather than as the main scene-aware result.

Table 7 shows that the SLAM-only FlowEgoMotion prior remains stronger than prior motion-only baselines even without scene input. On AMASS, it improves over EgoAllo by 15.2 mm in MPJPE and 20.5 mm in PA-MPJPE. On RICH, it improves over EgoAllo by 33.8 mm in MPJPE and 30.6 mm in PA-MPJPE, while slightly improving grounding. These results support the interpretation of the main paper: the gains on Nymeria and ADT come from explicit scene reasoning, but they are built on a SLAM-conditioned prior that is already stronger than prior baselines.

B.2 Qualitative Results

We provide additional qualitative results to complement the quantitative findings in Sec. 4. The sequences are deliberately drawn from activities that stress different aspects of the egocentric estimation problem: locomotion across height changes, transitions between standing and seated poses, and bending or reaching motions where most of the body is out of view. For each activity we describe the specific source of ambiguity, then point to the corresponding qualitative behavior in Fig. 4–6.

Descending stairs. Stair descent is one of the hardest scenarios in Nymeria. The wearer's head moves through a coarse, near-linear vertical trajectory, but the underlying gait alternates between supporting and swinging legs at each step, with foot placement happening entirely outside the field of view. A SLAM-only signal carries almost no information about which leg is loaded at any given instant, and a small error in foot height accumulates into visible floor penetration or floating. In the corresponding row of Fig. 5, FlowEgoMotion's predicted body remains tightly aligned with the ground truth across the descent, with feet landing on the correct step rather than clipping through the staircase geometry, consistent with the scene encoding the step heights.

Sitting down on a sofa. Sit-down transitions are severely underconstrained from head pose alone: the head descends along a smooth arc, but this is consistent with many distinct lower-body configurations (squatting, kneeling, sitting on a chair, sitting on the floor). The correct interpretation depends



Figure 4: **Scene-anchored predictions with the wearer's view.** Two challenging Nymeria scenarios shown across four timesteps (columns). **Top:** walking through a cluttered furnished room, where most of the wearer's body is out of view and arm motion must be inferred from the SLAM trajectory and scene context. **Bottom:** bending forward to reach a low surface, a motion that is locally indistinguishable from a deep squat in the head trajectory alone. The predicted SMPL body (magenta) is rendered inside the reconstructed scene point cloud; the synchronized egocentric image is shown as a small inset in the lower-left of each frame for visualization only; it is *not* a model input. In both scenarios, the body remains anchored to the floor and consistent with nearby geometry without any post-hoc refinement.

