MUTUALEGO: ESTIMATING 3D MOTION OF INTERACTING HUMANS FROM DUAL EGOCENTRIC VIEWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recovering the 3D motion of two interacting camera wearers is challenging because each egocentric view only partially observes the partner while moving with its own wearer. We introduce **MutualEgo**, a pair-centric diffusion framework that jointly estimates both participants' motion in a shared reference frame. By treating the interacting pair as a single prediction unit, MutualEgo integrates complementary evidence from both moving views while explicitly modeling their evolving spatial relationship. Its Reference-Aware Motion Tokenizer (RAMT) grounds dense DINOv3 features in calibrated camera geometry and uses anchor-guided attention to aggregate structured global and local conditioning tokens. To support training and evaluation, we introduce **MutualEgo-10K**, comprising 10,105 interaction sequences with rendered dual-egocentric RGB, calibrated camera geometry, and paired 3D motion. On MutualEgo-10K, MutualEgo achieves reconstruction and forecasting MPJPEs of 0.081 m and 0.287 m, compared with 0.261 m and 0.445 m for UniEgoMotion retrained on the same dataset. Ablations evaluate dual-view pair modeling and geometry-aware tokenization, while further experiments study training-data scale and adaptation to Harmony4D.

1 INTRODUCTION

Egocentric smart glasses offer a practical route to capturing human–human interactions beyond instrumented studios. Recovering full 3D motion for both participants from wearable video could support immersive replay, telepresence, and interaction analysis. In this work, we study *egocentric dual-human motion estimation*, which aims to jointly estimate the 3D motion of two interacting camera wearers in a shared reference frame from synchronized dual-egocentric videos and calibrated camera geometry. Each participant-mounted camera primarily observes the partner and surrounding context while moving with its own wearer, whose body lies largely outside its view. Mutual occlusion, limited fields of view, and rapid head motion make each view incomplete, motivating their joint use. Physical contact and action–reaction dynamics also couple the participants over time. We consider two tasks: reconstruction recovers paired motion from complete observations, whereas forecasting predicts future motion from an observed prefix while also reconstructing the observed history. Figure 1 previews the inputs and outputs of both settings.

This task lies at the intersection of egocentric human motion estimation and human–human interaction modeling. Existing egocentric motion estimation primarily focuses on a single person, recovering the wearer from body-facing imagery (Rhodin et al., 2016; Akada et al., 2022; Wang et al., 2022), from forward-facing observations or sparse head/hand tracking (Jiang et al., 2022; Li et al., 2023; Yi et al., 2025; Patel et al., 2025), or reconstructing a visible partner (Zhang et al., 2022; 2023b). Social cues can improve wearer estimation while retaining this single-person prediction target (Ng et al., 2020; Scofano et al., 2025), while multi-wearer and multi-camera systems broaden the available observations for 3D pose estimation and tracking (Khirodkar et al., 2023; Yang et al., 2026) or reconstruction of a designated wearer (Jiang et al., 2026). In contrast, human–human interaction research treats the pair as the modeling unit, capturing relative geometry and contact (Fieraru et al., 2020; Yin et al., 2023; Müller et al., 2024) as well as action–reaction dependencies in motion generation and forecasting (Liang et al., 2024; Xu et al., 2024b; Guo et al., 2022; Fang et al., 2025). These formulations, however, typically condition on external images, motion histories, audio, or text, rather than on complementary moving first-person views carried by the two participants whose motion is being estimated. Egocentric dual-human motion estimation brings these two lines together: both

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.

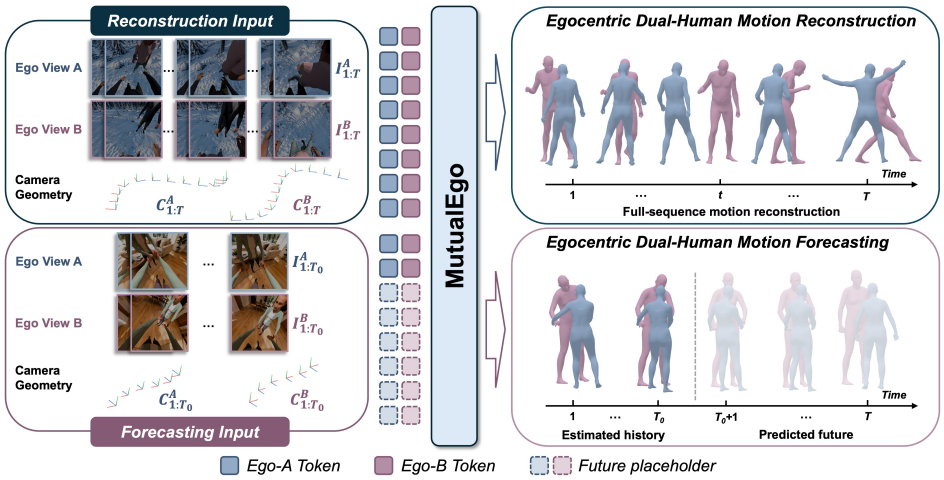


Figure 1: Overview of egocentric dual-human motion estimation. Reconstruction uses synchronized dual-egocentric videos and calibrated camera trajectories over the full observation window to recover both participants' 3D motion in a shared frame. Forecasting uses only the observed prefix of these inputs to reconstruct historical motion and predict future paired motion.

participants are simultaneously observers and prediction targets, and their wearable cameras provide complementary but incomplete views of the interaction. This setting requires jointly integrating visual evidence from two moving cameras in a shared reference frame while maintaining a consistent and evolving spatial relationship between the estimated participants.

To address this task, we develop **MutualEgo**, a pair-centric motion estimation framework that treats the interacting pair as a single prediction unit. Both participants are represented in a shared pair frame, with one temporal motion token jointly encoding their states at each frame. A key challenge is to connect visual evidence from the moving participant-mounted cameras to motion expressed in this shared frame. Dense image features capture appearance, but do not explicitly encode the calibrated camera poses and viewing rays that situate each observation in the common 3D reference frame. To bridge this gap, our Reference-Aware Motion Tokenizer (RAMT) conditions dense DINOv3 features (Siméoni et al., 2025) on pair-frame camera geometry and organizes them into structured global and local tokens. Conditioned on these tokens and the camera trajectories, a temporal diffusion Transformer integrates complementary evidence from both views and jointly denoises the paired motion. Reconstruction and forecasting share the same architecture and are trained independently, with observation-specific masking of visual and camera conditions.

To support the training and evaluation of motion estimators for this task, we construct **MutualEgo-10K**, a dual-egocentric dataset for human interaction and motion. We jointly standardize paired interaction motions from Dual-Human (Fang et al., 2024), Inter-X (Xu et al., 2024a), and additional two-person motion-capture recordings, preserving the relative configuration of each pair. From these motions, we render synchronized dual-egocentric RGB and calibrated camera geometry aligned with paired 3D motion in the shared frame. MutualEgo-10K contains 10,105 interaction sequences, together with synchronized exocentric RGB, depth, instance masks, and auxiliary annotations. Its shared-frame annotations support evaluating both individual motion accuracy and pair-relative geometry.

We evaluate egocentric dual-human motion estimation on MutualEgo-10K through reconstruction and forecasting, measuring both shared-frame body accuracy and pair-relative geometry. MutualEgo improves these measures over released pretrained egocentric estimators and UniEgoMotion trained on MutualEgo-10K. Ablations show that dual-view joint modeling reduces reconstruction relative MPJPE from 0.164 to 0.117 m over independent prediction, with RAMT further reducing it to 0.088 m. Training-data scaling shows continued gains with more paired supervision, while fine-tuning on Harmony4D evaluates adaptation to real capture.

Our contributions are threefold:

- **Egocentric dual-human motion estimation.** We formulate the task of jointly estimating the 3D motion of two interacting camera wearers in a shared reference frame from synchronized dual-egocentric videos and calibrated camera geometry, covering reconstruction from complete observations and forecasting from an observed prefix.
- **MutualEgo-10K dataset.** We introduce MutualEgo-10K, a dataset of 10,105 interaction sequences with synchronized rendered dual-egocentric RGB, calibrated camera geometry, and paired 3D motion, supporting training and evaluation under both settings.
- **MutualEgo framework.** We propose MutualEgo, a pair-centric diffusion framework that jointly estimates both participants' motion. Its Reference-Aware Motion Tokenizer conditions dense visual features on shared-frame camera geometry and aggregates them into global and local conditioning tokens through anchor-guided attention, enabling the denoiser to integrate complementary evidence from both views.

2 Related Work

2.1 Egocentric Human Motion Estimation

Egocentric methods mainly recover one wearer from body-facing imagery (Rhodin et al., 2016; Xu et al., 2019; Tome et al., 2019; Akada et al., 2022; 2024; Millerdurai et al., 2024; Yang et al., 2024; Li et al., 2026; Wang et al., 2022; 2023), sparse head/hand sensing (Jiang et al., 2022; Du et al., 2023; Aliakbarian et al., 2023; Jiang et al., 2024), or forward-facing and multimodal inputs (Camiletto et al., 2025; Lee et al., 2025; Wang et al., 2025a; Tran et al., 2025; Lee & Joo, 2024). Our experimental baselines EgoEgo, EgoAllo, and UniEgoMotion exemplify recent full-body recovery from egocentric observations (Li et al., 2023; Yi et al., 2025; Patel et al., 2025). Interaction-aware methods infer the wearer from partner cues (Ng et al., 2020; Scofano et al., 2025) or reconstruct the visible partner (Zhang et al., 2022; 2023b); multi-wearer perception and distributed reconstruction such as EgoExoMoCap broaden observation coverage (Khirodkar et al., 2023; Yang et al., 2026; Jiang et al., 2026). Related world-grounded and multiview methods reason about global human motion and camera geometry from external observations (Yuan et al., 2022; Sun et al., 2023; Wang et al., 2024; Shin et al., 2024; Shen et al., 2024; Müller et al., 2025; Choudhury et al., 2023). Most egocentric formulations still predict one wearer, one visible person, or independently defined subjects, rather than jointly treating two interacting camera wearers as one coupled motion target in a shared frame.

2.2 Human–Human Interaction and Generative Motion Modeling

Interaction research treats the pair as the modeling unit: reconstruction constrains relative placement and contact through learned proxemics and physics (Fieraru et al., 2020; 2025; Müller et al., 2024; Huang et al., 2024; Ugrinovic et al., 2024), while generation and reaction synthesis capture complementary roles and action–reaction dependencies (Liang et al., 2024; Ghosh et al., 2024; Xu et al., 2024b; Wang et al., 2025b; Javed et al., 2025; Geng et al., 2026). ExPI and PGformer extend pair dynamics to forecasting from 3D histories (Guo et al., 2022; Fang et al., 2025). General diffusion formulations and Transformer denoisers provide generative modeling foundations (Ho et al., 2020; Nichol & Dhariwal, 2021; Peebles & Xie, 2023); motion-specific methods address prediction, generation, completion, and probabilistic forecasting (Mao et al., 2020; Salzmann et al., 2022; Tevet et al., 2023; Zhang et al., 2023a; 2024; Karunratanakul et al., 2023; Chen et al., 2023; Barquero et al., 2023; Curreli et al., 2025). These priors support inferring unobserved motion while accounting for uncertainty. However, interaction models typically condition pair dynamics on external imagery, text, audio, partner motion, or observed 3D histories rather than complementary moving first-person views carried by both predicted participants.

2.3 Egocentric and Interaction Motion Datasets

As Table 1 summarizes, egocentric resources offer overlapping subsets of RGB, exocentric views, calibration, and motion annotations, complemented by multimodal wearable collections (Ma et al., 2024; Fan et al., 2025). Ego-Humans, Harmony4D, and CoMind provide multi-person or dual-view observations, while InterVLA studies instructed human collaboration using synchronized dual-egocentric and exocentric observations (Xu et al., 2025). Interaction resources provide social capture, paired motion, and richer pair structure in practice (Joo et al., 2015; Fieraru et al., 2025; Fang et al.,

2024; Xu et al., 2024a; Ho et al., 2025), with Dual-Human and Inter-X supplying our primary motion sources. MutualEgo-10K complements these resources with a common rendering and calibration protocol for paired-interaction reconstruction and forecasting.

Table 1: Dataset comparison by participants per sequence, scene coverage, egocentric RGB, synchronized multi-participant egocentric RGB, exocentric RGB, motion-capture data, and dataset size. Frame counts include all reported RGB camera streams. ✓, ○, and × denote a reported capability, restricted multi-wearer coverage, and no support, respectively.

Dataset	People	Scenes	Ego	Multi-Ego	Exo	MoCap	# Seq.	# RGB Frames
UnrealEgo (Akada et al., 2022)	Multi	Multi	✓	×	×	×	∼4.6k	900k
EgoBody (Zhang et al., 2022)	Dual	Multi	✓	×	✓	✓	125	199.1k
Aria Digital Twin (Pan et al., 2023)	Mixed	Multi	✓	○	×	✓	236	–
Ego-Humans (Khirodkar et al., 2023)	Multi	Multi	✓	✓	✓	✓	7	571k
Ego-Exo4D (Grauman et al., 2024)	Single	Multi	✓	×	✓	×	5k	–
CoMind (Gavryushin et al., 2026)	Dual	Multi	✓	✓	✓	×	80	–
Harmony4D (Khirodkar et al., 2024)	Dual	Multi	✓	✓	✓	✓	208	1.66M
CHI3D (Fieraru et al., 2020)	Dual	Single	×	×	✓	✓	631	485.8k
Hi4D (Yin et al., 2023)	Dual	Single	×	×	✓	✓	100	∼88k
MutualEgo-10K	Dual	Multi	✓	✓	✓	✓	**10k**	**4.20M**

3 METHOD

We consider egocentric dual-human motion estimation in two complementary settings. Reconstruction recovers the paired motion over a fully observed dual-egocentric window, whereas forecasting observes only a prefix of the dual-egocentric images and camera conditions and estimates the full paired motion sequence, reconstructing the observed history and predicting the unobserved future. Both settings treat the interacting pair as one prediction unit in a shared reference frame. Within this formulation, MutualEgo is our pair-centric estimator, while MutualEgo-10K supplies aligned observations and paired motion targets for training and evaluation. The complete notation and conditional task definitions are given in Sec. A in the Appendix.

3.1 MUTUALEGO OVERVIEW

The architecture of MutualEgo is illustrated in Figure 2. We represent each interaction in a gravity-aligned pair frame defined by the initial egocentric view of participant A, and express both participants and camera trajectories in this shared coordinate system. The state of each participant comprises global body pose, root motion, hand pose, foot-contact state, and body shape. Concatenating the two 243D participant states in fixed A/B order yields one 486D pair-motion token per frame, projected to 768D for denoising. Detailed motion and camera representations are provided in Sec. D.1 in the Appendix. The pair-frame construction uses synchronized camera trajectories in a common metric world frame with known gravity and floor geometry.

Two complementary streams condition denoising. Frozen DINOv3 ViT-B/16 extracts dense dual-egocentric patch features. RAMT grounds them in the shared pair frame using calibrated rays and poses, while the trajectory branch supplies explicit temporal motion of the body-coupled cameras. RAMT compresses each view into one global and four softly localized motion tokens, then forms a structured dual-view observation memory $\mathbf{V}$ by validity-aware temporal alignment. Noisy pair-motion tokens, conditioned by dual-camera trajectories, form temporal queries. A temporal diffusion Transformer alternates temporal self-attention and cross-attention to the complete RAMT memory. A shared linear head predicts each clean pair state, jointly denoising both bodies and their evolving relation.

