EgoCast: Forecasting Egocentric Human Pose in the Wild

Maria Escobar
Universidad de Los Andes
mc.escobar11@uniandes.edu.co

Juanita Puentes
Universidad de Los Andes
j.puentes@uniandes.edu.co

Cristhian Forigua
Universidad de Los Andes
cd.forigua@uniandes.edu.co

Jordi Pont-Tuset
Google DeepMind
jponttuset@google.com

Kevis-Kokitsi Maninis
Google DeepMind
kmaninis@google.com

Pablo Arbelaez
Universidad de Los Andes
pa.arbelaez@uniandes.edu.com

Abstract

Accurately estimating and forecasting human body pose is important for enhancing the user's sense of immersion in Augmented Reality. Addressing this need, our paper introduces EgoCast, a bimodal method for 3D human pose forecasting using egocentric videos and proprioceptive data. We study the task of human pose forecasting in a realistic setting, extending the boundaries of temporal forecasting in dynamic scenes and building on the current framework for current pose estimation in the wild. We introduce a current-frame estimation module that generates pseudo-groundtruth poses for inference, eliminating the need for past groundtruth poses typically required by current methods during forecasting. Our experimental results on the recent Ego-Exo4D and Aria Digital Twin datasets validate EgoCast for real-life motion estimation. On the Ego-Exo4D Body Pose 2024 Challenge, our method significantly outperforms the state-of-the-art approaches, laying the groundwork for future research in human pose estimation and forecasting in unscripted activities with egocentric inputs.

1. Introduction

As Augmented Reality (AR) continues to revolutionize our interactions within digital environments, accurately capturing the human body's motion becomes essential for delivering a smooth fusion of the real and virtual worlds [4, 32]. Beyond capturing the present pose, forecasting human motion is critical for anticipating user actions and needs. Forecasting will be necessary in applications such as patient rehabilitation and medical monitoring [23, 28, 35], augmented reality responsiveness, and sports coaching [48]. However, predicting future motion in real-world settings is challenging because people do not move in simple, predictable patterns. Unlike controlled benchmarks, where movements are designed to reach a specific goal through a limited range of actions, real-life movements are far more varied and complex. Therefore, forecasting methods must be capable of long-term prediction and generalization to different types of actions in the wild. Nonetheless, forecasting approaches [10, 16, 27] often focus on predefined movements, missing the unpredictable nature of everyday activities.

Besides sight, humans rely on *proprioception* to perceive and manage their body's position and movement. This internal sensory feedback system is crucial for everyday activities ranging from walking to complex physical motions in fields like gymnastics (*e.g.*a gymnast executing a triple jump significantly relies on proprioception to maintain balance and orientation in mid-air). This proprioceptive awareness is often enhanced by *visual cues*, providing a complete understanding of the surrounding environment and enabling more precise and controlled movements [3, 26]. This combination of internal sensory feedback and external visual information fits nicely in an egocentric setup that captures a visual stream and head pose as a form of mimicking proprioception.

In this paper, we present EgoCast, a novel method for human pose forecasting in realistic settings, that leverages both internal (proprioceptive) and external (visual) egocentric cues. EgoCast uses as input the 3D proprioceptive information from the head pose along with the visual stream from past observations (Fig. 1 left) to forecast future 3D human motion (Fig. 1 right). Our method uses a bimodal Transformer approach that mirrors human perception by mixing proprioceptive and visual information. To avoid relying on ground-truth body poses at test time for forecasting, we first design a current-frame estimation module to predict pseudo-groundtruth full-body poses, given the headset pose and the visual feed from the past. Second, we take the pseudo-groundtruth full-body poses for the past and estimate the future poses using a proprioception encoder that combines the temporal information across frames.

Current forecasting approaches [43, 52] evaluate performance in a short period, usually from 1-5 seconds at 2 FPS,