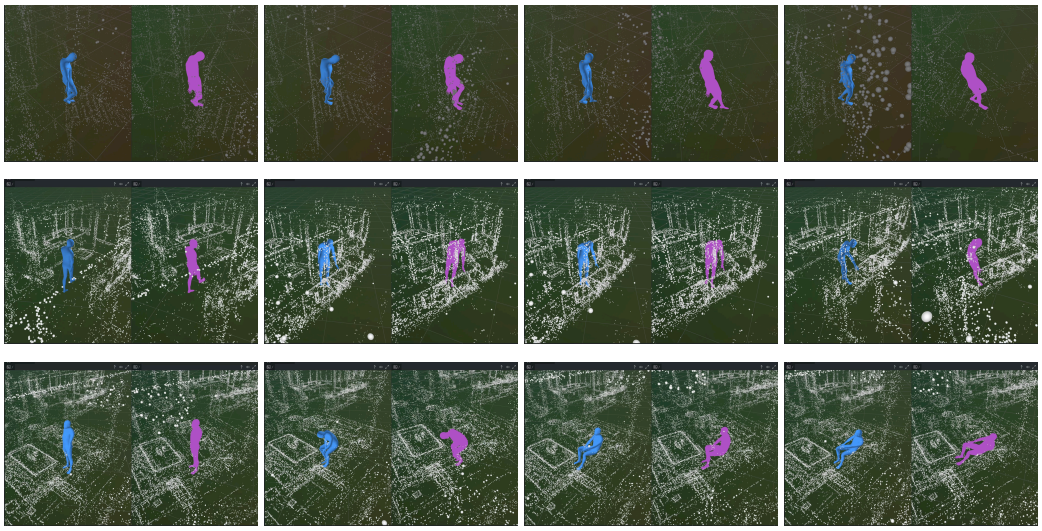


Figure 5: **Ground truth vs. prediction in the world frame.** Three Nymeria scenarios that stress different aspects of the estimation problem. **Top:** *descending stairs*, where foot placement is entirely out of view and small per-step height errors accumulate into visible penetration or floating. **Middle:** *walking through a cluttered room*, where the head trajectory is locally smooth and arm/torso articulation must be supplied by the prior. **Bottom:** *sitting down on a sofa*, a transition that is severely underconstrained from head pose alone and requires the surrounding furniture geometry to resolve. Each frame shows the ground-truth body (blue, left half) and the FlowEgoMotion prediction (magenta, right half) in the same world-aligned scene point cloud; global root motion of the prediction is recovered analytically from the SLAM trajectory (Eq. (1)).

on where in the room the descent occurs and what surface is nearby. In the corresponding row, the predicted body settles onto the sofa surface at the correct height, with the pelvis aligned to the seat rather than hovering above it or penetrating it, even though the sofa's height is not recoverable from the SLAM trajectory alone.

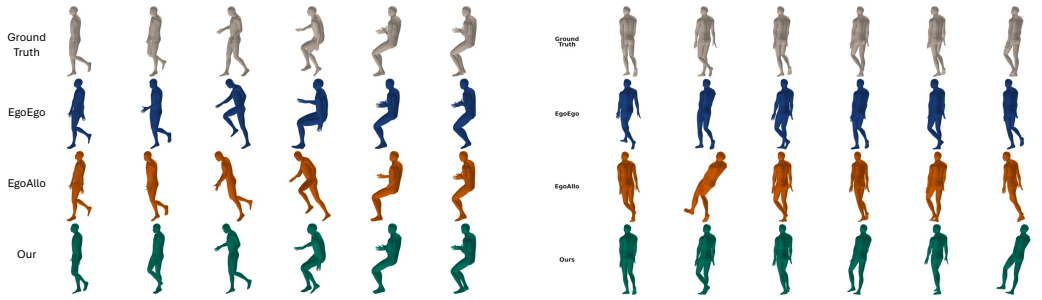


Figure 6: **Additional method comparison on Nymeria.** Two further test sequences extending Fig. 3 of the main paper. **Left:** *sitting down onto a sofa*—a sit-down transition where the head descent is consistent with many distinct lower-body configurations. **Right:** *standing-to-walking transition*—gait initiation with asymmetric weight shift that is nearly invisible to the SLAM signal. Rows show, from top to bottom: ground truth (gray), EgoEgo (Li et al., 2023) (blue), EgoAllo (Yi et al., 2025) (orange), and FlowEgoMotion (Ours, green). EgoEgo exhibits articulation drift in the limbs; EgoAllo recovers a plausible coarse posture but shows foot–floor inconsistencies and arm jitter when the body leaves the field of view; FlowEgoMotion more closely tracks the reference articulation and keeps the body grounded in both scenarios.

Walking through a cluttered room. Free walking through furniture-dense spaces tests whether scene conditioning regularizes articulation without overfitting layout priors. The head trajectory is locally smooth, so most of the articulated motion—arm swing, torso rotation around obstacles—must come from the prior. In Fig. 4, the predicted body navigates between objects with arm and torso configurations that remain plausible across the full sequence, even though the wearer's arms briefly leave the field of view between frames.

Reaching and bending. Bending over to reach for an object on a low surface produces a head trajectory that is locally indistinguishable from a deep squat: the head moves down and slightly forward in both cases, but the limb configurations differ entirely. Disambiguation requires knowing where the relevant surface is. In the corresponding row of Fig. 6, EgoEgo and EgoAllo collapse toward a generic squat-like prior, while FlowEgoMotion produces a forward-bend posture consistent with the table-height surface in the scene tokens.

Standing-to-walking transitions. Gait initiation involves an asymmetric weight shift that is nearly invisible to SLAM because the head barely moves. In the corresponding sequence of Fig. 6, FlowEgoMotion recovers the leading-leg asymmetry of the first step where both baselines produce a near-symmetric posture.