3.2 REFERENCE-AWARE MOTION TOKENIZER

RAMT uses the shared pair frame as its reference. For time t, participant-mounted view $p \in \{A, B\}$, and patch n, *Geometry-Conditioned Feature Modulation* (GCFM) modulates the DINOv3 feature

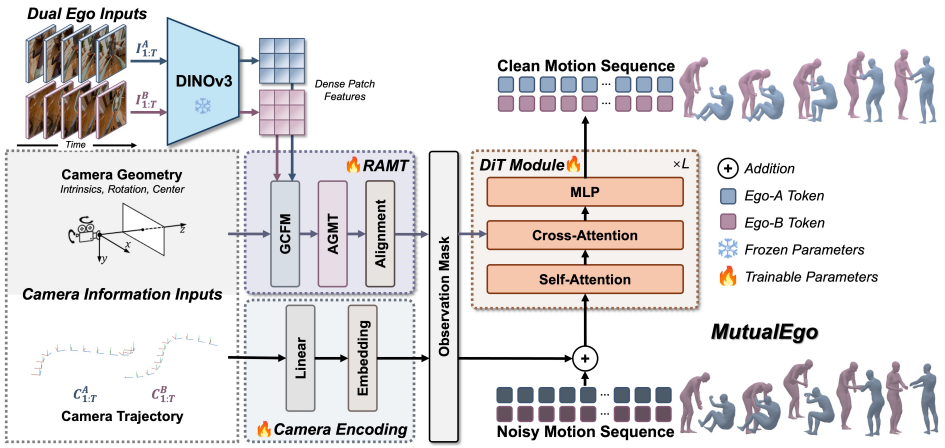


Figure 2: Overall architecture of MutualEgo. Frozen DINOv3 provides dual-egocentric patch features; RAMT applies GCFM and AGMT to form geometry-aware global and local memory. A separate camera encoder conditions noisy pair-motion queries. The temporal diffusion Transformer alternates temporal self-attention and cross-attention to the observation-masked dual-view memory, jointly predicting the clean pair sequence. The blue and purple motion blocks depict the A/B components concatenated into one pair-motion token per frame.

$\mathbf{F}_{t,p,n}$ using normalized image location $\bar{\mathbf{u}}_n$, ray $\mathbf{d}^q_{t,p,n}$, camera center $\mathbf{o}^q_{t,p}$, and orientation $R^q_{c_p,t}$:

$$\begin{aligned}\mathbf{g}_{t,p,n} &= [\bar{\mathbf{u}}_n; \mathbf{d}^q_{t,p,n}; \mathbf{o}^q_{t,p}; \rho_6(R^q_{c_p,t})],\\ \mathbf{z}^q_{t,p,n} &= \mathrm{FiLM}(\mathrm{LN}(W_F\mathbf{F}_{t,p,n}); \mathbf{g}_{t,p,n}) + \mathbf{e}^{\mathrm{time}}_t + \mathbf{e}^{\mathrm{view}}_p,\end{aligned} \tag{1}$$

Superscript q marks pair-frame geometry and the resulting reference-conditioned features. Each ray is back-projected from its patch center using calibrated camera intrinsics, then rotated into the shared pair frame. Here, ρ_6 is the continuous 6D rotation representation (Zhou et al., 2019), FiLM (Perez et al., 2018) applies learned channel-wise geometry modulation, and LN denotes layer normalization.

Anchor-Guided Motion Tokenization (AGMT) aggregates the conditioned patches with learned queries $\mathbf{q}_k$. Each local query uses a learnable image-plane anchor $\mathbf{a}_k$ and positive scale s_k to bias attention:

$$\begin{aligned}b_{k,n} &= -s_k\|\mathbf{a}_k - \bar{\mathbf{u}}_n\|_2^2,\\ \alpha_{k,n} &= \mathrm{softmax}_n\left(\frac{(W_Q\mathbf{q}_k)^\top(W_K\mathbf{z}^q_{t,p,n})}{\sqrt{d_h}} + b_{k,n}\right),\\ \mathbf{v}_{t,p,k} &= W_{\mathrm{out}}\,\mathrm{LN}\left(\mathbf{q}_k + W_O\sum_n \alpha_{k,n}W_V\mathbf{z}^q_{t,p,n}\right).\end{aligned} \tag{2}$$

Here, W denotes learned projections and d_h the head width; head concatenation is implicit. The global query uses $b_{0,n} = 0$. AGMT compresses the geometry-conditioned patches from each view into one global token and four local tokens with separately learned spatial biases. Anchors provide soft spatial priors: every query can attend to every patch, with initialization encouraging different spatial preferences. Time, view, and role embeddings and validity-aware alignment yield $\mathbf{V}$; Sec. B.3 in the Appendix describes the implementation and temporal alignment.

3.3 Pair Motion Estimation with Diffusion

Writing the clean paired sequence as $\mathbf{X}_0$, a Gaussian forward process (Ho et al., 2020) with a cosine schedule (Nichol & Dhariwal, 2021) corrupts it at diffusion step τ as

$$\mathbf{X}_\tau = \sqrt{\bar{\alpha}_\tau}\mathbf{X}_0 + \sqrt{1-\bar{\alpha}_\tau}\boldsymbol{\epsilon}, \qquad \widehat{\mathbf{X}}_0 = f_\theta(\mathbf{X}_\tau, \tau, \mathbf{c}_{1:T}, \mathbf{V}), \tag{3}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$ and $\mathbf{c}_{1:T}$ is the dual-camera trajectory conditioning sequence. The diffusion timestep conditions both the temporal queries and flattened RAMT memory. The denoiser predicts

the clean sequence using valid-element MSE,

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{X}_0,\tau,\boldsymbol{\epsilon}}\big[\|\mathbf{X}_0 - \widehat{\mathbf{X}}_0\|_M^2\big], \tag{4}$$

where $\|\cdot\|_M^2$ denotes masking and averaging squared errors over valid motion elements. Reconstruction conditions on the complete observation window and denoises all valid pair states. Forecasting conditions on dual-egocentric RGB and calibrated camera geometry from the observed prefix. Learned placeholder tokens replace visual and trajectory conditions for future frames. Starting from Gaussian noise, the sampler jointly denoises the full pair-motion window, estimating the historical motion and predicting the future. Details of the diffusion sampler, loss mask, and future-condition masking protocol are provided in Sec. D in the Appendix.

3.4 MUTUALEGO-10K DATASET

To support training and evaluation for egocentric dual-human motion estimation, we construct MutualEgo-10K with synchronized participant-mounted views, calibrated moving-camera geometry, and paired interaction motion on one timeline. It draws primarily from Dual-Human and Inter-X (Fang et al., 2024; Xu et al., 2024a), supplemented with two-person motion-capture recordings. We harmonize body and hand parameterizations, frame rates, body shapes, and coordinate conventions jointly for both participants, preserving the original pair configuration and interaction geometry. Each egocentric camera follows a per-frame upper-body coordinate frame derived from the head, neck, and shoulders of the corresponding participant; a synchronized dynamic exocentric camera provides a reference view. Both bodies and all cameras share a world frame, and Blender Cycles renders synchronized RGB, depth, and instance masks from their shared states. Each rendered frame has calibrated camera intrinsics and pose, temporally aligned with dual-egocentric RGB and paired motion. The dataset contains 10,105 interaction sequences yielding approximately 35k clips, with aligned motion labels. Under reconstruction, dual-egocentric RGB and calibrated egocentric camera geometry are observed over the full window; under forecasting, they are observed only over the prefix. Figure 3(a) shows the construction of MutualEgo-10K dataset, while Figure 3(b), (c), and (d) visualize synchronized exocentric RGB frames, dual-egocentric RGB frames, and paired motion. Sec. C in the Appendix documents dataset construction, modality definitions, and exact statistics.

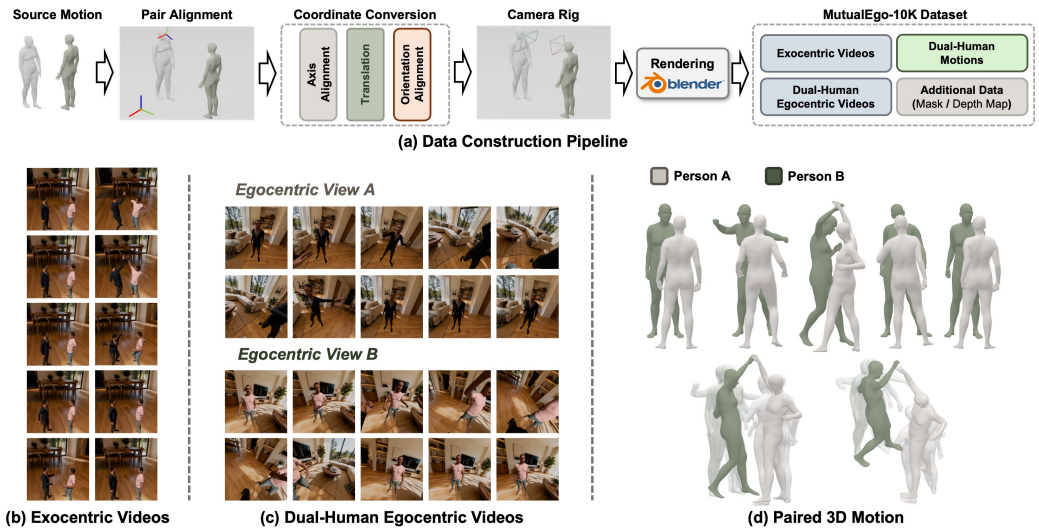


Figure 3: MutualEgo-10K construction and synchronized modalities. (a) Paired source motions are jointly standardized and rendered with two participant-mounted cameras and an exocentric reference camera. Visualizations show (b) exocentric RGB frames, (c) dual-egocentric RGB frames, and (d) paired SMPL-X motion. Gray and green meshes denote participants A and B.

4 EXPERIMENTS

We evaluate egocentric dual-human motion estimation on MutualEgo-10K across reconstruction and forecasting through comparisons with existing egocentric methods, formulation and component ablations, training-data scaling, and qualitative results. We additionally evaluate MutualEgo after adaptation to Harmony4D.

4.1 EXPERIMENTAL SETTINGS

MutualEgo-10K contains 35,001 clips from an 8:2 source-interaction split, comprising 28,059 training clips and 6,942 validation clips. Both tasks use 40-frame clips at 10 fps. Reconstruction uses the fully observed window. Forecasting observes only the first 10 frames of dual-egocentric RGB and camera conditions and estimates the full 40-frame motion sequence, reconstructing the first 10 frames and predicting the remaining 30 frames. All forecasting results are evaluated over the full 40-frame window. On MutualEgo-10K, our reconstruction and forecasting models, including ablation variants and data-scaling runs, are trained for 350 epochs on a single NVIDIA RTX 6000 Ada GPU using AdamW with a batch size of 64 and an initial learning rate of 3×10^{-5}. We also fine-tune MutualEgo on Harmony4D (Khirodkar et al., 2024) for 350 epochs with a learning rate of 10^{-5}. Full implementation and evaluation details appear in Sec. D in the Appendix.

4.2 COMPARISON WITH EGOCENTRIC METHODS

Table 2 compares MutualEgo with EgoEgo (Li et al., 2023), EgoAllo (Yi et al., 2025), and UniEgoMotion (Patel et al., 2025) with released pretrained weights. We additionally retrain UniEgoMotion on MutualEgo-10K for 350 epochs per task using its original model configuration and visual backbone. Each participant is estimated independently from their egocentric observations and camera trajectory, after which the two predictions are mapped into the shared pair frame following Sec. D.7.

Table 2: Comparison of egocentric motion estimators on MutualEgo-10K. EgoEgo, EgoAllo, and UniEgoMotion use released pretrained weights, while UniEgoMotion, *Retrained* is trained on MutualEgo-10K using the original UniEgoMotion configuration. Spatial mapping is applied after prediction to express both participants in the shared pair frame. MutualEgo is trained on MutualEgo-10K. Distances are in meters and rotations in degrees. **Bold** and underlined values denote the best and second-best results within each task, respectively.

Method	MPJPE ↓	PA-MPJPE ↓	Rel. MPJPE ↓	Rel. Trans. ↓	Rel. Rot. ↓	Head Trans. ↓	Head Rot. ↓	Foot F1 ↑
Reconstruction								
EgoEgo (Li et al., 2023)	0.274	0.118	0.430	0.366	29.9	0.187	6.7	0.774
EgoAllo (Yi et al., 2025)	0.310	0.139	0.419	0.374	36.7	0.224	6.4	0.765
UniEgoMotion (Patel et al., 2025)	0.162	0.085	0.229	0.207	19.0	0.106	6.2	0.882
UniEgoMotion, *Retrained*	0.261	0.052	0.305	0.284	12.8	0.247	6.5	0.900
MutualEgo	**0.081**	**0.043**	**0.088**	**0.058**	**10.7**	**0.058**	**4.8**	**0.905**
Forecasting								
UniEgoMotion (Patel et al., 2025)	0.429	0.107	0.529	0.508	40.1	0.429	41.9	**0.840**
UniEgoMotion, *Retrained*	0.445	0.082	0.547	0.519	31.8	0.443	25.7	0.821
MutualEgo	**0.287**	**0.073**	**0.329**	**0.289**	**29.1**	**0.282**	**23.8**	0.832

Reconstruction. Among the methods evaluated with released pretrained weights, UniEgoMotion obtains the lowest MPJPE and relative MPJPE, at 0.162 m and 0.229 m, respectively. Retraining UniEgoMotion on MutualEgo-10K improves PA-MPJPE from 0.085 m to 0.052 m and relative rotation error from 19.0° to 12.8°, while its translation-sensitive errors increase. In comparison, MutualEgo achieves an MPJPE of 0.081 m and a relative MPJPE of 0.088 m. It further reduces relative translation error to 0.058 m and head translation error to 0.058 m, while obtaining relative and head rotation errors of 10.7° and 4.8°, respectively.

Forecasting. Retraining UniEgoMotion on MutualEgo-10K improves PA-MPJPE from 0.107 m to 0.082 m and reduces relative and head rotation errors from 40.1° and 41.9° to 31.8° and 25.7°, respectively, while its translation-sensitive errors remain close to those of the released model. MutualEgo reduces MPJPE and relative MPJPE to 0.287 m and 0.329 m, respectively. It further lowers relative translation error to 0.289 m and relative rotation error to 29.1°. The released UniEgoMotion model obtains the highest Foot F1 of 0.840, while MutualEgo achieves 0.832.

4.3 Ablation Study

We organize the ablation around dual-view pair modeling and reference-aware motion tokenization. The single-view variant jointly predicts both participants from view A and its camera trajectory. The independent variant applies a shared-weight estimator separately to each participant, conditioned on their own view and camera trajectory, and combines the two predictions in the shared pair frame. The joint baseline, MutualEgo w/o RAMT, uses both views to predict the paired motion jointly. These variants use the same MutualEgo-10K split and DINOv3 ViT-B/16 visual features. We then ablate the components of RAMT within the joint estimator. Training details and complete variant metrics are provided in Sec. E.4 in the Appendix.

Table 3: Ablation of dual-view and joint pair modeling on the MutualEgo-10K dataset. Distances are in meters. **Bold** and underlined values denote the best and second-best results within each task, respectively.

	Reconstruction			Forecasting		
Variant	MPJPE ↓	Rel. MPJPE ↓	Rel. Trans. ↓	MPJPE ↓	Rel. MPJPE ↓	Rel. Trans. ↓
MutualEgo w/o RAMT, *Single-View*	0.226	0.313	0.274	0.461	0.553	0.521
MutualEgo w/o RAMT, *Independent*	0.118	0.164	0.125	<u>0.345</u>	0.425	0.387
MutualEgo w/o RAMT	<u>0.116</u>	<u>0.117</u>	<u>0.067</u>	0.347	<u>0.375</u>	<u>0.336</u>
MutualEgo	**0.081**	**0.088**	**0.058**	**0.287**	**0.329**	**0.289**

Pair formulation. Table 3 indicates that one egocentric view provides insufficient evidence for reliable pair recovery. The single-view variant yields reconstruction MPJPE of 0.226 m and relative MPJPE of 0.313 m; the corresponding forecasting errors reach 0.461 m and 0.553 m. Independent prediction substantially improves individual accuracy, with reconstruction MPJPE of 0.118 m approaching the 0.116 m achieved by the joint baseline. Its forecasting MPJPE is slightly lower at 0.345 m, compared with 0.347 m for the joint baseline. The dual-view joint baseline reduces relative MPJPE from 0.164 to 0.117 m in reconstruction and from 0.425 to 0.375 m in forecasting. Relative translation also decreases from 0.125 to 0.067 m and from 0.387 to 0.336 m, respectively. The independent and joint variants both use the two participant views, but only the joint variant exchanges information across views and predicts a single pair state. MutualEgo further improves both individual and relative accuracy over the joint baseline.

