# UniEgoMotion: A Unified Model for Egocentric Motion Reconstruction, Forecasting, and Generation

Supplementary Material

## A. Qualitative Comparison

See Fig. 1 for a qualitative visualization of egocentric motion reconstruction, with vertex errors color-coded. Please refer to the supplementary video to view UniEgoMotion's results on egocentric motion reconstruction, forecasting, and generation, as well as comparisons with baselines.

## B. Baselines

### Egocentric Motion Reconstruction

We compare the egocentric motion reconstruction capabilities of UniEgoMotion with task-specific prior works: EgoEgo [13], EgoAllo [22], and AvatarPoser [12]. To ensure a fair evaluation, we retrain each method on the EE4D-Motion dataset using their publicly available code. Following EgoEgo's experimental setup, we exclude hand tracking from AvatarPoser and instead provide a constant input for hand trajectories. Both EgoEgo and AvatarPoser use head trajectories derived from motion annotations rather than from the Aria device's SLAM system, resulting in perfect head tracking by design. Therefore, we omit their head tracking metrics from the evaluation. For EgoAllo, we evaluate the output of the motion diffusion model directly, without applying the post-processing optimization step.

Although EgoEgo and EgoAllo also adopt diffusion-based formulation for motion reconstruction, their approach differ from ours in their choice of motion representation and model architecture. For instance, EgoEgo assumes a constant body shape and uses a global motion representation, whereas EgoAllo uses a head-centric representation that explicitly includes the head-to-pelvis transformation and preserves the kinematic chain. More importantly, none of these baselines utilize semantic information from egocentric video for motion prediction. We compare these methods on the reconstruction task and also ablate their design choices separately within our UniEgoMotion framework in a consistent manner.

### Egocentric Motion Forecasting & Generation

For egocentric motion forecasting and generation, the most relevant baselines [3, 19] are two-stage models that generate or forecast human motion from third-person RGB images. They first predict the root trajectory (typically pelvis) and then generate the full-body human motion using a global motion representation. To replicate these baselines faithfully, we train a separate UniEgoMotion variant that uses global motion representation and predicts only the root trajectory. This output is then provided as an additional conditioning input to the standard UniEgoMotion model (also with global motion representation) for full-body motion prediction. We also train separate autoregressive LSTM-based baselines with a comparable model capacity for both forecasting and generation tasks. Since these models lack a generative component, their outputs tend to regress toward the mean of all plausible futures. As a result, they show lower error in direct comparison metrics such as MPJPE. However, their 'averaged' prediction suffer from reduced motion diversity and realism, as shown in semantic metrics and qualitative visualization (see supplementary video).

## C. Ablation on Conditioning Inputs

We evaluate UniEgoMotion under two ablation settings: without trajectory input and without video input. Additionally, we train two single-modality variants of UniEgoMotion. Egocentric reconstruction results in Tab. 1 shows that both signals are useful for optimal reconstruction performance, thereby validating our use of video input, unlike prior baselines. Interestingly, the separately trained single-modality variants offer no significant advantage over the original UniEgoMotion model when evaluated under the same conditions. Without video input, UniEgoMotion still outperforms baselines on most metrics. However, when the trajectory input is removed, the model is forced to implicitly solve visual odometry problem (a significantly harder task), leading to large errors on absolute metrics (head tracking, MPJPE, MPJPE-H). Despite this, it maintains accuracy in local pose metrics (MPJPE-PA, semantic similarity) and realism (FID), showing its ability to infer plausible motion from video alone.

EgoEgo [13] employed an off-the-shelf monocular visual SLAM on egocentric video and trained an additional module to predict scale and the gravity vector to derive gravity-aligned metric SLAM trajectory. Their results showed that using predicted metric SLAM trajectory leads to only a minor degradation in pose metrics compared to using ground-truth trajectories. In our work, we assume access to inertial SLAM trajectories for both our method and the baselines to decouple motion analysis from trajectory estimation and focus our evaluation on motion tasks.