Summary. Across these activity types, the failure modes of the baselines follow the predictions of Sec. 3: SLAM-only conditioning loses the articulated detail that depends on the surrounding environment, and image-conditioned methods recover coarse posture but cannot enforce contact with surfaces that are out of view. FlowEgoMotion's factorized local/global scene pathways target both failure modes. These examples are illustrative; the quantitative evidence is in Sec. 4 and App. A.

C Dataset Details

We use two native real-egocentric benchmarks (Nymeria, ADT) and two synthesized-egocentric benchmarks (AMASS, RICH). The distinction matters for interpreting our results: Nymeria and ADT provide real wearable sensing and directly evaluate the scene-aware setting targeted by FlowEgoMotion, whereas AMASS and RICH are converted into egocentric benchmarks by synthesizing device trajectories and are used only to evaluate the underlying SLAM-conditioned prior under a controlled setting. Across all four datasets we follow SMPL-H conventions (Romero et al., 2022) and the SLAM-conditioning protocol of Yi et al. (2025), so that the four benchmarks are mutually compatible at the metric level.

Nymeria. Nymeria (Ma et al., 2024) is our primary real-egocentric dataset and our only training source. Each recording provides synchronized Project Aria streams (RGB, mono SLAM cameras, IMU), the on-device closed-loop SLAM trajectory, an aggregated semi-dense 3D scene point cloud, and full-body motion ground truth captured by an Xsens inertial suit aligned to the metric world frame, all exposed under SMPL-H joint conventions. We split the data by subject identity into 130 training and 55 test subjects, yielding 595 training sequences and 253 test sequences ($\approx$149 h train / $\approx$63 h test, mean sequence length $\approx$15 min at 30 Hz). Preprocessing keeps the released SMPL-H body parameters and shape $\boldsymbol{\beta}$ unchanged; we resample SLAM poses to the 30 Hz motion clock by linear interpolation in translation and SLERP in rotation, drop the first and last 0.5 s of each sequence to avoid SLAM warm-up artefacts, and read the floor plane directly from the dataset's scene-graph metadata.

Aria Digital Twin (ADT). ADT (Pan et al., 2023) is our second real-egocentric benchmark and is used *evaluation-only*: no FlowEgoMotion weights are trained or fine-tuned on it. ADT provides native Project Aria sequences with aligned RGB, SLAM trajectories, scene meshes, object 6-DoF poses, and full-body motion captured by an OptiTrack system following the Biomech57 joint template (27 body joints + 30 hand joints). Two preprocessing steps are needed for parity with the SMPL-H protocol used elsewhere: (i) joint correspondence — following Yi et al. (2025), we manually map the 27 Biomech57 body joints to their SMPL-H counterparts where the 1:1 correspondence is unambiguous (feet, knees, hips, shoulders, elbows, wrists), and mask out joints whose conventions differ (head and collar joints), so that MPJPE / PA-MPJPE / Wrist-PE are computed only over matched joints; and (ii) floor extraction — since ADT does not annotate a metric floor plane, we run RANSAC on the high-confidence semi-dense SLAM points to fit the dominant horizontal plane $z = z_{\text{floor}}$, again following Yi et al. (2025).

AMASS. AMASS (Mahmood et al., 2019) is a unified collection of motion-capture sequences expressed in SMPL-H parameters, with no native egocentric sensing. Following the synthesis protocol of Yi et al. (2025); Li et al. (2023), for each motion we anchor a central pupil frame (CPF) between the SMPL-H vertices corresponding to the left and right eyes in the blend-skinned mesh, propagate the CPF along the motion to obtain a synthetic $SE(3)$ head trajectory $\mathbf{T}^W_{\text{cpf},t}$, and treat this trajectory as the device input $\mathbf{T}^W_{D,t}$ used by our model. The floor plane is read directly from the AMASS metadata ($z = 0$ in the canonical frame). The available modalities are therefore body motion (SMPL-H), shape $\boldsymbol{\beta}$, and the synthesised CPF trajectory; no images or scene point cloud are used. We follow the public AMASS test split shared by Yi et al. (2025); Li et al. (2023) and use windows of $T = 128$ frames at 30 Hz.