Table 4: Component ablation on the MutualEgo-10K dataset. Distances are in meters and rotations in degrees. **Bold** and underlined values denote the best and second-best results within each task, respectively.

Dense DINO	GCFM	AGMT	MPJPE ↓	PA-MPJPE ↓	Rel. MPJPE ↓	Rel. Trans. ↓	Rel. Rot. ↓	Head Trans. ↓	Head Rot. ↓	Foot F1 ↑
Reconstruction										
			<u>0.116</u>	<u>0.062</u>	<u>0.117</u>	<u>0.067</u>	<u>13.3</u>	<u>0.080</u>	<u>5.7</u>	<u>0.881</u>
✓			0.222	0.075	0.148	0.089	15.7	0.188	7.0	0.863
✓	✓		0.139	0.069	0.138	0.088	14.7	0.100	6.6	0.867
✓		✓	0.223	0.078	0.150	0.089	16.3	0.189	7.0	0.857
✓	✓	✓	**0.081**	**0.043**	**0.088**	**0.058**	**10.7**	**0.058**	**4.8**	**0.905**
Forecasting										
			0.347	0.079	0.375	0.336	32.0	0.342	25.8	0.818
✓			0.357	0.080	<u>0.370</u>	0.329	32.2	0.351	26.1	<u>0.828</u>
✓	✓		<u>0.317</u>	<u>0.076</u>	0.371	0.335	<u>30.6</u>	<u>0.310</u>	<u>25.2</u>	0.822
✓		✓	0.372	0.079	0.370	<u>0.327</u>	31.3	0.366	25.3	0.824
✓	✓	✓	**0.287**	**0.073**	**0.329**	**0.289**	**29.1**	**0.282**	**23.8**	**0.832**

Component ablation. Table 4 evaluates the contribution of geometric conditioning and structured token aggregation within the joint estimator. The dense-feature variants retain five tokens per view and frame while varying GCFM and AGMT. Without either component, dense features increase reconstruction MPJPE from 0.116 m for the pooled-feature baseline to 0.222 m. Adding GCFM reduces this error by 37.3%, whereas AGMT alone provides no improvement. Using both components improves every reconstruction metric over either alone, indicating that structured aggregation benefits from geometry-conditioned features. For forecasting, GCFM reduces MPJPE from 0.357 to 0.317 m, while AGMT alone has small, mixed effects. Combining GCFM and AGMT further reduces MPJPE to 0.287 m and improves all eight metrics over either component alone, showing that geometric conditioning and structured aggregation are complementary. MutualEgo achieves the lowest MPJPE in both tasks. The input-gradient sensitivity maps in Sec. E.1 in the Appendix show stronger responses

around visible people and interaction-relevant regions for MutualEgo than for MutualEgo w/o RAMT in the illustrated examples.

4.4 Additional Evaluations

To assess the effect of training-data volume, we train on label-stratified 20%, 40%, 60%, 80%, and 100% subsets and evaluate all models on the same MutualEgo-10K validation set. From 20% to 100%, reconstruction MPJPE falls from 0.134 m to 0.081 m and full-window forecasting MPJPE from 0.401 m to 0.287 m. All positional errors decrease monotonically; forecasting head-rotation error and Foot F1 are non-monotonic. Sec. E.2 in the Appendix reports the complete scaling results.

To study adaptation to real capture, we initialize MutualEgo from its MutualEgo-10K checkpoint and fine-tune on Harmony4D. Reconstruction MPJPE is 0.3049 m versus 0.3114 m for EgoEgo, and PA-MPJPE is 0.0947 m versus 0.1267 m for UniEgoMotion; Foot F1 is 0.5910 versus 0.5405 for UniEgoMotion. Pair-relative results are mixed: EgoEgo has lower relative MPJPE and translation error, while UniEgoMotion has lower relative rotation and head-translation errors. Sec. D.5 gives the fine-tuning protocol and Sec. E.3 reports the full table.

4.5 Qualitative Results

Figure 4 shows a hugging interaction. In this example, MutualEgo more closely matches the ground-truth arm placement and relative body configuration during reconstruction, and its forecast follows the illustrated approach toward a hug. The close-up egocentric views show limited body coverage. Additional examples appear in Sec. E.6.

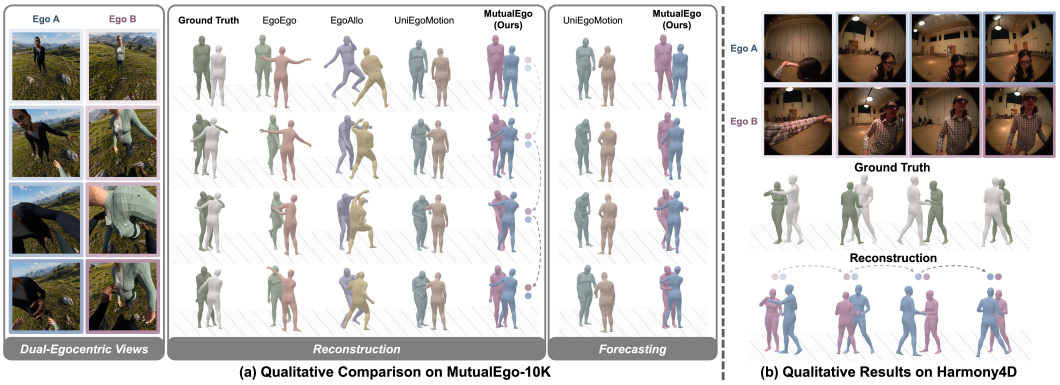


Figure 4: (a) Qualitative comparison on MutualEgo-10K; time progresses from top to bottom. (b) Results on Harmony4D; time progresses from left to right.

5 Conclusion

We formulate egocentric dual-human motion estimation as joint motion estimation of two interacting camera wearers in a shared reference frame, covering both reconstruction and forecasting from complementary first-person observations. MutualEgo addresses this task by treating the interacting pair as a single prediction unit and using RAMT to ground dense dual-view appearance in calibrated camera geometry. MutualEgo-10K provides synchronized dual-egocentric observations, calibrated camera geometry, and paired 3D interaction motion for training and evaluation. Experiments on MutualEgo-10K show lower body and pair-relative errors than both released pretrained estimators and UniEgoMotion trained on MutualEgo-10K, while formulation and component ablations characterize the effects of dual-view joint modeling and reference-aware motion tokenization. Training-data scaling and adaptation to Harmony4D further evaluate the approach across data scales and capture settings. Together, these results establish a foundation for joint motion estimation from complementary egocentric views during human–human interaction.

AI USE STATEMENT

In this work, generative AI tools were used only to aid and polish the writing of this paper, including language refinement and improvements to clarity and presentation. They were not used to conduct experiments or determine scientific conclusions. All AI-assisted text was reviewed and revised by the authors. We take responsibility for the final content of this work.

REFERENCES

Hiroyasu Akada, Jian Wang, Soshi Shimada, Masaki Takahashi, Christian Theobalt, and Vladislav Golyanik. UnrealEgo: A new dataset for robust egocentric 3d human motion capture. In *European Conference on Computer Vision*, pp. 1–17, 2022. doi: 10.1007/978-3-031-20068-7_1.

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*, pp. 767–776, 2024.

Sadegh Aliakbarian, Fatemeh Saleh, David Collier, Pashmina Cameron, and Darren Cosker. HMD-NeMo: Online 3d avatar motion generation from sparse observations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9622–9631, 2023.

German Barquero, Sergio Escalera, and Cristina Palmero. BeLFusion: Latent diffusion for behavior-driven human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2317–2327, 2023.

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*, pp. 17497–17507, 2025.

Ling-Hao Chen, Jiawei Zhang, Yewen Li, Yiren Pang, Xiaobo Xia, and Tongliang Liu. HumanMAC: Masked motion completion for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9544–9555, 2023.

Rohan Choudhury, Kris M. Kitani, and László A. Jeni. TEMPO: Efficient multi-view pose estimation, tracking, and forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14750–14760, 2023.

Cecilia Curreli, Dominik Muhle, Abhishek Saroha, Zhenzhang Ye, Riccardo Marin, and Daniel Cremers. Nonisotropic gaussian diffusion for realistic 3d human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1871–1882, 2025.

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*, pp. 481–490, 2023.

Zhen Fan, Peng Dai, Zhuo Su, Xu Gao, Zheng Lv, Jiarui Zhang, Tianyuan Du, Guidong Wang, and Yang Zhang. EMHI: A multimodal egocentric human motion dataset with HMD and body-worn IMUs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(3):2879–2887, 2025. doi: 10.1609/aaai.v39i3.32294.

Qi Fang, Yinghui Fan, Yanjun Li, Junting Dong, Dingwei Wu, Weidong Zhang, and Kang Chen. Capturing closely interacted two-person motions with reaction priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 655–665, 2024.

Yanwen Fang, Wenqi Jia, Xu Cao, Peng-Tao Jiang, Guodong Li, and Jintai Chen. Proxy-bridged game transformer for interactive extreme motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13912–13921, 2025.

Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7214–7223, 2020.

Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Reconstructing three-dimensional models of interacting humans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(12):10870–10881, 2025. doi: 10.1109/TPAMI.2025.3601974.

Alexey Gavryushin, Dingxi Zhang, Zhao Huang, Alexandros Delitzas, Jiaqi Chen, Ben Ellis, Cedric Zöllner, Manthan Patel, Manuel Kaufmann, Marc Pollefeys, and Xi Wang. CoMind: Understanding collaborative human activity from multiple minds and views. In *Proceedings of the European Conference on Computer Vision*, pp. 502–523, 2026. doi: 10.1007/978-3-032-37550-6_29.

Zichen Geng, Zeeshan Hayder, Bo Miao, Jian Liu, Wei Liu, and Ajmal Mian. Disentangled hierarchical VAE for 3d human-human interaction generation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=53eIDko6N5.

Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. ReMoS: 3d motion-conditioned reaction synthesis for two-person interactions. In *European Conference on Computer Vision*, pp. 418–437, 2024. doi: 10.1007/978-3-031-72764-1_24.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J. Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, David Crandall, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C. V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19383–19400, 2024.

Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Multi-person extreme motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13053–13064, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Leo Ho, Yinghao Huang, Dafei Qin, Mingyi Shi, Wangpok Tse, Wei Liu, Junichi Yamagishi, and Taku Komura. InterAct: A large-scale dataset of dynamic, expressive and interactive activities between two people in daily scenarios. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 8(4):1–27, 2025. doi: 10.1145/3747871.

Buzhen Huang, Chen Li, Chongyang Xu, Liang Pan, Yangang Wang, and Gim Hee Lee. Closely interactive human reconstruction with proxemics and physics-guided adaption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1011–1021, 2024.

Muhammad Gohar Javed, Chuan Guo, Li Cheng, and Xingyu Li. InterMask: 3d human interaction generation via collaborative masked modeling. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ZAyuwJYN8N.

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 *Proceedings of the European Conference on Computer Vision*, pp. 443–460, 2022. doi: 10.1007/978-3-031-20065-6_26.

Jiaxi Jiang, Paul Streli, Manuel Meier, and Christian Holz. EgoPoser: Robust real-time egocentric pose estimation from sparse and intermittent observations everywhere. In *Proceedings of the European Conference on Computer Vision*, pp. 277–294, 2024. doi: 10.1007/978-3-031-72627-9_16.

Jiaxi Jiang, Bharat Lal Bhatnagar, Nan Yang, Lingni Ma, Sebastian Starke, Robin Kips, Nadine Bertsch, Christian Holz, and Federica Bogo. EgoExoMoCap: Distributed ego-exo human motion capture. In *European Conference on Computer Vision*, 2026. URL `https://siplab.org/projects/EgoExoMoCap`.

Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3334–3342, 2015.

Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2151–2162, 2023.

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

Rawal Khirodkar, Jyun-Ting Song, Jinkun Cao, Zhengyi Luo, and Kris Kitani. Harmony4D: A video dataset for in-the-wild close human interactions. In *Advances in Neural Information Processing Systems*, volume 37, pp. 107270–107285, 2024. doi: 10.52202/079017-3407.

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

Jiye Lee and Hanbyul Joo. Mocap everyone everywhere: Lightweight motion capture with smartwatches and a head-mounted camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1091–1100, 2024.

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*, pp. 17142–17151, 2023.

Zhenyu Li, Sai Kumar Dwivedi, Filip Maric, Carlos Chacón, Nadine Bertsch, Filippo Arcadu, Tomas Hodan, Michael Ramamonjisoa, Peter Wonka, Amy Zhao, Robin Kips, Cem Keskin, Anastasia Tkach, and Chenhongyi Yang. EgoPoseFormer v2: Accurate egocentric human motion estimation for AR/VR. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21121–21131, 2026.

Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. InterGen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, 132(9):3463–3483, 2024. doi: 10.1007/s11263-024-02042-6.

Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, Kevin Bailey, David Soriano Fosas, C. Karen Liu, Ziwei Liu, Jakob Engel, Renzo De Nardi, and Richard Newcombe. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *Proceedings of the European Conference on Computer Vision*, pp. 445–465, 2024. doi: 10.1007/978-3-031-72691-0_25.

Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *Proceedings of the European Conference on Computer Vision*, pp. 474–489, 2020. doi: 10.1007/978-3-030-58568-6_28.

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*, pp. 1186–1195, 2024.

Lea Müller, Vickie Ye, Georgios Pavlakos, Michael J. Black, and Angjoo Kanazawa. Generative proxemics: A prior for 3d social interaction from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9687–9697, 2024.

Lea Müller, Hongsuk Choi, Anthony Zhang, Brent Yi, Jitendra Malik, and Angjoo Kanazawa. Reconstructing people, places, and cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21948–21958, 2025.

Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. You2Me: Inferring body pose in egocentric video via first and second person interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9890–9900, 2020.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171, 2021.

Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard 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*, 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*, pp. 10318–10329, 2025.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10975–10985, 2019.

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

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11671.

Helge Rhodin, Christian Richardt, Dan Casas, Eldar Insafutdinov, Mohammad Shafiei, Hans-Peter Seidel, Bernt Schiele, and Christian Theobalt. EgoCap: Egocentric marker-less motion capture with two fisheye cameras. *ACM Transactions on Graphics*, 35(6):1–11, 2016. doi: 10.1145/2980179.2980235.

Tim Salzmann, Marco Pavone, and Markus Ryll. Motron: Multimodal probabilistic human motion forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6457–6466, 2022.

Luca Scofano, Alessio Sampieri, Edoardo De Matteis, Indro Spinelli, and Fabio Galasso. Social egomesh estimation. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 5948–5958, 2025.

Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia Conference Papers*, pp. 144:1–144:11, 2024. doi: 10.1145/3680528.3687565.

Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing world-grounded humans with accurate 3d motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2070–2080, 2024.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. arXiv preprint arXiv:2508.10104, 2025.

Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J. Black. TRACE: 5d temporal regression of avatars with dynamic cameras in 3d environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8856–8866, 2023.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=SJ1kSyO2jwu.

Denis Tome, Patrick Peluse, Lourdes Agapito, and Hernan Badino. xR-EgoPose: Egocentric 3d human pose from an HMD camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7728–7738, 2019.

Minh Tran, Hongda Mao, Qingshuang Chen, and Yelin Kim. Head2Body: Body pose generation from multi-sensory head-mounted inputs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6849–6858, 2025.

Nicolas Ugrinovic, Boxiao Pan, Georgios Pavlakos, Despoina Paschalidou, Bokui Shen, Jordi Sanchez-Riera, Francesc Moreno-Noguer, and Leonidas Guibas. MultiPhys: Multi-person physics-aware 3d motion estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2331–2340, 2024.

Jian Wang, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, Diogo Luvizon, and Christian Theobalt. Estimating egocentric 3d human pose in the wild with external weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13157–13166, 2022.

Jian Wang, Diogo 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*, pp. 13031–13040, 2023.

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*, pp. 22668–22679, 2025a.

Yabiao Wang, Shuo Wang, Jiangning Zhang, Ke Fan, Jiafu Wu, Zhucun Xue, and Yong Liu. TIMotion: Temporal and interactive framework for efficient human-human motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7169–7178, 2025b.

Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. TRAM: Global trajectory and motion of 3d humans from in-the-wild videos. In *Proceedings of the European Conference on Computer Vision*, pp. 467–487, 2024. doi: 10.1007/978-3-031-73247-8_27.

Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, Yunhui Liu, Wenjun Zeng, and Xiaokang Yang. Inter-X: Towards versatile human-human interaction analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22260–22271, 2024a.

Liang Xu, Yizhou Zhou, Yichao Yan, Xin Jin, Wenhan Zhu, Fengyun Rao, Xiaokang Yang, and Wenjun Zeng. ReGenNet: Towards human action-reaction synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1759–1769, 2024b.

Liang Xu, Chengqun Yang, Zili Lin, Fei Xu, Yifan Liu, Congsheng Xu, Yiyi Zhang, Jie Qin, Xingdong Sheng, Yunhui Liu, Xin Jin, Yichao Yan, Wenjun Zeng, and Xiaokang Yang. Perceiving and acting in first-person: A dataset and benchmark for egocentric human-object-human interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12535–12548, 2025.

Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Pascal Fua, Hans-Peter Seidel, and Christian Theobalt. Mo2Cap2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera. *IEEE Transactions on Visualization and Computer Graphics*, 25(5):2093–2101, 2019. doi: 10.1109/TVCG.2019.2898650.

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*, pp. 401–417, 2024. doi: 10.1007/978-3-031-73001-6_23.

Nan Yang, Julian Straub, Fan Zhang, Richard Newcombe, Jakob Engel, and Lingni Ma. LAMP: Localization aware multi-camera people tracking in metric 3d world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21208–21220, 2026.

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*, pp. 7072–7084, 2025.

Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4D: 4d instance segmentation of close human interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17016–17027, 2023.

Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. GLAMR: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11038–11049, 2022.

Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. ReMoDiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 364–373, 2023a.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. MotionDiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4115–4128, 2024. doi: 10.1109/TPAMI.2024.3355414.

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*, pp. 180–200, 2022. doi: 10.1007/978-3-031-20068-7_11.

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*, pp. 7989–8000, 2023b.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5745–5753, 2019.

A Task Formulation

Given synchronized cameras worn by two interacting participants, the goal is to recover their complete paired motion in one common reference frame. The pair is the prediction unit: the model estimates two coupled motion streams and their evolving spatial relation rather than solving independent single-person problems. Let $p \in \{A, B\}$ index both a participant and the camera worn by that participant. We denote RGB by I_t^p and calibrated camera information by C_t^p, which comprises the intrinsics and camera-to-pair pose. The dual-egocentric observation and paired motion sequences are

$$\mathcal{O}_{1:T} = \{I_{1:T}^A, I_{1:T}^B, C_{1:T}^A, C_{1:T}^B\}, \qquad \mathbf{X}_{1:T} = \{(\mathbf{x}_t^A, \mathbf{x}_t^B)\}_{t=1}^T. \tag{A.1}$$

Both motion streams are expressed in a shared pair frame. $\mathcal{O}$ comprises dual-egocentric images and calibrated egocentric camera geometry.

In *reconstruction*, the complete observation window is available and the target is the paired motion over the same interval:

$$p_\theta(\mathbf{X}_{1:T} \mid \mathcal{O}_{1:T}). \tag{A.2}$$

In *forecasting*, dual-egocentric images and calibrated camera conditions are available only over the prefix $1 : T_o$. The model estimates the paired motion over the full sequence, reconstructing the observed history and predicting the unobserved future:

$$p_\theta(\mathbf{X}_{1:T} \mid \mathcal{O}_{1:T_o}). \tag{A.3}$$

Ground-truth motion is unavailable at inference, including over the observed prefix. Learned placeholders represent future visual and camera conditions.

B Reference-Aware Motion Tokenizer Details

RAMT converts dual-egocentric visual features and calibrated camera geometry into structured conditioning tokens for the motion denoiser. As shown in Figure B.1, it combines Geometry-Conditioned Feature Modulation (GCFM) with Anchor-Guided Motion Tokenization (AGMT). This section details these two components and the temporal alignment of their output tokens.

B.1 Geometry-Conditioned Feature Modulation

For patch-grid position (i, j) of the 16×16 DINOv3 grid, the pixel center is $u = (j + 0.5)16$, $v = (i + 0.5)16$ in the resized 256×256 image, and $\bar{\mathbf{u}} = (2u/W - 1, 2v/H - 1) \in [-1, 1]^2$. Let $\widetilde{\mathbf{u}} = (u, v, 1)^\top$, K_t^p be the effective intrinsics after resizing, and $R_{c_p,t}^q$ be the camera-to-pair rotation. Back-projection first produces an OpenCV ray; $S = \mathrm{diag}(1, -1, -1)$ converts it to the Blender camera convention before rotation into the pair frame:

$$\mathbf{d}_{t,p,n}^q = R_{c_p,t}^q \frac{S(K_t^p)^{-1}\widetilde{\mathbf{u}}_n}{\|S(K_t^p)^{-1}\widetilde{\mathbf{u}}_n\|_2}, \qquad \mathbf{g}_{t,p,n} = [\bar{\mathbf{u}}_n; \mathbf{d}_{t,p,n}^q; \mathbf{o}_{t,p}^q; \rho_6(R_{c_p,t}^q)] \in \mathbb{R}^{14}, \tag{B.1}$$

where $\mathbf{o}_{t,p}^q$ is the pair-frame camera center. Image location and ray direction vary across patches, while all patches from the same image share a camera center and orientation. The rays are derived from the calibrated rendering intrinsics, rescaled consistently with the image resolution.

DINOv3 patch features are projected from 768 to the 256D RAMT width and Layer-normalized. A two-layer geometry MLP and a two-layer FiLM MLP then perform the exact modulation

$$\begin{aligned} \mathbf{z}_{t,p,n} &= \mathrm{LN}(W_F \mathbf{F}_{t,p,n}), \\ \mathbf{h}_{t,p,n}^g &= W_{g,2}\,\mathrm{SiLU}(W_{g,1}\mathbf{g}_{t,p,n}), \\ [\boldsymbol{\gamma}_{t,p,n}; \boldsymbol{\beta}_{t,p,n}] &= W_{f,2}\,\mathrm{SiLU}(W_{f,1}\mathbf{h}_{t,p,n}^g), \\ \mathbf{z}_{t,p,n}^q &= \mathrm{LN}((1 + \boldsymbol{\gamma}_{t,p,n}) \odot \mathbf{z}_{t,p,n} + \boldsymbol{\beta}_{t,p,n}) + \mathbf{e}_t^{\mathrm{time}} + \mathbf{e}_p^{\mathrm{view}}. \end{aligned} \tag{B.2}$$

The geometry MLP has widths $14 \to 128 \to 256$, and the FiLM MLP has widths $256 \to 256 \to 512$. The final FiLM layer is initialized to zero, so $(1 + \boldsymbol{\gamma})$ starts near the projected appearance feature. A 256D sinusoidal absolute-time encoding and learned A/B view embedding are added after modulation. The two views are encoded independently inside RAMT; cross-view fusion occurs in the diffusion Transformer through the combined memory.

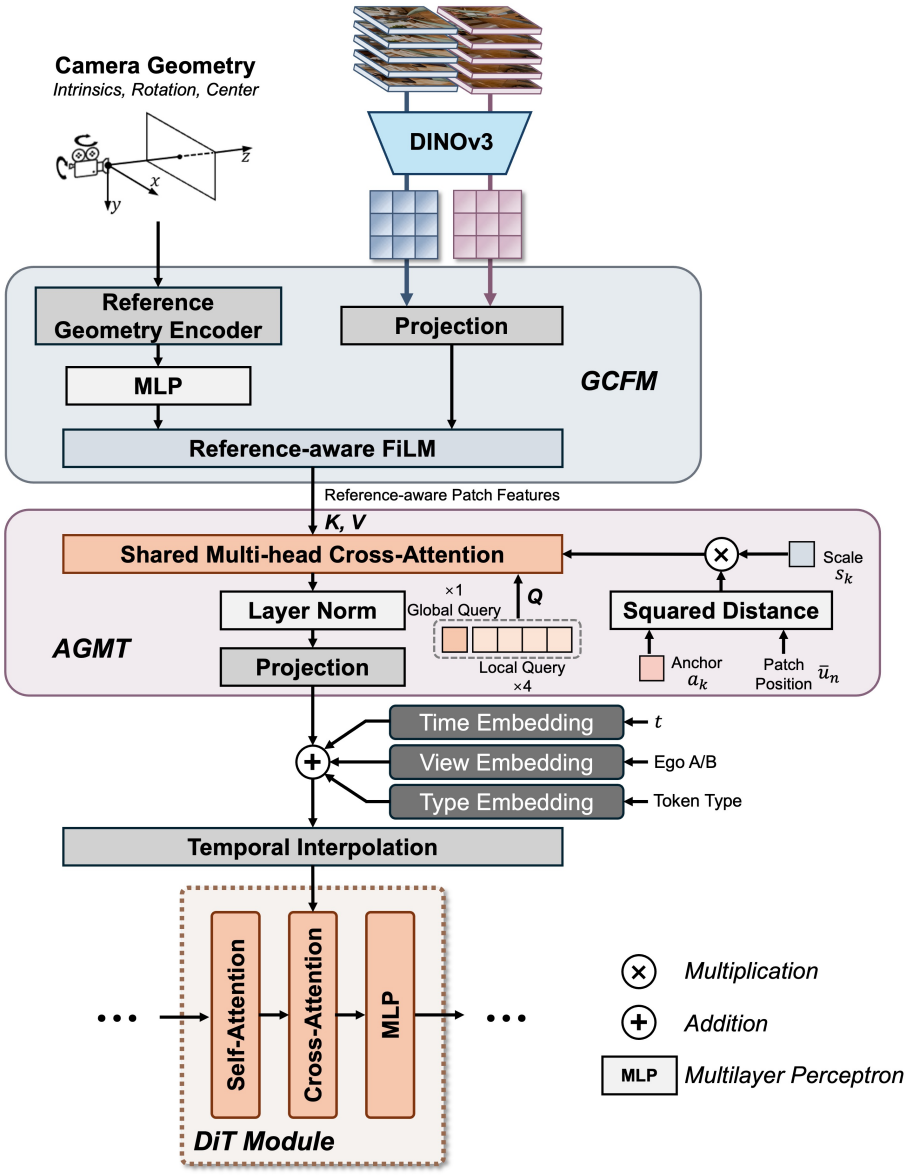


Figure B.1: Detailed RAMT architecture. Geometry-Conditioned Feature Modulation (GCFM) modulates dense visual patches with calibrated camera geometry through FiLM. Anchor-Guided Motion Tokenization (AGMT) uses one unbiased global query and four softly biased local queries to extract structured tokens. The distance branch supplies a negative spatial bias to the local-query attention logits, as defined in Eq. B.3. Anchors are learnable soft priors; view and type identities are preserved, temporal alignment is validity-aware, and the two views are encoded independently inside RAMT.

B.2 Anchor-Guided Motion Tokenization

For each source time and view, one learned global query and four learned local queries attend to all 256 conditioned patches. The local anchors are initialized in the four image quadrants at $(-0.5, -0.5)$, $(0.5, -0.5)$, $(-0.5, 0.5)$, and $(0.5, 0.5)$. Effective anchors are constrained to $[-1, 1]^2$ by $\tanh$, and their positive scales are $s_k = \mathrm{softplus}(\widehat{s}_k)$ with $\widehat{s}_k = 0$ at initialization. For each

attention head,

$$
\begin{aligned}
b_{k,n} &= -s_k \|\mathbf{a}_k - \bar{\mathbf{u}}_n\|_2^2, \\
\alpha_{k,n} &= \mathrm{softmax}_n\left(\frac{(W_Q \mathbf{q}_k)^\top (W_K \mathbf{z}^q_{t,p,n})}{\sqrt{d_h}} + b_{k,n}\right), \\
\mathbf{v}_{t,p,k} &= W_{\mathrm{out}}\,\mathrm{LN}\left(\mathbf{q}_k + W_O \sum_n \alpha_{k,n} W_V \mathbf{z}^q_{t,p,n}\right).
\end{aligned}
\tag{B.3}
$$

The global query uses $b_{0,n} = 0$. The soft spatial bias preserves access to the full patch field while encouraging each local query to specialize around its learned anchor. All five queries share query, key, value, attention-output, and final-output projections. Their roles differ through query values, local biases, and token-type embeddings. Attention has width 256 and eight heads; the five outputs are projected to 768D, then receive 768D time, A/B view, and global/local token-type embeddings.

B.3 Architecture, Masks, and Temporal Alignment

The frozen DINOv3 ViT-B/16 extracts a 16×16 grid of 768D patch features from each image. RGB images are resized to 256×256, with corresponding rescaling of camera intrinsics. For each time and view, RAMT summarizes the 256 patch features into five 768D tokens. Missing or padded observations are masked in both RAMT attention and denoiser cross-attention, and their token representations are set to zero.

Visual features are sampled with temporal stride two. For target motion frame t, let $l = \lfloor t/2 \rfloor$, $r = \min(l+1, T_v - 1)$, and $\alpha = (t \bmod 2)/2$, with T_v the number of visual observations and m_l, m_r their validity indicators. Interpolation is validity-aware:

$$
w_l = (1-\alpha) m_l, \quad w_r = \alpha m_r, \qquad \mathbf{V}_t = \frac{w_l \mathbf{V}_l + w_r \mathbf{V}_r}{\max(w_l + w_r, \varepsilon)}. \tag{B.4}
$$

Interpolation weights are normalized over valid observations. Tokens at invalid motion frames are masked after interpolation. Temporal alignment yields five tokens per view at each of the 40 motion frames. The resulting 400 visual tokens are ordered by time, view, and role; every motion query can cross-attend globally to this complete memory.

C MutualEgo-10K Dataset Details

C.1 Motion Sources and Standardization

MutualEgo-10K converts paired SMPL-X interaction motion (Pavlakos et al., 2019) into synchronized first-person visual–geometric data. The principal public sources are Dual-Human and Inter-X, supplemented by additional two-person motion-capture recordings. For Dual-Human, we retain interactions with motion available for both participants over a common time interval; Inter-X sequences preserve their original participant pairing. We harmonize source-specific pose and hand parameterizations, frame rates, body shapes, and coordinate conventions into a common motion representation before rendering. Temporal sampling and spatial conversion are applied jointly to both participants, preserving their relative placement and interaction geometry.

We distinguish full interaction sequences from the fixed-length motion clips used for learning and evaluation. Each sequence includes both participant-mounted egocentric views and an exocentric view, synchronized with paired motion and camera calibration. The source motions vary in duration, and many last less than five seconds. We therefore sample at 10 fps and use 40-frame windows to retain broad coverage of these shorter interactions while still supporting a 10-frame observation prefix and a 30-frame forecasting horizon. Windows are extracted with stride 10, and incomplete tails are excluded. Source interactions are partitioned 8:2 before windowing, with all clips from each interaction assigned to the same partition. This procedure yields 35,001 motion clips: 28,059 for training and 6,942 for validation.

C.2 Camera and Rendering Setup

Each egocentric camera is attached to a per-frame upper-body coordinate frame derived from the SMPL-X head, neck, and shoulder joints. With up vector $\mathbf{u}$ from neck to head and preliminary right vector $\mathbf{r}_0$ from left to right shoulder, the forward and re-orthogonalized right directions are

$$\mathbf{f} = \mathrm{norm}(\mathbf{u} \times \mathbf{r}_0), \qquad \mathbf{r} = \mathrm{norm}(\mathbf{f} \times \mathbf{u}). \tag{C.1}$$

For degenerate joint configurations, fixed world axes define the camera frame. Each egocentric camera is placed 0.12 m forward and 0.02 m above the head joint, with a 15° upward pitch and no additional yaw or roll. The two egocentric cameras therefore move with their respective participants. A third, dynamic exocentric camera tracks a target 0.7 m above the pelvis midpoint from a horizontal offset of 2.6 m perpendicular to the projected A-to-B pelvis axis and a height of 1.2 m above that target. Its local $-Z$ axis points toward the target, with local $+Y$ kept upright and a 45° horizontal field of view.