Forecasting and generation results follow similar trends, with both input modalities contributing to optimal performance. Notably, the model without video input performs worse, as it lacks scene context necessary for generating or forecasting relevant motion.

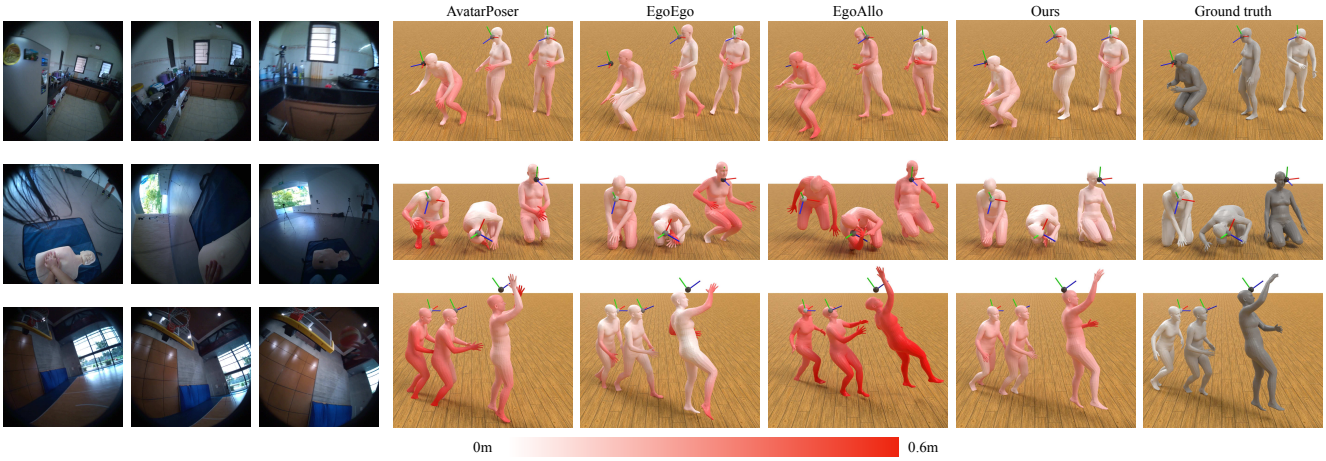


Figure 1. Qualitative comparison of Egocentric Reconstruction, with absolute vertex errors color-coded. The input egocentric images are shown on the left, with the corresponding ego-device trajectory visualized alongside the predictions.

Table 1. **Ablation on Conditioning Inputs**: We evaluate UniEgoMotion in two ablation settings–without video and without trajectory input. Additionally, we train two single-modality variants of UniEgoMotion by conditioning only on trajectory or only on video.

| | Egocentric Motion Reconstruction | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | Head Rot. Err. | Head Trans. Err. | MPJPE | MPJPE-PA | MPJPE-H | Foot Slide | Foot Contact | Semantic Sim.(↑) | FID |
| UniEgoMotion | **0.260** | 0.058 | **0.100** | **0.053** | **0.180** | 3.62 | 0.027 | **0.918** | 0.027 |
| w/o video | 0.278 | **0.057** | 0.115 | 0.066 | 0.234 | 3.64 | 0.026 | 0.878 | 0.030 |
| w/o trajectory | 0.539 | 0.280 | 0.290 | 0.059 | 0.352 | 2.95 | 0.024 | 0.885 | 0.033 |
| UniEgoMotion (w/o video) | 0.293 | 0.063 | 0.119 | 0.067 | 0.239 | 3.49 | 0.025 | 0.877 | **0.026** |
| UniEgoMotion (w/o trajectory) | 0.535 | 0.292 | 0.299 | 0.060 | 0.362 | **2.70** | **0.023** | 0.886 | 0.035 |