RICH. RICH (Huang et al., 2022) is a human–scene interaction dataset that pairs detailed SMPL-X body motion with calibrated 3D scene scans, again without native egocentric sensing. We convert SMPL-X parameters to SMPL-H by dropping facial-expression and detailed-hand DoFs and copying the shared body and global-orientation parameters, then synthesise CPF trajectories with the same procedure used for AMASS. The floor plane is taken from each sequence's scene scan. Compared with AMASS, RICH has richer body–scene contact (sitting, lying, leaning) and cluttered environments; we follow the test split released with the dataset and otherwise apply identical windowing and preprocessing.

Every dataset exposes the SLAM/CPF trajectory and SMPL body motion used by every model. Only Nymeria and ADT additionally provide a real-world scene point cloud / mesh ($\mathcal{S}$) and synchronized egocentric RGB; AMASS and RICH provide only synthetic trajectories with a flat floor. Accordingly, the full FlowEgoMotion model (with both local and global scene streams) is evaluated on Nymeria and ADT, while the motion-only variant is evaluated on AMASS and RICH (App. B.1).

The main paper focuses on Nymeria and ADT because they directly test our central claim: *explicit scene conditioning improves egocentric motion estimation on real wearable data*. AMASS and RICH appear in App. B as supporting evidence that the proposed rectified-flow prior remains strong even when the problem is reduced to SLAM-only estimation with synthesized device trajectories.

D IMPLEMENTATION DETAILS

This appendix collects the implementation choices used in all experiments. The goal is to make the reported results easy to interpret and reproduce while keeping the main paper focused on the method itself. Notation follows Sec. 3.

D.1 STATE REPRESENTATION

The model predicts a compact body-centric state consisting of sequence-level body shape coefficients together with per-frame local body rotations and contact variables. Concretely, the packed state dimension is

$$d_{\text{state}} = 16 + 21 \times 9 + 21 = 226, \tag{18}$$

corresponding to 16 shape coefficients, 21 body-joint rotation matrices represented in flattened 3×3 form, and 21 contact logits. Wrist-based evaluation metrics (Wrist-PE) are computed from the wrist joint positions obtained by forward kinematics through this body representation; we do not regress finger-level hand articulation.

Note that the state vector packs nine values per joint rotation rather than three (axis-angle) or four (quaternion). This choice avoids the discontinuities of axis-angle and quaternion representations during flow training; we project predicted matrices back to $SO(3)$ via polar decomposition (SVD) at decoding time.

Local scene encoder. The local stream provides time-aligned geometric evidence around the wearer. We use the pretrained LitePT-S backbone (Yue et al., 2025) as a frozen point-cloud encoder; for each device-pose timestamp t it consumes a fixed-size point set extracted from the scene cloud and produces a 504-dimensional descriptor $\mathbf{l}_t$ as in Eq. 4 of the main text. The procedure is fully deterministic given a fixed seed and is executed once, offline; at training time the dataloader reads only the cached descriptors.

(I) SPATIAL REGION OF INTEREST. Let $\mathbf{p}_t^D \in \mathbb{R}^3$ denote the translational component of the device pose $\mathbf{T}_{D,t}^W \in SE(3)$, and let $\mathcal{S}_W = \{\mathbf{x}_i\}_{i=1}^N$ denote the scene point cloud expressed in the world frame. The local region of interest at time t is the open metric ball

$$\mathcal{R}_t := \big\{\mathbf{x} \in \mathcal{S}_W \,:\, \|\mathbf{x} - \mathbf{p}_t^D\|_2 < r\big\}, \qquad r = 1.5 \text{ m}. \tag{19}$$

We do not use a frustum mask, k-nearest-neighbour query, or sensor-cone restriction: the region is purely a Euclidean sphere of fixed radius around the device origin.

(II) COORDINATE FRAME. Points retained by Eq. (19) are expressed in a *head-centric, world-axis-aligned* frame by translating the world origin to the device:

$$\tilde{\mathbf{x}}_{i,t} := \mathbf{x}_i - \mathbf{p}_t^D, \qquad \mathbf{x}_i \in \mathcal{R}_t. \tag{20}$$

We deliberately do *not* apply the rotational component $\mathbf{R}_{D,t}^W$, nor the yaw-only rotation $\mathbf{R}_{\hat{D}_t}^W$ used to canonicalize the conditioning representation in §3.1. Gravity and world yaw are therefore preserved in the input to LitePT, so the encoder sees a translation-only normalisation of the scene. We found this important in preliminary experiments: rotating the local point cloud into the canonical frame $\hat{D}_t$ destroys the cue that "down" is consistently aligned with $-z$, which the encoder otherwise uses to disambiguate floor, table, and ceiling surfaces.