Both body meshes and all cameras share a common world coordinate system and are synchronized at every timestamp. We render RGB, depth, and person instance masks from the same scene using Blender Cycles at 512×512 resolution with 16 samples, GPU acceleration, and no denoising. Each image is associated with its rendering intrinsics, resolution, field of view, and full 4×4 camera-to-world transform in the Blender convention. These calibrated quantities define the projection geometry used by the model. The egocentric cameras have a 100° horizontal field of view. All visual modalities, camera trajectories, and paired motion are temporally aligned.

C.3 Dataset Modalities and Model Inputs

Visual observations. Each interaction sequence provides synchronized dual-egocentric and exocentric RGB. The dynamic exocentric stream supports reference visualization and quality assurance.

Camera geometry. Camera intrinsics and camera-to-world poses are available at every frame for all three views. The participant-mounted cameras define the shared pair reference and relate image appearance to moving-view geometry.

Paired motion. Both participants are represented by synchronized SMPL-X body and hand motion, root translation and orientation, and body shape in the shared world frame. These quantities define the prediction target and the body meshes used for rendering. Appendix D defines the motion representation used by the model.

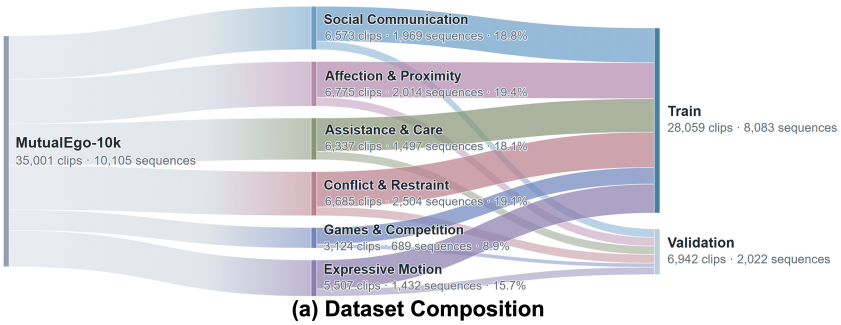
Auxiliary annotations. Auxiliary annotations comprise aligned 16-bit depth, person instance masks, and one of 43 motion-type labels. Mask values 0, 1, and 2 denote background, person A, and person B. Depth is exported from the Blender Cycles Depth/Z pass as floating-point EXR values in meters and stored as 16-bit PNG values quantized in millimeters: positive stored values are divided by 1,000 to recover meters. Background, non-finite values, and depths outside $0 < d < 65.535$ m are encoded as zero.

Task inputs. Reconstruction uses dual-egocentric RGB and calibrated egocentric camera geometry over the complete observation window. Forecasting uses these modalities only over the observed prefix. Paired motion annotations provide training targets and evaluation references. The remaining modalities are provided as auxiliary annotations and visualization resources.

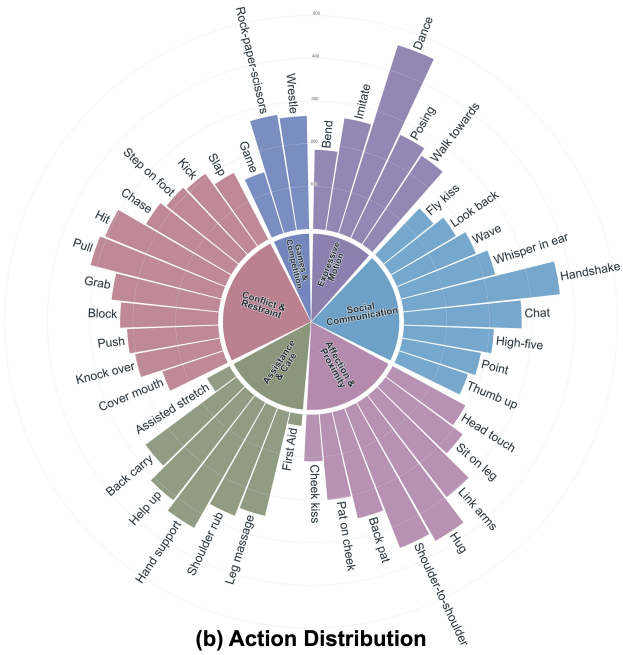
C.4 Dataset Statistics

Figure C.1 summarizes the dataset composition and action distribution, while Tables C.1 and C.2 provide the exact counts. Each source sequence is a standardized paired interaction, and each motion clip is a 40-frame window extracted from one such sequence. The 8:2 source-interaction partition produces 35,001 clips in total, comprising 28,059 training and 6,942 validation clips. Their proportions differ slightly from an exact 8:2 clip split because source durations vary and all windows from one interaction remain in the same partition.

Table C.2 gives the complete action-label distribution. First Aid contains one source sequence and five training clips, with no validation sequence.



(a) Dataset Composition



(b) Action Distribution

Figure C.1: Composition and action distribution of MutualEgo-10K. (a) The 35,001 motion clips and 10,105 source sequences are grouped into six interaction families and partitioned into training and validation sets. Percentages are shares of all motion clips. (b) Source-sequence counts for the 43 actions, grouped by the same families.

Table C.1: Sequence and clip counts for MutualEgo-10K.

Unit	Total	Train	Validation
Source sequences	10,105	8,083	2,022
Motion clips	35,001	28,059	6,942

Table C.2: Per-label sequence and clip counts. "Tr" and "Va" denote training and validation.

Label	Activity	Clips	Tr	Va	Seqs.	Tr	Va	Label	Activity	Clips	Tr	Va	Seqs.	Tr	Va
A000	Hug	1242	999	243	373	297	76	A022	Link arms	1290	1041	249	302	242	60
A001	Handshake	1075	882	193	369	295	74	A023	Shoulder to shoulder	1367	1096	271	352	282	70
A002	Wave	496	393	103	201	161	40	A024	Bend	411	331	80	186	148	38
A003	Grab	668	536	132	253	203	50	A025	Carry on back	1182	948	234	266	214	52
A004	Hit	942	760	182	305	243	62	A026	Massaging shoulder	1213	965	248	271	217	54
A005	Kick	433	347	86	209	169	40	A027	Massaging leg	896	718	178	250	200	50
A006	Posing	842	672	170	268	214	54	A028	Hand wrestling	1173	941	232	265	211	54
A007	Push	522	419	103	214	172	42	A029	Chat	2011	1620	391	276	220	56
A008	Pull	852	682	170	316	252	64	A030	Pat on cheek	499	403	96	202	162	40
A009	Sit on leg	906	731	175	225	179	46	A031	Thumb up	407	325	82	170	136	34
A010	Slap	414	331	83	176	142	34	A032	Touch head	467	372	95	194	154	40
A011	Pat on back	695	554	141	256	204	52	A033	Imitate	1088	873	215	266	214	52
A012	Point finger at	448	362	86	187	151	36	A034	Kiss on cheek	309	247	62	110	88	22
A013	Walk towards	705	572	133	250	200	50	A035	Help up	1242	996	246	311	249	62
A014	Knock over	547	437	110	202	162	40	A036	Cover mouth	359	287	72	150	120	30
A015	Step on foot	527	415	112	217	175	42	A037	Look back	519	416	103	177	141	36
A016	High-five	584	462	122	211	169	42	A038	Block	611	496	115	230	186	44
A017	Chase	810	649	161	232	184	48	A039	Fly kiss	300	240	60	150	120	30
A018	Whisper in ear	733	580	153	228	182	46	A040	Assisted Stretch	226	187	39	60	48	12
A019	Support with hand	1573	1246	327	338	270	68	A041	First Aid	5	5	0	1	1	0
A020	Rock-paper-scissors	1295	1032	263	274	218	56	A042	Game	656	550	106	150	120	30
A021	Dance	2461	1941	520	462	368	94								

D Experimental Protocol and Model Specifications

D.1 Pair Motion and Camera-Trajectory Representations

We use a Z-up world coordinate system, with the XY plane representing the ground plane. The shared pair frame is defined by camera A at the first frame: its position projected onto the ground plane sets the origin, and the horizontal projection of the local $-Z$ viewing direction of the Blender camera is yaw-aligned with pair-frame $+Y$. The same world-to-pair transform is applied to both participants and both cameras, preserving their relative spatial configuration. The transform requires both cameras to be localized in a common metric world frame with known gravity and floor geometry. MutualEgo-10K supplies rendering-based camera transforms, and Harmony4D supplies calibrated capture geometry.

Each participant is represented by the 243D motion state in Table D.1. The rotations and positions of the 22 body joints are expressed globally in the pair frame. The root trajectory is represented by an absolute transform at the first frame and frame-to-frame increments thereafter. Binary left/right foot-contact labels indicate whether the ankle velocity is below 0.02 and its height is below the estimated sequence floor plus 0.08. Concatenating the states of participant A and participant B in this order yields the 486D pair token used by diffusion. Both joint rotations and positions are directly predicted; geometric metrics use the corresponding global-joint components. The ten shape coefficients are repeated across frames and predicted within each frame-level state under the same element-wise MSE

Table D.1: Components of the 243D per-participant motion representation.

Component	Dim.	Representation
Body joints	198	22 global rotations (6D) and positions (3D)
Root motion	9	Absolute/incremental rotation (6D) and translation
Left hand	12	SMPL-X hand PCA coefficients
Right hand	12	SMPL-X hand PCA coefficients
Foot contact	2	Left/right binary contact indicators
Body shape	10	SMPL-X shape coefficients

Each egocentric camera has the 18D representation in Table D.2. It combines a 9D absolute camera-to-pair transform with a 9D frame-to-frame motion increment; at the first frame, the absolute transform also serves as the incremental component. Concatenating the two camera representations in participant order yields a 36D trajectory condition, which is projected and added to temporal motion queries. RAMT directly uses the absolute camera centers and orientations in the pair frame to form the geometry in Eq. B.1.

Table D.2: Components of the 18D per-camera trajectory representation.

Component	Dim.	Representation
Absolute pose	9	Camera-to-pair rotation (6D) and translation
Motion increment	9	Frame-to-frame rotation (6D) and translation

D.2 Denoiser Architecture and Diffusion Process

Noisy pair motion is projected from 486 to 768 dimensions, receives a 768D sinusoidal temporal encoding, and is added to the projected 36D camera trajectory. A learned embedding of the diffusion index is prepended as a timestep token. RAMT outputs are flattened in $[\text{time}, \text{view}, \text{role}]$ order and preceded by the same timestep embedding, giving 401 context tokens: one timestep token plus $40 \times 2 \times 5$ visual tokens. Each frame is represented by one pair-motion token.

The denoiser has 12 blocks, hidden width 768, 12 attention heads, feed-forward width 1,536, and dropout 0.1. Each block applies query-normalized temporal self-attention, query-normalized cross-attention to the complete RAMT context, and a feed-forward residual update. After the leading timestep state is discarded, a shared $768 \to 486$ linear head predicts the clean pair state for every frame.

The process uses $L = 1000$ diffusion steps. With

$$f(s) = \cos^2\left(\frac{s + 0.008}{1.008}\frac{\pi}{2}\right), \qquad \beta_\tau = \min\left(1 - \frac{f((\tau+1)/L)}{f(\tau/L)}, 0.999\right), \tag{D.1}$$

the network is trained for direct $\mathbf{X}_0$ prediction. If M_{tj} is the broadcast valid-element mask, the expanded form of Eq. 4 is

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{X}_0,\tau,\boldsymbol{\epsilon}}\left[\frac{\sum_{t,j} M_{tj}(X_{0,tj} - \widehat{X}_{0,tj})^2}{\sum_{t,j} M_{tj}}\right]. \tag{D.2}$$

Diffusion steps are sampled uniformly with unit importance weights. The model is optimized with the diffusion objective in Eq. D.2. Inference uses the full 1,000-step ancestral DDPM sampler with fixed-small posterior variance.

D.3 Training Details

Our models on MutualEgo-10K follow the training and inference settings in Table D.3, with temporal stride two for visual features. Reconstruction and forecasting are trained independently for 350 epochs. The DINOv3 ViT-B/16 visual encoder remains frozen throughout training. MutualEgo and its RAMT component ablations use dense patch features, while the single-view, independent, and joint-baseline variants use global features. All trainable modules other than the visual encoder are optimized from scratch, and the epoch-350 checkpoints are used for evaluation.

D.4 Forecasting Protocol and Future-Condition Masking

Forecasting uses a 40-frame window with a 10-frame RGB and camera observation prefix and a 30-frame prediction horizon. RAMT samples visual features, camera poses, and intrinsics from the observed prefix at temporal stride two. Future visual and trajectory conditions are replaced with learned placeholder tokens.

Training applies Eq. D.2 over all 40 valid motion frames, with ground-truth motion serving as supervision. At inference, sampling starts from Gaussian noise over the full window and estimates both historical and future motion using only the prefix RGB and camera conditions. The eight primary forecasting metrics for our models are computed over all 40 frames.

D.5 Evaluation Protocol

MutualEgo-10K. Results are evaluated on the 6,942-clip validation split. Geometric errors are aggregated with equal weight per valid evaluated element. Foot-contact true positives, false positives,

Table D.3: Training and inference settings for MutualEgo-10K.

Item	Setting
Optimizer	AdamW
Learning rate / weight decay	3×10^{-5} / 0.01
Epochs	350
Learning-rate schedule	$\times 0.1$ after epoch 300
Training batch size	64
Training precision	bfloat16 mixed precision
Gradient clipping	global ℓ_2 norm 1.0
Visual encoder	Frozen DINOv3 ViT-B/16
Random seed	62
Diffusion-step sampling	Uniform
Generation sampling	1,000-step ancestral DDPM
Evaluation batch / precision	32 / float32
Hardware	$1\times$ NVIDIA RTX 6000 Ada

and false negatives are aggregated over the complete validation split before computing the F1 score for each participant. Appendix D.6 defines the participant-level averaging and pair-relative metrics.

Harmony4D. Harmony4D motions use a Z-up world frame with the XY plane as ground and the Blender camera-axis convention. Egocentric images are rectified using their Fisheye624 fisheye calibration to the pinhole representation used for MutualEgo-10K before extracting dense DINOv3 ViT-B/16 patch features. A reconstruction model is initialized from the MutualEgo checkpoint trained on MutualEgo-10K and fine-tuned on Harmony4D for 350 epochs using AdamW with batch size 64, learning rate 10^{-5}, and weight decay 0.01. Evaluation uses batch size 64 and BF16 mixed precision, with metrics aggregated over all valid evaluated elements of the 268-clip validation split.

D.6 Evaluation Metrics

Let i index clips, t frames, $p \in \{A, B\}$ participants, and $j \in \{0, \dots, J-1\}$ the $J = 22$ body joints. Predicted and target joint positions in the shared pair frame are $\widehat{\mathbf{x}}_{itpj}$ and $\mathbf{x}_{itpj}$, with rotations $\widehat{R}_{itpj}$ and R_{itpj}. The root and head indices are $r = 0$ and $h = 15$. Let $\mathcal{V}$ contain every valid evaluated (i, t) pair across the validation split. All geometric errors are aggregated over the complete set $\mathcal{V}$ before division.

MPJPE ↓. Mean Per-Joint Position Error measures the absolute 3D joint accuracy of both reconstructed participants in the shared pair frame. For each participant, we average the Euclidean position error over all J joints and all evaluated frames, and then average the two participant-level values:

$$\mathrm{MPJPE} = \frac{1}{2} \sum_{p \in \{A,B\}} \frac{1}{J|\mathcal{V}|} \sum_{(i,t) \in \mathcal{V}} \sum_{j=0}^{J-1} \left\| \widehat{\mathbf{x}}_{itpj} - \mathbf{x}_{itpj} \right\|_2 . \tag{D.3}$$

MPJPE is computed directly in the shared pair frame before any alignment. The metric is reported in meters, and lower values indicate more accurate shared-frame body reconstruction.