| | Egocentric Motion Forecasting | | | | | | | Egocentric Motion Generation | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | J (2-4s) | J-PA (2-4s) | J-H (2-4s) | FS | FC | SS (↑) | FID | J (0-2s) | J-PA (0-2s) | J-H (0-2s) | FS | FC | SS (↑) | FID |
| UniEgoMotion | **0.206** | 0.071 | **0.308** | 2.60 | 0.026 | **0.849** | **0.047** | **0.226** | **0.070** | **0.321** | 2.89 | 0.025 | 0.817 | **0.043** |
| w/o video | 0.255 | 0.090 | 0.378 | **2.43** | 0.028 | 0.782 | 0.058 | 0.356 | 0.100 | 0.449 | **2.36** | 0.027 | 0.696 | 0.065 |
| w/o trajectory | 0.322 | **0.070** | 0.414 | 2.66 | 0.025 | 0.838 | **0.047** | **0.226** | **0.070** | **0.321** | 2.89 | 0.025 | 0.816 | **0.043** |
| UniEgoMotion (w/o video) | 0.276 | 0.095 | 0.400 | 2.69 | 0.028 | 0.767 | 0.067 | 0.379 | 0.108 | 0.483 | 3.03 | 0.027 | 0.684 | 0.044 |
| UniEgoMotion (w/o trajectory) | 0.318 | **0.070** | 0.404 | 2.50 | **0.024** | 0.842 | 0.050 | 0.228 | **0.070** | **0.321** | 2.71 | **0.024** | **0.820** | 0.044 |

## D. Why Not Text Conditioning

Many motion generation approaches [5, 17, 20] rely on text-based conditioning, where a clear textual prompt defines the intended motion or action. This explicit guidance simplifies the generation process. In contrast, our work focuses on passive conditioning using sensor data (e.g., video and device trajectory), where motion must be inferred without direct user input. While this introduces greater ambiguity, it also enables broader applicability in real-world scenarios such as continuous gait monitoring or fall prediction, where explicit user inputs are typically unavailable. Nonetheless, we believe that egocentric motion generation and forecasting from text prompts are promising future directions for many assistive applications. Our work, along with datasets like EE4D-Motion (with action narrations from EgoExo4D) and Nymeria [6], offers a promising starting point for such research.

## E. Training Details

We train UniEgoMotion for 350 epochs using a batch size of 64 and the AdamW optimizer with a weight decay of 0.01. The learning rate is initialized at 3e-5 and decayed to 3e-6 after epoch 300. The model follows a standard transformer architecture [18], comprising 12 decoder layers with a latent

dimension of 768. Training is conducted on 8-second motion sequences (80 steps at 10 fps), enabling long-horizon motion prediction. To improve training efficiency, DINOv2 features are precomputed and cached. End-to-end training takes approximately 2 days on a single NVIDIA L40S GPU. For diffusion, we use cosine noise scheduling with 1000 steps, consistent with prior works [13, 17], though effective motion synthesis has been demonstrated with very few diffusion steps [11, 17]. During training, we alternate between reconstruction and generation tasks with equal 0.5 probability by randomly masking the input sequence.

## F. Motion Representation

Although SMPL-X parameters $X_i = (R_i^r, t_i^r, \theta_i, \beta_i)$ are sufficient to represent 3D body motion, they are not always ideal for learning [10, 22]. The global parameterization of the root trajectory $(R_i^r, t_i^r)$, defined at the pelvis, does not exploit motion invariances, forcing the model to learn all movements in every direction separately. Moreover, a mismatch exists between the conditioning information $(T_i, I_i)$, defined in the egocentric frame, and the SMPL-X parameters $X_i$, defined in the pelvis-centric frame. This misalignment complicates the reasoning between pelvis-centric motion and egocentric conditioning inputs. Additionally, using local joint angles forces the model to reason complex forward kinematics of the SMPL-X skeleton, often resulting in suboptimal motion with noticeable artifacts such as foot-floor penetration and foot sliding.