(III) SUB-SAMPLING, VOXELIZATION, AND NORMALIZATION. The cropped cloud is forced to a constant cardinality $K = 256$ by uniform random resampling,

$$\mathcal{C}_t := \begin{cases} \text{sample without replacement of size } K \text{ from } \mathcal{R}_t, & |\mathcal{R}_t| \geq K, \\ \mathcal{R}_t \cup \text{sample with replacement of size } K - |\mathcal{R}_t| \text{ from } \mathcal{R}_t, & 0 < |\mathcal{R}_t| < K, \\ \mathbf{0}_{K\times 3}, & \mathcal{R}_t = \varnothing, \end{cases} \tag{21}$$

yielding $\mathcal{C}_t \in \mathbb{R}^{K\times 3}$. No mean-centring, per-axis scaling, or unit-norm rescaling is applied beyond the translation in Eq. (20). Voxelization is delegated to LitePT-S, which performs a grid-sample with edge length

$$g = 0.02 \text{ m}, \tag{22}$$

retaining one representative point per occupied voxel before the sparse-convolutional stages (Yue et al., 2025). The per-voxel features at the most compressed scale (504-D) are mean-pooled across the voxels of frame t:

$$\mathbf{l}_t = \frac{1}{|\mathcal{V}_t|} \sum_{v \in \mathcal{V}_t} \mathbf{h}_v \ \in \ \mathbb{R}^{504}, \tag{23}$$

where $\mathcal{V}_t$ is the set of occupied voxels at time t and $\mathbf{h}_v$ is the LitePT backbone output at voxel v.

(IV) CACHING STRATEGY. Because $\mathbf{l}_t$ depends only on the device trajectory and the static scene, features are computed offline once per trajectory and never recomputed during training. For every Aria recording the full sequence

$$\mathbf{F}_{\text{seq}} = [\mathbf{l}_1, \ldots, \mathbf{l}_T]^\top \in \mathbb{R}^{T \times 504} \tag{24}$$

is stored in `float16` on disk and aggregated across the corpus into a single HDF5 archive. At training time the dataloader reads subsequences directly from this cache; only the learnable linear projection $\mathbf{P}_{\text{loc}} \in \mathbb{R}^{d \times 504}$ defined in Eq. 5 of the main text is applied on the fly.

The descriptor sequence $\mathbf{l}_{1:T}$ is the input to the gated fusion module of Eqs. (5)–(6). During training the descriptors are cached; at inference they must be computed for each new trajectory ($\approx$0.2 ms per frame, App. A.2).

Global scene encoder. The global stream summarizes room-scale layout and affordance structure. We use SceneScript (Avetisyan et al., 2024) as a frozen, pre-trained scene encoder that produces structured, autoregressively decoded scene tokens describing walls, floors, doors, and large furniture as parametric primitives. Each scene yields a variable-length token sequence $\mathbf{G}^{\text{glob}} = (\mathbf{g}_m^{\text{glob}})_{m=1}^{M_s}$ with $\mathbf{g}_m^{\text{glob}} \in \mathbb{R}^{256}$. We zero-pad this sequence to a fixed maximum length $M = 512$ and supply a per-token validity mask $\boldsymbol{\mu} \in \{0,1\}^M$ used in masked pooling (Eq. 8) and key-padding for cross-attention (Eq. (35)). The pooled summary $\bar{\mathbf{g}}^{\text{glob}} \in \mathbb{R}^{256}$ feeds the AdaLN modulation pathway.

D.2 VECTOR-FIELD NETWORK

The conditional vector field $\mathbf{v}_\theta$ is a continuous-time Transformer operating on noisy motion tokens. We describe each component below. All learned linear layers are initialised with the default PyTorch scheme except where stated otherwise.