PA-MPJPE ↓. PA-MPJPE isolates articulated pose error by fitting an independent orientation-preserving similarity transform to each predicted person at each evaluated frame. Specifically,

$$(s^*_{itp}, Q^*_{itp}, \mathbf{a}^*_{itp}) = \underset{s>0,\, Q \in \mathrm{SO}(3),\, \mathbf{a} \in \mathbb{R}^3}{\arg\min} \sum_{j=0}^{J-1} \left\| s Q \widehat{\mathbf{x}}_{itpj} + \mathbf{a} - \mathbf{x}_{itpj} \right\|_2^2 , \tag{D.4}$$

and

$$\text{PA-MPJPE} = \frac{1}{2J|\mathcal{V}|} \sum_{p \in \{A,B\}} \sum_{(i,t) \in \mathcal{V}} \sum_{j=0}^{J-1} \left\| s^*_{itp} Q^*_{itp} \widehat{\mathbf{x}}_{itpj} + \mathbf{a}^*_{itp} - \mathbf{x}_{itpj} \right\|_2 . \tag{D.5}$$

The transform removes scale, rotation, and translation separately for each body and frame; hence PA-MPJPE measures pose quality but not shared-frame placement or inter-person geometry. It is reported in meters, and lower is better.

Relative MPJPE ↓. Relative MPJPE evaluates whether the predicted pair preserves the joint-wise spatial arrangement between participants. Define the target and predicted A–B displacement vectors for joint j as $\mathbf{d}_{itj} = \mathbf{x}_{itAj} - \mathbf{x}_{itBj}$ and $\widehat{\mathbf{d}}_{itj} = \widehat{\mathbf{x}}_{itAj} - \widehat{\mathbf{x}}_{itBj}$. The metric averages their Euclidean discrepancy:

$$\mathrm{RelMPJPE} = \frac{1}{J|\mathcal{V}|} \sum_{(i,t)\in\mathcal{V}} \sum_{j=0}^{J-1} \left\| \widehat{\mathbf{d}}_{itj} - \mathbf{d}_{itj} \right\|_2 . \tag{D.6}$$

This is a pair-level metric: a prediction can have plausible individual poses yet receive a large error if the two bodies are incorrectly arranged relative to one another. Relative MPJPE is measured in meters, and lower is better.

Relative Translation Error ↓. This metric evaluates the displacement between the two root joints. Let $\mathbf{u}_{it} = \mathbf{x}_{itAr} - \mathbf{x}_{itBr}$ and $\widehat{\mathbf{u}}_{it} = \widehat{\mathbf{x}}_{itAr} - \widehat{\mathbf{x}}_{itBr}$. We compute

$$\mathrm{RelTrans} = \frac{1}{|\mathcal{V}|} \sum_{(i,t)\in\mathcal{V}} \left\| \widehat{\mathbf{u}}_{it} - \mathbf{u}_{it} \right\|_2 . \tag{D.7}$$

Relative Translation Error directly measures whether the predicted participants occupy the correct pairwise locations, independently of any common global shift. It is reported in meters, and lower is better.

Relative Rotation Error ↓. Relative Rotation Error measures the orientation of participant A with respect to participant B. We define $Q_{it} = R_{itBr}^{\top} R_{itAr}$ and $\widehat{Q}_{it} = \widehat{R}_{itBr}^{\top} \widehat{R}_{itAr}$. Rotation discrepancy is the SO(3) geodesic angle

$$d_{\mathrm{SO}(3)}(R_1, R_2) = \frac{180}{\pi} \cos^{-1}\left(\mathrm{clip}\left(\frac{\mathrm{tr}(R_1^{\top} R_2) - 1}{2}, -1, 1 \right) \right), \tag{D.8}$$

and the reported pair-level error is

$$\mathrm{RelRot} = \frac{1}{|\mathcal{V}|} \sum_{(i,t)\in\mathcal{V}} d_{\mathrm{SO}(3)}(\widehat{Q}_{it}, Q_{it}). \tag{D.9}$$

The metric is reported in degrees. Lower values indicate that the predicted bodies preserve the correct relative facing direction.

Head Translation Error ↓. Head Translation Error measures the shared-frame position error of the head joint h for each participant. We first aggregate over all evaluated frames for each person and then average the two person-level errors:

$$\mathrm{HeadTrans} = \frac{1}{2} \sum_{p\in\{A,B\}} \frac{1}{|\mathcal{V}|} \sum_{(i,t)\in\mathcal{V}} \left\| \widehat{\mathbf{x}}_{itph} - \mathbf{x}_{itph} \right\|_2 . \tag{D.10}$$

Because the egocentric cameras follow head motion, this metric also reflects how well the model connects visual evidence and calibrated camera motion to global body placement. It is reported in meters, and lower is better.

Head Rotation Error ↓. Head Rotation Error evaluates the head orientation of each participant using the same SO(3) geodesic distance as Relative Rotation Error:

$$\mathrm{HeadRot} = \frac{1}{2} \sum_{p\in\{A,B\}} \frac{1}{|\mathcal{V}|} \sum_{(i,t)\in\mathcal{V}} d_{\mathrm{SO}(3)}(\widehat{R}_{itph}, R_{itph}). \tag{D.11}$$

The metric is reported in degrees, and lower is better. Unlike Relative Rotation Error, it evaluates each head orientation directly in the shared frame before averaging over participants.

Foot Contact F1 ↑. Each participant has two predicted contact values, which we threshold directly at 0.5. We accumulate true positives, false positives, and false negatives over both feet and every $(i,t) \in \mathcal{V}$. We compute a person-level F1 score and then average over the pair:

$$\mathrm{F1}_p = \frac{2\mathrm{TP}_p}{2\mathrm{TP}_p + \mathrm{FP}_p + \mathrm{FN}_p}, \qquad \mathrm{FootF1} = \frac{\mathrm{F1}_A + \mathrm{F1}_B}{2}. \tag{D.12}$$

If a participant has neither predicted nor target-positive contact, we define the person-level F1 as one. Foot Contact F1 ranges from 0 to 1, and higher values indicate more accurate contact-state prediction.

Table D.4: Summary of primary metrics. Arrows indicate the preferred direction. "Level" identifies whether a metric directly evaluates the coupled pair or averages participant-level quantities. Each metric is aggregated over all valid evaluated elements in the complete validation split.

Metric	Level	Quantity	Unit
MPJPE ↓	Per-person average	Shared-frame joint position	m
PA-MPJPE ↓	Per-person average	Similarity-aligned joint position	m
Rel. MPJPE ↓	Pair	A–B joint displacement layout	m
Rel. Trans. ↓	Pair	Relative root translation	m
Rel. Rot. ↓	Pair	Relative root orientation	degree
Head Trans. ↓	Per-person average	Shared-frame head position	m
Head Rot. ↓	Per-person average	Head orientation	degree
Foot F1 ↑	Per-person average	Binary foot contact	–

D.7 Adaptation of Compared Methods

Shared-frame composition. We evaluate the released pretrained EgoEgo, EgoAllo, and UniEgoMotion models together with UniEgoMotion trained on MutualEgo-10K. Each model retains its single-person formulation and estimates the two participants independently:

$$(I^A, C^A) \to \widehat{P}^A, \qquad (I^B, C^B) \to \widehat{P}^B, \tag{D.13}$$

For each estimator, we invert its canonicalization, convert its predictions to world-frame SMPL-X parameters, and transform both participants into the shared pair frame used for evaluation. The origin is the first-frame position of camera A projected onto the ground plane, and yaw aligns camera A's horizontal viewing direction with pair-frame $+Y$. This frame is independent of predicted body placement.

UniEgoMotion. The official UniEgoMotion (Patel et al., 2025) model represents one person with a 243D state in segment/head-canonical coordinates; our 486D pair state uses the shared pair frame. We invert its segment/head canonicalization and restore the sequence floor reference, following the camera-axis convention of UniEgoMotion, before applying the common conversion above. This procedure applies to both reconstruction and forecasting. For forecasting, the model receives only prefix RGB and camera conditions; the motion prefix required by its repainting procedure is reconstructed from these observations.

Independent variant. The independent variant applies the same shared-weight estimator separately to the two participants. Each branch is conditioned on its own egocentric view and camera trajectory and predicts the corresponding participant motion without cross-view or cross-participant information exchange. The two predictions are then expressed in the common pair frame and combined for pair-level evaluation.

EgoAllo. EgoAllo (Yi et al., 2025) uses a central-pupil-frame (CPF) trajectory, in which the CPF is located between the eyes and follows the SMPL head orientation. We convert each camera trajectory to this head/CPF representation using the dataset-level camera-to-head calibration and the model-dependent eye offset. EgoAllo recovers a world-frame root transform from its CPF-conditioned output; its SMPL-H body and hand parameters are converted to SMPL-X as described below. Its required preceding trajectory pose is obtained by backward extrapolation from the first observed motion.

EgoEgo. EgoEgo (Li et al., 2023) is conditioned on a head trajectory. We derive this trajectory from the calibrated camera path, shift it to the floor convention of EgoEgo, and convert 10 fps inputs to its 30 fps training rate using linear translation interpolation and rotational SLERP. Its canonicalization removes the first-frame horizontal translation, and its original inverse transform restores rotation. We recover the missing horizontal translation by aligning the predicted first-frame head with the head pose supplied to the model, restoring its world-frame placement, then subsample the motion to 10 fps before the common conversion. EgoEgo retains its original mean-shape assumption.

SMPL-H to SMPL-X conversion. For EgoEgo and EgoAllo, the compatible 21-joint body pose and MANO hand components transfer directly because their joint order agrees. SMPL-H and SMPL-X shape coefficients parameterize different linear spaces. We fit the ten SMPL-X shape coefficients to the source rest-pose joints, preserving the skeletal representation for joint-based evaluation. EgoEgo retains its method-specific mean body.

Rigid transforms of SMPL-X translation. The SMPL-X translation parameter $\mathbf{t}_{\mathrm{smplx}}$ differs from the root-joint position. If $\mathbf{J}_0$ is the rest-pose root for the converted body shape, the model represents a point as

$$\mathbf{x} = R_o(\mathbf{v}_{\mathrm{posed}} - \mathbf{J}_0) + \mathbf{J}_0 + \mathbf{t}_{\mathrm{smplx}}. \tag{D.14}$$

Applying a world rigid transform $(R, \mathbf{t})$ therefore requires

$$R'_o = RR_o, \qquad \mathbf{t}'_{\mathrm{smplx}} = R(\mathbf{J}_0 + \mathbf{t}_{\mathrm{smplx}}) + \mathbf{t} - \mathbf{J}_0, \tag{D.15}$$

where the pelvis-offset terms convert the origin-based update $\mathbf{t}'_{\mathrm{smplx}} = R\mathbf{t}_{\mathrm{smplx}} + \mathbf{t}$ to the SMPL-X parameter convention. Likewise, a world-frame root-joint position $\mathbf{r}$ corresponds to SMPL-X translation $\mathbf{r} - \mathbf{J}_0$, with $\mathbf{J}_0$ computed from the converted SMPL-X shape.

E Additional Experimental Results and Analysis

This appendix supplements the main results with input-gradient sensitivity analysis, training-data scaling, reconstruction on Harmony4D, quantitative diagnostics, and qualitative examples.

E.1 Input-Gradient Sensitivity Analysis

Figure E.1 compares input-gradient sensitivity maps for MutualEgo w/o RAMT and MutualEgo on the same egocentric images under matched diffusion conditions. Warmer colors indicate greater sensitivity of the denoising objective to the corresponding input pixels.

For each selected frame and view, we differentiate the denoising MSE against the clean pair-motion target over valid task-evaluation frames with respect to the resized RGB input, backpropagating through frozen DINOv3 only for the selected observation while treating features from all other observations as fixed. We average absolute gradients over the three RGB channels and apply 9×9 average pooling. The two models share the input, target, evaluation mask, diffusion timestep, and sampled noise. Within each example, both maps use their pooled 55th and 99th percentiles l, h to normalize saliency S as $(1 - \exp[-\max((S-l)/(h-l), 0)])^{0.85}$; normalization is computed separately for each example and shared by both methods.

In the illustrated examples, MutualEgo w/o RAMT produces weaker and more localized responses. MutualEgo responds more broadly over the head, torso, and limbs of the partner, including visible hands and forearms during close contact. This suggests greater input-gradient sensitivity to visible body and interaction cues under the matched conditions.

E.2 Training-Data Scaling

For each activity label, we retain the indicated proportion of its training clips to construct label-stratified subsets. Tables E.1 and E.2 report the reconstruction and forecasting results, respectively. Their 100% rows correspond to the MutualEgo models in Table 2.

Table E.1: Reconstruction results with increasing training data on MutualEgo-10K.

Training data	Train clips	MPJPE ↓	PA-MPJPE ↓	Rel. MPJPE ↓	Rel. Trans. ↓	Rel. Rot. ↓	Head Trans. ↓	Head Rot. ↓	Foot F1 ↑
20%	5,612	0.134	0.067	0.132	0.084	13.8	0.100	6.6	0.865
40%	11,224	0.108	0.054	0.110	0.070	12.5	0.079	5.7	0.878
60%	16,835	0.095	0.049	0.099	0.063	11.8	0.069	5.3	0.891
80%	22,447	0.088	0.045	0.092	0.060	11.2	0.064	5.0	0.898
100%	**28,059**	**0.081**	**0.043**	**0.088**	**0.058**	**10.7**	**0.058**	**4.8**	**0.905**

From 20% to 100% training data, reconstruction MPJPE falls by 39.1% and forecasting MPJPE by 28.4%. All positional errors improve monotonically in both tasks. Reconstruction angular errors

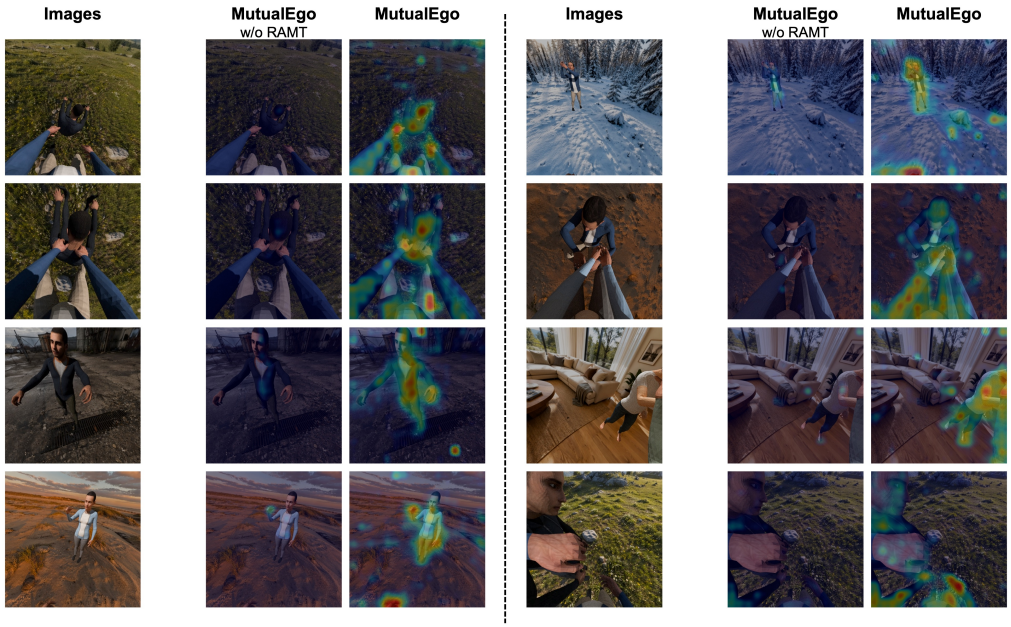


Figure E.1: Input-gradient sensitivity maps on representative MutualEgo-10K validation frames. Each triplet shows the input image, MutualEgo w/o RAMT, and MutualEgo. Warmer colors indicate greater sensitivity of the denoising objective to input pixels. Each example uses a separately normalized color scale shared by both methods. In these examples, MutualEgo shows stronger responses over visible body regions, including hands and forearms during close interactions.

Table E.2: Forecasting results with increasing training data on MutualEgo-10K.