To address these issues, we adopt a head-centric motion representation instead of a pelvis-centric one. We transform the SMPL-X parameters $X_i = (R_i^r, t_i^r, \theta_i, \beta_i)$ into $(M_i^h, \boldsymbol{M}_i^j)$ using forward kinematics where $M_i^h \in \mathbb{R}^{4\times4}$ is the global SE(3) transform of the head joint, and $\boldsymbol{M}_i^j \in \mathbb{R}^{21\times4\times4}$ are the global SE(3) transforms of other joints. This eliminates the dependency of each joint on its parent in the kinematic chain. Next, we derive a canonical reference frame ${}_cM_i$ for each frame by projecting the head transform $M_i^h$ onto the floor. In particular, ${}_cM_i$ represents the *global* 3D transform of the head joint after removing the pitch and roll angle (keeping only yaw) and removing its height $t_z$ relative to the floor (+Z direction). We then express the motion $(M_i^h, \boldsymbol{M}_i^j)$ as $({}_cM_i, {}_cM_i \odot M_i^h, {}_cM_i \odot \boldsymbol{M}_i^j)$, where ${}_cM_i$ captures the head's global trajectory projected onto the floor, and $({}_cM_i \odot M_i^h, {}_cM_i \odot \boldsymbol{M}_i^j)$ encode local canonicalized pose information. To achieve trajectory invariance, we represent ${}_cM_i$ as its residual relative to the previous frame ${}_cM_{i-1}^{-1} \odot {}_cM_i$. Following standard practice, we incorporate additional redundant information, such as joint locations and foot contact labels, into our motion representation.

While our motion representation is similar to the canonicalization in [22], it differs in that [22] retains the kinematic chain and defines local joint rotations relative to parent joints. Since all body joint information in our approach is defined relative to the floor, it naturally facilitates better reasoning about foot-floor contact. We validate the effectiveness of our motion representation through ablation studies and demonstrate that while [22] exhibits significant foot-floor penetration or floating artifacts, UniEgoMotion produces high-quality motion.

## G. EE4D-Motion Dataset

Training UniEgoMotion requires paired egocentric videos and 3D human motion data within real-world environments. However, capturing 3D human motion in everyday activity settings—such as kitchens, offices, and sports fields—is challenging due to the cumbersome setup of motion capture systems. Existing large-scale 3D motion datasets [15, 16] lack paired egocentric videos, while most egocentric datasets either lack 3D motion annotations [8, 9], are small-scale [23], or have limited scene-motion correlation and diversity [13, 24]. The Nymeria dataset [6] stands out with 200+ hours of daily activity egocentric videos paired with motion capture of simple skeleton sequences, but it does not provide the standard SMPL motion representation.

To bridge this gap, we process the large-scale EgoExo4D dataset [9] to generate pseudo-ground-truth 3D motion data. We refer to this processed dataset as EE4D-Motion, which consists of 110+ hours of time-synchronized 3D motion data and egocentric videos, alongside other EgoExo4D annotations. This dataset serves as an extensive benchmark for multimodal motion research.

### EgoExo4D Source Data

EgoExo4D provides synchronized egocentric and exocentric video recordings of diverse activities, including cooking, dance, sports, music, healthcare, and bike repair. Egocentric videos were captured using Project Aria glasses [4] along with the 3D trajectory of the ego camera. While EgoExo4D includes 3D body joint annotations for a subset of the dataset, these annotations are sparse, noisy, discontinuous, and lack joint angle information, making them unsuitable for motion tasks. Thus, we develop a processing pipeline to fit the SMPL-X body model to the continuous frames of EgoExo4D captures.

### Fitting Pipeline

Our pipeline leverages off-the-shelf models for pose estimation and follows a two-stage fitting approach [1, 15] to obtain 3D-accurate motion groundtruth. We exclude rock climbing sequences to focus on motions occurring on a flat surface. Our pipeline consists of the following steps.

**Detection & Tracking**: We detect [14] and track the egocentric camera wearer in each exo view. When multiple people are present, we use the Aria 3D trajectory to identify the person of interest.

**Pose Estimation**: For each bounding box, we estimate 2D keypoints [21] and obtain an initial SMPL-X parameter estimate using an off-the-shelf HMR model [2]. However, single-view HMR estimates suffer from depth ambiguity and jitter in 3D translation.