Conditioning lifting Φ. The per-timestep conditioning vector $\mathbf{u}_t \in \mathbb{R}^{12}$ from Eq. (2) is lifted to dimension $d = 768$ by $K = 6$ sinusoidal Fourier bands followed by a two-layer MLP with GELU activation:

$$\Phi(\mathbf{u}) = \text{MLP}\big([\sin(2^k \pi \mathbf{u}), \cos(2^k \pi \mathbf{u})]_{k=0}^{K-1}\big). \tag{25}$$

Egomotion encoder E_{ego}. E_{ego} in Eq. 3 is a single token-wise affine projection. Given the per-frame conditioning vector $\mathbf{c}_t \in \mathbb{R}^{d_{\text{cond}}}$ assembled from the device pose stream, it produces

$$\mathbf{h}_t^{\text{ego}} \ = \ W_{\text{ego}}\, \mathbf{c}_t \ + \ \mathbf{b}_{\text{ego}}, \qquad W_{\text{ego}} \in \mathbb{R}^{d \times d_{\text{cond}}}, \tag{26}$$

with token width $d = 768$. Stacking over time yields $\mathbf{H}^{\text{ego}} = [\mathbf{h}_1^{\text{ego}}; \ldots; \mathbf{h}_T^{\text{ego}}] \in \mathbb{R}^{T \times d}$. Temporal mixing of these tokens is performed inside the denoiser via cross-attention rather than by a dedicated egomotion Transformer; we therefore do not pre-contextualise $\mathbf{c}_{1:T}$ before injection.

Trajectory encoder E_{traj}. E_{traj} in Eq. (14) projects the noisy state $\mathbf{x}_\tau \in \mathbb{R}^{T \times d_{\text{state}}}$ to the latent token width via a linear layer followed by a learned positional encoding:

$$\mathbf{Z}_t^{(0)} = \text{Linear}_{d_{\text{state}} \to d}(\mathbf{x}_{\tau,t}) + \mathbf{p}_t, \tag{27}$$

where $\mathbf{p}_t \in \mathbb{R}^d$ are learned position embeddings.

Time embedding $\phi(\tau)$. Transport time $\tau \in [0,1]$ is embedded with sinusoidal Fourier features followed by a two-layer MLP, identical in form to Φ but with input dimension 1:

$$\phi(\tau) = \text{MLP}\big([\sin(2^k \pi \tau), \cos(2^k \pi \tau)]_{k=0}^{5}\big) \in \mathbb{R}^d. \tag{28}$$

Pooling operators. Pool_T in Eqs. (3) and (15) is mean pooling over the temporal axis. $\mathrm{MaskPool}$ in Eq. (8) is mean pooling over valid tokens only:

$$\mathrm{MaskPool}(\mathbf{G}^{\mathrm{glob}}, \boldsymbol{\mu}) = \frac{\sum_{m=1}^{M} \mu_m \, \mathbf{g}_m^{\mathrm{glob}}}{\sum_{m=1}^{M} \mu_m}. \tag{29}$$

Modulation function f. f in Eq. (15) concatenates its three inputs and applies a two-layer MLP with output dimension d:

$$\boldsymbol{\Upsilon} = \mathrm{MLP}\big([\mathrm{Pool}_T(\mathbf{Q}_{1:T}); \, \bar{\mathbf{g}}^{\mathrm{glob}}; \, \mathbf{e}_\tau]\big). \tag{30}$$

Decoder D. D in Eq. (16) is a single linear projection from token width d to the state dimension d_{state}, applied independently per timestep.

Backbone configuration. The full vector-field Transformer uses

$$L = 8, \qquad d_{\mathrm{latent}} = 768, \qquad d_{\mathrm{ff}} = 3072, \qquad H = 8, \tag{31}$$

where L is the number of layers, d_{latent} the token width, d_{ff} the feed-forward width, and H the number of attention heads. Including all scene projections, lifting maps, and the modulation MLPs, the full model has approximately 145M parameters.

Initialization of conditioning pathways. Scene-specific pathways are initialized conservatively so that optimization starts from a stable motion-conditioned model. Specifically, the local residual projection $\mathbf{W}_\delta$ is zero-initialized, the local-fusion gate bias is set to $b_\alpha = -4$ (giving an initial gating value $\sigma(-4) \approx 0.018$ that effectively disables the local stream at step zero), and the projections producing the residual gates $\gamma_{\mathrm{sa}}, \gamma_{\mathrm{ego}}, \gamma_{\mathrm{scene}}, \gamma_{\mathrm{ff}}$ are all zero-initialized. As a result, the model starts behaving as a pure motion-only flow and gradually integrates scene information as training progresses.