Training data	Train clips	MPJPE ↓	PA-MPJPE ↓	Rel. MPJPE ↓	Rel. Trans. ↓	Rel. Rot. ↓	Head Trans. ↓	Head Rot. ↓	Foot F1 ↑
20%	5,612	0.401	0.094	0.444	0.400	35.6	0.401	29.2	0.820
40%	11,224	0.349	0.083	0.394	0.354	32.5	0.342	26.5	**0.840**
60%	16,835	0.341	0.082	0.385	0.342	32.4	0.337	26.9	0.833
80%	22,447	0.320	0.078	0.359	0.318	31.1	0.316	25.3	0.823
100%	**28,059**	**0.287**	**0.073**	**0.329**	**0.289**	**29.1**	**0.282**	**23.8**	0.832

also improve monotonically; forecasting head-rotation error is non-monotonic. Reconstruction Foot F1 also increases monotonically, whereas forecasting Foot F1 is non-monotonic and peaks at 40%. These results show that increasing the training data consistently improves motion estimation accuracy in both reconstruction and forecasting.

E.3 Additional Evaluation on Harmony4D

Table E.3 compares the reconstruction performance of MutualEgo and the baselines on Harmony4D.

Table E.3: Reconstruction on Harmony4D after fine-tuning the MutualEgo checkpoint trained on MutualEgo-10K; external baselines use their released weights. Distances are in meters and angles in degrees. **Bold** and underlined values mark the best and second-best reported results respectively.

Method	MPJPE ↓	PA-MPJPE ↓	Rel. MPJPE ↓	Rel. Trans. ↓	Rel. Rot. ↓	Head Trans. ↓	Head Rot. ↓	Foot F1 ↑
Reconstruction								
EgoEgo	0.3114	0.1326	**0.3510**	**0.2543**	44.71	0.2089	16.39	0.4222
EgoAllo	0.3526	0.1565	0.3787	0.2815	45.26	0.2354	17.24	0.4601
UniEgoMotion	0.3127	0.1267	0.5005	0.4659	**28.54**	**0.1958**	38.00	0.5405
MutualEgo	**0.3049**	**0.0947**	0.3614	0.2702	34.72	0.2838	**11.58**	**0.5910**

MutualEgo has the lowest MPJPE, PA-MPJPE, and head-rotation error and the highest Foot F1 among the compared methods. Pair-relative geometry and head translation remain mixed: EgoEgo has lower pair-relative positional errors, while UniEgoMotion has lower relative-rotation and head-translation errors.

E.4 Additional Quantitative Diagnostics

Table E.4 reports supplementary diagnostics on MutualEgo-10K. *Root Trans.* measures root-translation error for each participant. *Pair Motion MSE* averages squared errors over the complete 486D pair-motion state on valid evaluated frames. *Rot6D MSE* is restricted to the six rotation entries of the 22 body joints for each participant on those frames. All three reconstruction diagnostics decrease from MutualEgo w/o RAMT to MutualEgo. For forecasting, MutualEgo achieves a root-translation error of 0.261 m over the full 40-frame window.

Table E.4: Additional pair-motion diagnostics on MutualEgo-10K for reconstruction and forecasting. Root Trans. is measured in meters. Forecasting diagnostics use all 40 frames for both MutualEgo w/o RAMT and MutualEgo.

Method	Task	Root Trans. ↓	Pair Motion MSE ↓	Rot6D MSE ↓
MutualEgo w/o RAMT	Reconstruction	0.088	0.0716	0.0447
MutualEgo	Reconstruction	0.064	0.0468	0.0240
MutualEgo w/o RAMT	Forecasting	0.323	0.1501	0.1354
MutualEgo	Forecasting	0.261	0.1290	0.1186

Pair-formulation variants. Table E.5 reports all eight metrics for the four variants in Table 3: the single-view and independent variants, the joint baseline, and MutualEgo. The Independent variant applies a shared-weight estimator separately to the two participants using their respective views and camera trajectories, without cross-view or cross-participant information exchange during prediction. The single-view and independent variants use DINOv3 ViT-B/16 global features. All four variants are trained for 350 epochs, and the final checkpoints are used for evaluation.

Table E.5: Complete ablation of dual-view and joint pair modeling on MutualEgo-10K. Distances are in meters and rotations in degrees. **Bold** and underlined values denote the best and second-best results within each task, respectively.

Variant	MPJPE ↓	PA-MPJPE ↓	Rel. MPJPE ↓	Rel. Trans. ↓	Rel. Rot. ↓	Head Trans. ↓	Head Rot. ↓	Foot F1 ↑
Reconstruction								
MutualEgo w/o RAMT, *Single-View*	0.226	0.075	0.313	0.274	24.3	0.205	16.0	0.853
MutualEgo w/o RAMT, *Independent*	0.118	0.059	0.164	0.125	13.2	0.087	5.7	0.900
MutualEgo w/o RAMT	0.116	0.062	0.117	0.067	13.3	0.080	5.7	0.881
MutualEgo	**0.081**	**0.043**	**0.088**	**0.058**	**10.7**	**0.058**	**4.8**	**0.905**
Forecasting								
MutualEgo w/o RAMT, *Single-View*	0.461	0.087	0.553	0.521	44.0	0.453	34.9	0.812
MutualEgo w/o RAMT, *Independent*	0.345	0.085	0.425	0.387	34.1	0.338	26.5	0.821
MutualEgo w/o RAMT	0.347	0.079	0.375	0.336	32.0	0.342	25.8	0.818
MutualEgo	**0.287**	**0.073**	**0.329**	**0.289**	**29.1**	**0.282**	**23.8**	**0.832**

E.5 Robustness to Camera and Synchronization Errors

We evaluate reconstruction on the MutualEgo-10K validation set under controlled perturbations to camera calibration and temporal synchronization. Individual settings apply 3° camera rotation, 3 cm camera translation, 5% focal-length and 2% principal-point perturbations, or a one-frame temporal offset. Combined applies all four perturbation types simultaneously, while Clean uses the original calibrated and synchronized inputs.

Table E.6: Robustness of MutualEgo to camera calibration and temporal synchronization perturbations on MutualEgo-10K reconstruction. Distances are in meters.

Perturbation	MPJPE ↓	Rel. MPJPE ↓
Clean	0.081	0.088
Rotation 3°	0.083	0.090
Translation 3 cm	0.085	0.095
Intrinsics 5% focal / 2% pp	0.082	0.088
Temporal 1 frame	0.088	0.099
Combined	0.093	0.107

Temporal misalignment produces the largest degradation among individual perturbations, increasing MPJPE from 0.081 to 0.088 m and relative MPJPE from 0.088 to 0.099 m. Translation increases MPJPE by 4.2% and relative MPJPE by 8.2%, indicating a larger effect on pair-relative geometry than on overall body accuracy. Intrinsic perturbation produces comparatively small changes, with MPJPE of 0.082 m and relative MPJPE of 0.088 m. Combined produces the largest overall degradation, reaching 0.093 m MPJPE and 0.107 m relative MPJPE.

E.6 Additional Qualitative Results

Figures E.2–E.5 extend the reconstruction comparison with four additional sequences. Each figure shows the synchronized dual-egocentric views, ground-truth paired motion, and reconstructions from EgoEgo, EgoAllo, UniEgoMotion, and MutualEgo, together with the corresponding error maps. In the error maps, darker red indicates larger position errors relative to the ground truth. Figures E.6–E.9 show four additional forecasting sequences, comparing UniEgoMotion and MutualEgo with the ground truth. Figure E.10 presents reconstruction examples on Harmony4D, showing the synchronized dual-egocentric images, ground-truth paired motion, and reconstructions from MutualEgo.

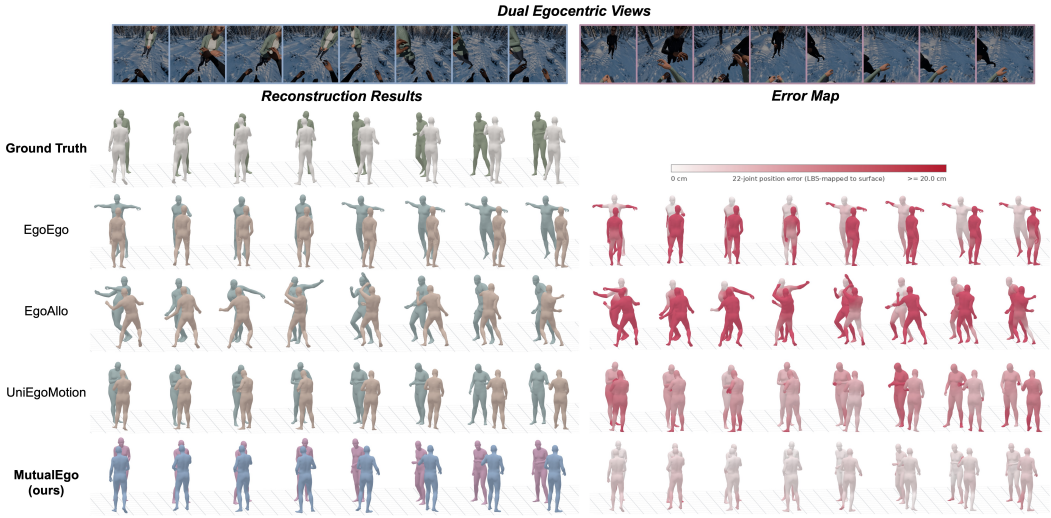


Figure E.2: Additional reconstruction comparison on MutualEgo-10K (1/4). **Top:** synchronized dual-egocentric views. **Left:** ground truth and paired motions reconstructed by EgoEgo, EgoAllo, UniEgoMotion, and MutualEgo. **Right:** 22-joint position errors mapped to the mesh surface using linear blend skinning (LBS) weights. Darker red indicates larger errors; the scale saturates at 20 cm. Time progresses from left to right within each view and result panel.

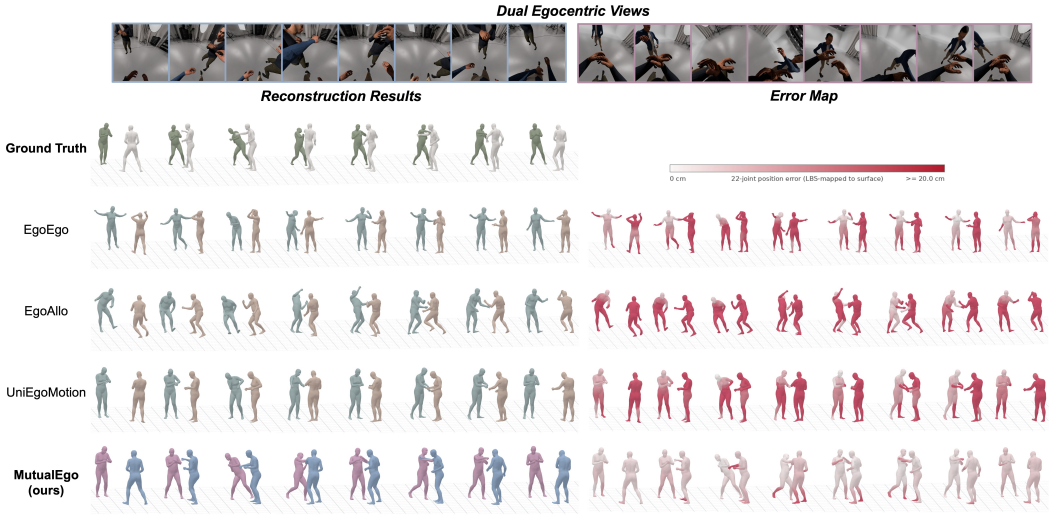


Figure E.3: Additional reconstruction comparison on MutualEgo-10K (2/4). **Top:** synchronized dual-egocentric views. **Left:** ground truth and paired motions reconstructed by EgoEgo, EgoAllo, UniEgoMotion, and MutualEgo. **Right:** 22-joint position errors mapped to the mesh surface using linear blend skinning (LBS) weights. Darker red indicates larger errors; the scale saturates at 20 cm. Time progresses from left to right within each view and result panel.

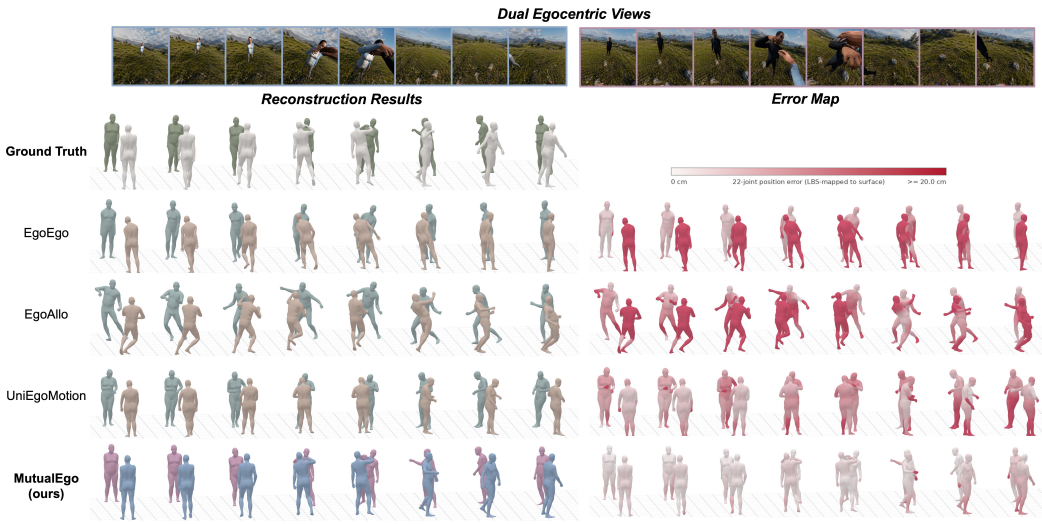


Figure E.4: Additional reconstruction comparison on MutualEgo-10K (3/4). **Top:** synchronized dual-egocentric views. **Left:** ground truth and paired motions reconstructed by EgoEgo, EgoAllo, UniEgoMotion, and MutualEgo. **Right:** 22-joint position errors mapped to the mesh surface using linear blend skinning (LBS) weights. Darker red indicates larger errors; the scale saturates at 20 cm. Time progresses from left to right within each view and result panel.

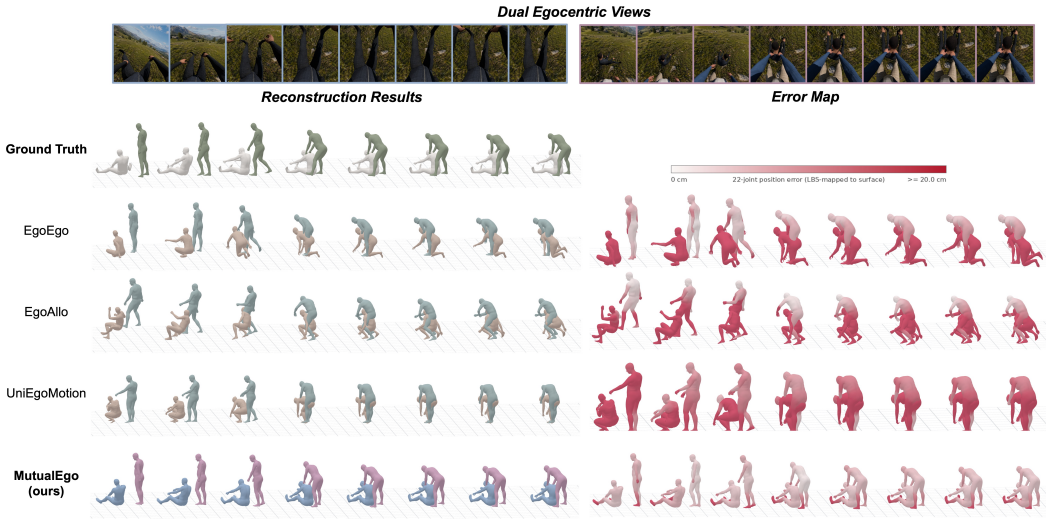


Figure E.5: Additional reconstruction comparison on MutualEgo-10K (4/4). **Top:** synchronized dual-egocentric views. **Left:** ground truth and paired motions reconstructed by EgoEgo, EgoAllo, UniEgoMotion, and MutualEgo. **Right:** 22-joint position errors mapped to the mesh surface using linear blend skinning (LBS) weights. Darker red indicates larger errors; the scale saturates at 20 cm. Time progresses from left to right within each view and result panel.