**Per-Frame Fitting**: We initialize SMPL-X fitting by averaging HMR estimates across exo views. The fitting optimizes SMPL-X parameters $(R^r, t^r, \theta, \beta)$ using the following energy term [1]:

$$\mathcal{L}_{\text{fitting}} = \lambda_\theta E_\theta(\theta) + \lambda_\beta E_\beta(\beta) + \lambda_{2d} \sum_v \sigma\left(\pi_v(J(R^r, t^r, \theta, \beta)) - K_v^{2d}\right)$$

where $E_\theta$ and $E_\beta$ are priors for pose and shape, respectively, $J$ is the SMPL-X 3D joint regressor, $\pi_v$ is the 2D projection operator using known camera intrinsics and extrinsics of view $v$, $K_v^{2d}$ represents detected 2D joints, $\sigma$ is the robust Geman-McClure function [1, 7], and $\lambda_*$ are energy weights.

**Sequence-Level Optimization**: After per-frame fitting, we refine results at the sequence level by fixing the body shape $\beta$ as the average across the sequence, incorporating egocentric view detections, and adding a temporal jitter penalty to enforce smooth motion.

**Filtering & Quality Control**: We filter out segments with excessive jitter caused by erroneous device trajectories, suboptimal off-the-shelf model predictions, or severe occlusions across all exo views.

Through this pipeline, EE4D-Motion provides 3D-accurate motion annotations aligned with egocentric video, enabling us to train and evaluate UniEgoMotion model.

## Motion Annotations Quality

EE4D-Motion annotations can be noisy in scenes with poor exocentric visibility (e.g., kitchen, COVID testing) or large camera distances (e.g., basketball). EgoExo4D's own pose annotations are sparse and jittery, resulting in high pose error of $\sim$0.24m for EgoEgo, as reported by the authors of EgoExo4D [9], compared to $\sim$0.16m on our smoother and denser annotations. Unlike EgoEgo's synthetic dataset, where motions are scene-agnostic, EE4D-Motion provides contextually grounded motion aligned with real-world environments, which is essential for both generation and forecasting tasks.

## References


[1] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 561–578. Springer, 2016. 3, 4

[2] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36:11454–11468, 2023. 4

[3] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 387–404. Springer, 2020. 1

[4] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. 3

[5] Hong et al. Egolm: Multi-modal language model of egocentric motions. *CVPR*, 2025. 2

[6] Ma et al. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *ECCV*, 2024. 2, 3

[7] Stuart Geman and Donald E. McClure. Statistical methods for tomographic image reconstruction. 1987. 4

[8] 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*, pages 18995–19012, 2022. 3

[9] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 3, 4

[10] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5152–5161, 2022. 3

[11] Vladimir Guzov, Yifeng Jiang, Fangzhou Hong, Gerard Pons-Moll, Richard Newcombe, C Karen Liu, Yuting Ye, and Lingni Ma. Hmd2: Environment-aware motion generation from single egocentric head-mounted device. *arXiv preprint arXiv:2409.13426*, 2024. 3

[12] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *European conference on computer vision*, pages 443–460. Springer, 2022. 1

[13] Jiaman Li, 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*, pages 17142–17151, 2023. 1, 3

[14] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European conference on computer vision*, pages 280–296. Springer, 2022. 3

[15] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36: 25268–25280, 2023. 3

[16] 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*, pages 5442–5451, 2019. 3

[17] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 2, 3

[18] 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. 2

[19] Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. Scene-aware generative network for human motion synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12206–12215, 2021. 1

[20] 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 Computer Vision and Pattern Recognition Conference*, pages 22668–22679, 2025. 2

[21] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems*, 2022. 4

[22] Brent Yi, Vickie Ye, Maya Zheng, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. Estimating body and hand motion in an ego-sensed world. *arXiv preprint arXiv:2410.03665*, 2024. 1, 3

[23] Ye Yuan and Kris Kitani. Ego-pose estimation and forecasting as real-time pd control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10082–10092, 2019. 3

[24] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *European conference on computer vision*, pages 180–200. Springer, 2022. 3