D.3 Detailed Transformer block updates

We adopt the DiT parameterization (Peebles & Xie, 2023): for a token $\mathbf{z}$, AdaLN is defined as

$$\mathrm{AdaLN}(\mathbf{z}; \boldsymbol{\Upsilon}) = \big(1 + s(\boldsymbol{\Upsilon})\big) \odot \mathrm{LN}(\mathbf{z}) + b(\boldsymbol{\Upsilon}), \tag{32}$$

where $s(\cdot)$ and $b(\cdot)$ are learned linear projections from $\boldsymbol{\Upsilon}$ to dimension d. Each Transformer block contains four sublayers (self-attention; cross-attention to $\mathbf{Q}_{1:T}$; cross-attention to $\mathbf{G}^{\mathrm{glob}}$; feed-forward). Every sublayer is wrapped in AdaLN and augmented with a residual gate predicted from $\boldsymbol{\Upsilon}$:

$$\tilde{\mathbf{Z}}^\ell = \mathbf{Z}^\ell + \gamma_{\mathrm{sa}}(\boldsymbol{\Upsilon}) \odot \mathrm{SA}\big(\mathrm{AdaLN}(\mathbf{Z}^\ell; \boldsymbol{\Upsilon})\big), \tag{33}$$

$$\hat{\mathbf{Z}}^\ell = \tilde{\mathbf{Z}}^\ell + \gamma_{\mathrm{ego}}(\boldsymbol{\Upsilon}) \odot \mathrm{CA}_{\mathrm{ego}}\big(\mathrm{AdaLN}(\tilde{\mathbf{Z}}^\ell; \boldsymbol{\Upsilon}), \, \mathbf{Q}_{1:T}\big), \tag{34}$$

$$\bar{\mathbf{Z}}^\ell = \hat{\mathbf{Z}}^\ell + \gamma_{\mathrm{scene}}(\boldsymbol{\Upsilon}) \odot \mathrm{CA}_{\mathrm{scene}}\big(\mathrm{AdaLN}(\hat{\mathbf{Z}}^\ell; \boldsymbol{\Upsilon}), \, \mathbf{G}^{\mathrm{glob}}; \, \boldsymbol{\mu}\big), \tag{35}$$

$$\mathbf{Z}^{\ell+1} = \bar{\mathbf{Z}}^\ell + \gamma_{\mathrm{ff}}(\boldsymbol{\Upsilon}) \odot \mathrm{MLP}\big(\mathrm{AdaLN}(\bar{\mathbf{Z}}^\ell; \boldsymbol{\Upsilon})\big). \tag{36}$$

Residual gating is especially important in the few-step flow regime because it lets the network regulate update magnitude as a function of transport time and conditioning strength. After L blocks, the decoder produces the vector-field prediction $\mathbf{v}_\theta = D(\mathbf{Z}^L)$ as in Eq. (16).

D.4 Training objective

The main text defines the training loss as $\mathcal{L} = \mathcal{L}_{\mathrm{RF}} + \lambda_{\mathrm{FK}} \mathcal{L}_{\mathrm{FK}}$ (Eq. (13)). The rectified-flow term decomposes as a modality-weighted sum over the body-shape, body-rotation, and contact components,

$$\mathcal{L}_{\mathrm{RF}} = \lambda_\beta \mathcal{L}_\beta + \lambda_R \mathcal{L}_R + \lambda_c \mathcal{L}_c, \tag{37}$$

which is the explicit form of the block-diagonal $\boldsymbol{\Lambda}$ weighting in Eq. (10) when $\boldsymbol{\Lambda}$ has constant blocks $\lambda_\beta \mathbf{I}, \lambda_R \mathbf{I}, \lambda_c \mathbf{I}$. The total objective is therefore

$$\mathcal{L} = \lambda_\beta \mathcal{L}_\beta + \lambda_R \mathcal{L}_R + \lambda_c \mathcal{L}_c + \lambda_{\mathrm{FK}} \mathcal{L}_{\mathrm{FK}}. \tag{38}$$

We use weights

$$(\lambda_\beta,\, \lambda_R,\, \lambda_c,\, \lambda_{\mathrm{FK}}) = (0.1,\, 1.0,\, 0.1,\, 5.0). \tag{39}$$

The relatively large $\lambda_{\mathrm{FK}} = 5.0$ reflects the fact that joint-position errors in metric world coordinates are the ultimate evaluation criterion; the lower weights on shape and contact prevent these auxiliary outputs from dominating the rotation signal.