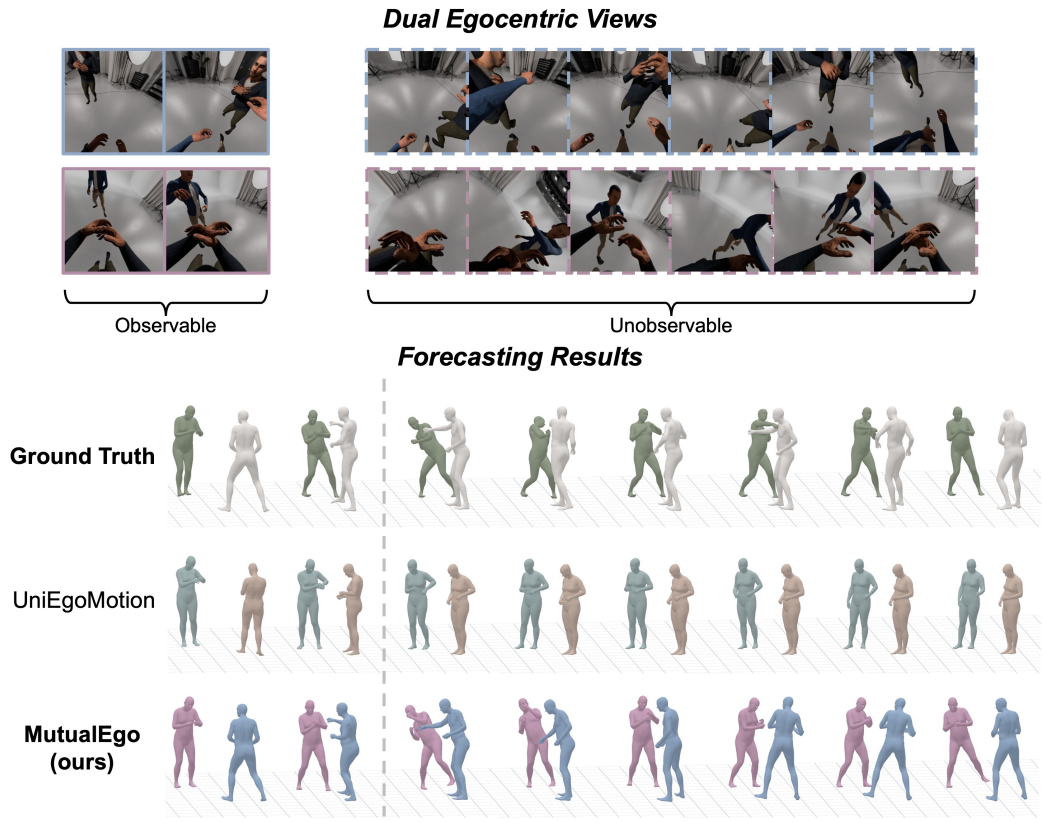


Figure E.6: Additional forecasting comparison on MutualEgo-10K (1/4). **Top:** observed dual-egocentric RGB on the left and future RGB shown for reference on the right. **Bottom:** ground-truth paired motion and estimates from UniEgoMotion and MutualEgo. The dashed line separates estimated history (left) from predicted future (right). Time progresses from left to right.

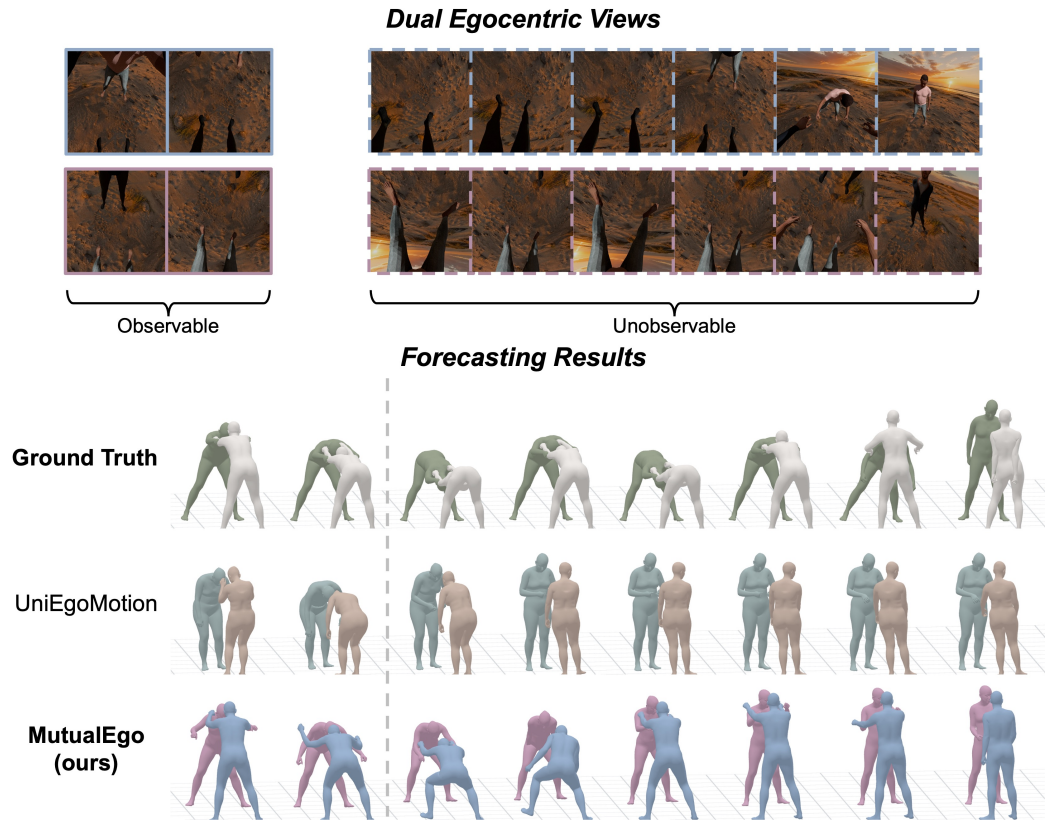


Figure E.7: Additional forecasting comparison on MutualEgo-10K (2/4). **Top:** observed dual-egocentric RGB on the left and future RGB shown for reference on the right. **Bottom:** ground-truth paired motion and estimates from UniEgoMotion and MutualEgo. The dashed line separates estimated history (left) from predicted future (right). Time progresses from left to right.

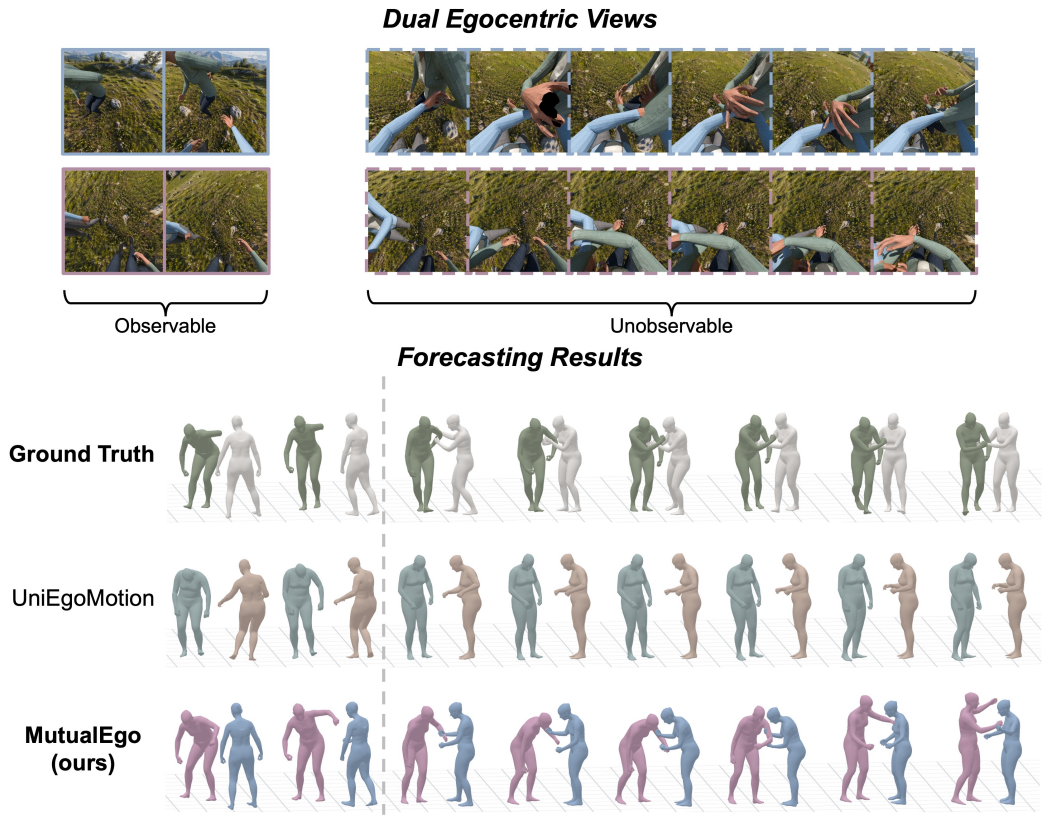


Figure E.8: Additional forecasting comparison on MutualEgo-10K (3/4). **Top:** observed dual-egocentric RGB on the left and future RGB shown for reference on the right. **Bottom:** ground-truth paired motion and estimates from UniEgoMotion and MutualEgo. The dashed line separates estimated history (left) from predicted future (right). Time progresses from left to right.

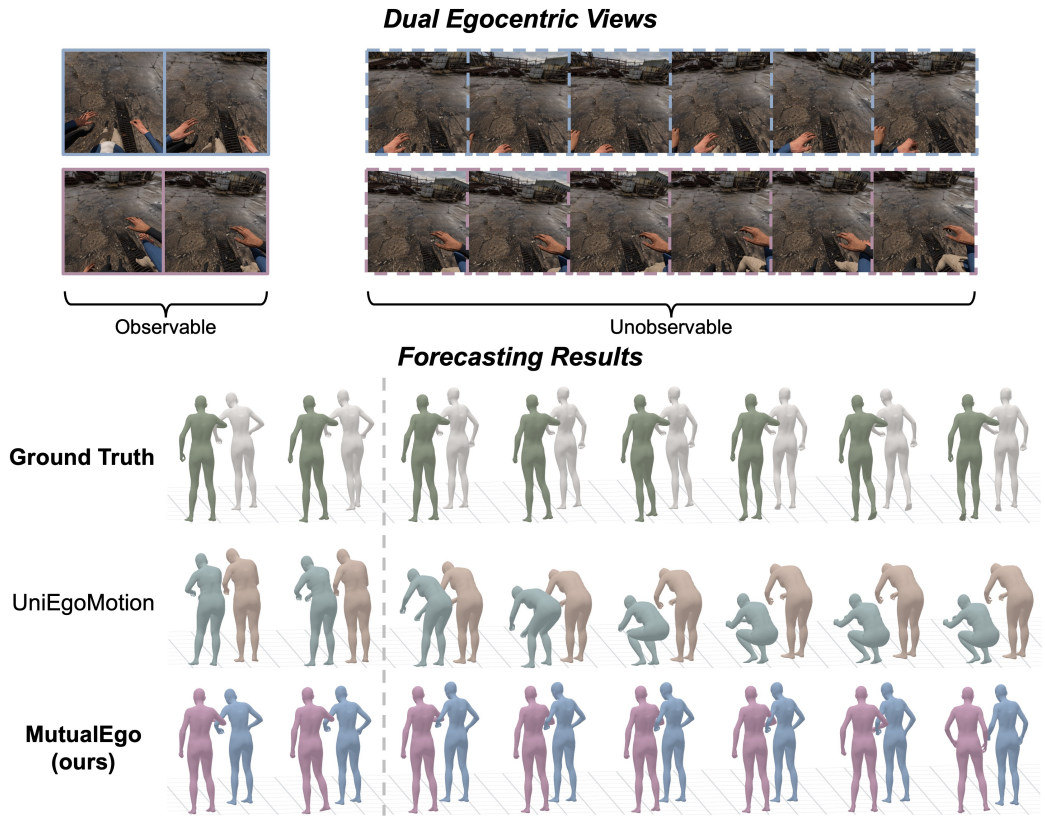


Figure E.9: Additional forecasting comparison on MutualEgo-10K (4/4). **Top:** observed dual-egocentric RGB on the left and future RGB shown for reference on the right. **Bottom:** ground-truth paired motion and estimates from UniEgoMotion and MutualEgo. The dashed line separates estimated history (left) from predicted future (right). Time progresses from left to right.

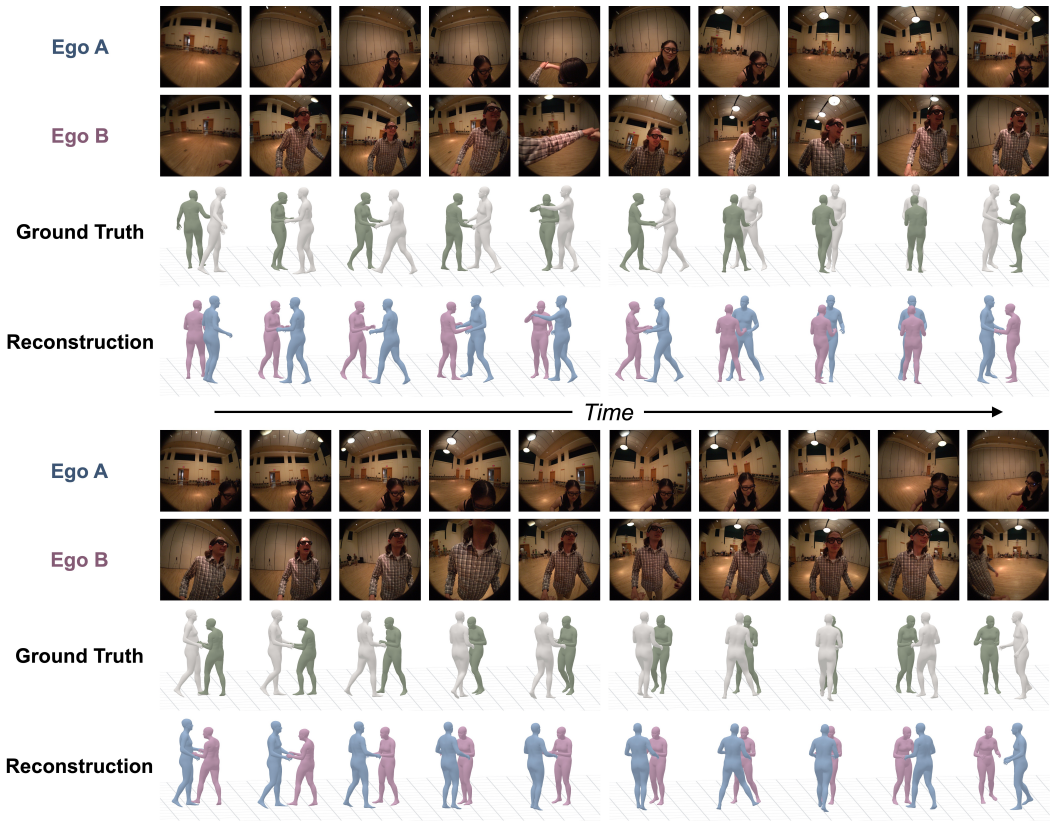


Figure E.10: Qualitative reconstruction results on Harmony4D. Each block shows the synchronized egocentric views from participants A and B, the ground-truth paired motion, and the reconstruction produced by MutualEgo after fine-tuning on Harmony4D. Time progresses from left to right within each block.

F LIMITATIONS

MutualEgo-10K uses rendered observations, creating a domain gap to real wearable capture. Its upper-body camera model approximates head motion and can miss independent head turns. MutualEgo requires synchronized cameras localized in a shared metric frame with calibrated intrinsics, gravity, and floor geometry. The perturbation study in Sec. E.5 covers bounded calibration and synchronization errors. Inference uses 1,000 ancestral DDPM steps, making sampling computationally expensive. Evaluation uses a source-interaction validation split and measures body motion, pair-relative geometry, and foot–ground contact. Generalization to unseen participants and sessions, human–human contact quality, and multi-sample forecast diversity remain to be evaluated.

G FUTURE WORK

Future work will focus on robust cross-camera alignment and synchronization, realistic head-camera motion, and interaction-specific supervision for contact, penetration, and contact timing. Faster sampling and additional real wearable data can support deployment and broader evaluation of generalization and forecast diversity.