Transport time is sampled using the logit-normal parameterization (Esser et al., 2024)

$$z \sim \mathcal{N}(0, 1), \qquad \tau = \sigma(z), \tag{40}$$

which concentrates training signal near the middle of the transport path. The clean-sample estimate $\hat{\mathbf{y}}$ used in $\mathcal{L}_{\mathrm{FK}}$ is defined in Eq. (11) of the main text; it is not detached, so $\mathcal{L}_{\mathrm{FK}}$ back-propagates through $\mathbf{v}_\theta$, the $SO(3)$ projection, and the SMPL-H forward-kinematics layer.

D.5 Training procedure

Optimization. We train with AdamW using

$$\text{learning rate} = 10^{-4}, \qquad \text{weight decay} = 10^{-4}, \qquad \text{batch size} = 128. \tag{41}$$

We use $1{,}000$ warmup steps followed by cosine decay to 10^{-5}, clip gradients to $\|\nabla\|_2 \leq 1.0$, and maintain an exponential moving average (EMA) of model parameters with decay 0.9999. All reported evaluations use the EMA weights.

Iterations and compute. Training runs for ≈ 400k steps on a single 40 GB NVIDIA A100 GPU, taking approximately four days of wall-clock time for each run.

D.6 Data preparation and augmentation

Splits. We use Nymeria as the primary real-world training source. We split the data by subject identity into 130 training and 55 test subjects, yielding 595 training sequences and 253 test sequences of synchronized SLAM, scene, and motion data, with each sequence lasting approximately 15 minutes (≈ 149 hours of training data and ≈ 63 hours of test data).

Subsequence sampling. Training uses subsequences of length $T = 128$ frames ($\approx$4.3 s of motion at 30 Hz). At each training step we sample subsequences with random-uniform slicing within each sequence. For deterministic evaluation on the test split, we slice with a fixed stride.

Scene dropout. During training, we apply stochastic scene dropout with probability

$$p_{\mathrm{drop}} = 0.2. \tag{42}$$

When dropout is triggered for a training example, both local and global scene streams are zeroed out simultaneously for the entire sample. This regularizes the model against over-reliance on scene cues and improves robustness when scene features are noisy, incomplete, or unavailable. Section 4.2 ablates $p_{\mathrm{drop}} \in \{0.0, 0.2, 0.3\}$ and finds $p_{\mathrm{drop}} = 0.2$ to be the best operating point. We do not use classifier-free guidance or any conditional/unconditional extrapolation at test time.

D.7 Inference

Solver and time schedule. Unless otherwise stated, inference uses a first-order Euler solver with

$$S = 2 \tag{43}$$

transport steps. We evaluate both uniform and cosine time schedules, with the cosine schedule serving as a drop-in alternative that places more samples near the middle of the transport interval where the velocity field carries the most information.

Overlapping-window blending. The model is trained on fixed windows of length $T = 128$. For longer sequences we run inference on overlapping windows with stride 96 (overlap 32) and blend predictions using linear ramp weights: each frame in the overlap region is a convex combination of predictions from the two windows, weighted by distance to the window boundaries. This produces a coherent full-length trajectory without autoregressive error accumulation.

Post-processing. The only deterministic post-processing steps are: (i) projection of predicted rotation matrices back to $SO(3)$ via SVD polar decomposition; and (ii) analytic recovery of the world-space root transform from the observed device trajectory through Eq. (1) of the main text. No test-time optimization, classifier-free guidance, or learned refinement is used.

Latency. The 40 ms figure is the cost of the generative stage for one window: two Euler steps of the vector-field network over $T = 128$ frames, measured on a single NVIDIA A100 (40 GB) and averaged over 10000 runs, with scene features precomputed as in HMD^2. With stride-96 windows at 30 Hz this amounts to ≈ 12.5 ms of network compute per second of motion, or ≈ 18.5 ms including local LitePT descriptors (App. A.2).