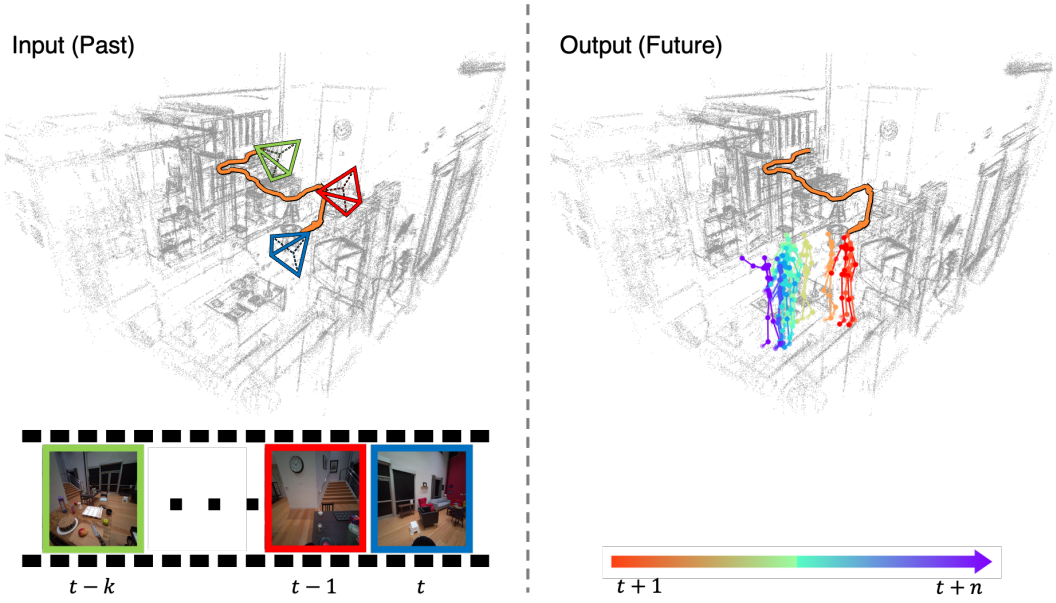

Figure 1. **3D Human Pose Forecasting.** Our forecasting approach focuses on studying human movement from egocentric inputs in a realistic setting. Given the headset trajectory (3D position and rotation) of the past $(t-k, t)$, represented as the orange line in our figure, and the visual cues gathered during the past trajectory, the goal is to forecast the 3D full-body human pose in a future temporal window $(t-t_n)$, as shown in the right side of the figure. Note that we do not receive as input ground-truth historical poses.

which is not a realistic framework for an AR setting where the user can be constantly moving. In contrast, we propose to assess the forecasting performance in a future span of $\{0.5,1,2,3,4,5\}$ seconds at 30 FPS. Using this setup, we calculate the Mean Per Joint Position Error (MPJPE) at each timeframe and then propose to create an MPJPE curve and use the area under as a new metric to measure the average performance of the methods across all timeframes.

Our results on the egocentric datasets Ego-Exo4D [13] and Aria Digital Twin (ADT) [30] show that EgoCast is suited for realistic 3D human pose forecasting. Our Current-Frame Estimation Module achieves state-of-the-art performance in the BodyPose challenge of Ego-Exo4D [13], outperforming the baseline approach by 4.15 cm and surpassing the previous state-of-the-art by approximately 1 cm, as shown by the results available on the official leaderboard. Furthermore, our forecasting approach achieves AUC values of 24.41 cm on Ego-Exo4D and 26.69 cm on ADT, showing the benefits of exploiting proprioception alongside visual cues.

Our main contributions can be summarized as follows:

- We formulate a 3D human pose forecasting benchmark for the study of 'in the wild' poses with realistic estimation timeframes and a new evaluation metric. Unlike previous approaches, our forecasting formulation does not rely on ground truth body poses at inference. Instead, we create pseudo-groundtruth poses through our current frame estimation module.
- We present EgoCast, a temporally aware transformer-based system that integrates sparse proprioceptive inputs and egocentric visual data for 3D human pose estimation and forecasting.

Find our full project on bcv-uniandes.github.io/egocast-wp/.

2. Related Work

Egocentric Vision: AR and robotics applications demand a comprehension of the world from a first-person perspective; thus, egocentric vision has gained significant attention in recent years. Most relevant datasets which collected daily-life activities to study human-object interactions include EPIC-Kitchens [6–8], UT Ego [20, 36], EGTEA Gaze+ [22] and Ego4D [12]. However, none of these datasets provides 3D human pose annotations, particularly of the camera wearer. Additional works provide 3D body annotations and visual data; however, their reliance on chest-mounted devices for data collection restricts their practical application in real-world scenarios [17, 29]. EgoBody [50] introduces a dataset designed for estimating human pose, shape, and motion from first-person perspectives, specifically focusing on social interactions. Nonetheless, the human pose in this dataset is dependent on a specific body model is not suitable for a more generic approach. Furthermore, the Aria Digital Twin (ADT) dataset [30, 34] provides densely annotated 3D human poses and visual data from real-world activities in two realistic environments—an apartment and an office. ADT [30] is

specifically designed to focus on realistic long-term activities, serving as a foundational database for developing systems for real-world applications. Recently, Ego-Exo4D [13] further enriches the landscape of human pose and activity datasets by offering a broad spectrum of human actions captured *in the wild* across diverse environments. We leverage Ego-Exo4D [13] and ADT [30] to examine forecasting in realistic settings.

Human Pose Prediction from Sparse Inputs: Estimating full-body poses from sparse inputs using head- and hand-tracking devices has become an area of considerable interest within the community. Prior works rely on Inertial Measurements Units (IMUs) to estimate the whole body pose [15, 46, 47]. Huang *et al.* trained a bidirectional LSTM to predict the human body from 6 IMU inputs. Moreover, Jiang *et al.* proposed AvatarPoser [19], a transformer-based architecture that integrates traditional Inverse Kinematics (IK) to estimate 3D full-body pose from head-mounted devices and hand controllers. Some recent work [4,9] leveraged the potential of generative diffusion models for full-body pose estimation. Du *et al.* introduced AGRoL [9], an MLP-based diffusion model that tracks entire bodies given sparse upper-body tracking signals. Recently, Gong *et al.* introduced DiffPose [11], a framework that conceptualizes 3D pose estimation through a reverse diffusion process. Lately, Zhang *et al.* proposed DynaIP, a human pose estimation method using sparse inertial sensors and part-based motion dynamics [51]. However, all these models ignore the external awareness that egocentric visual cues offer.

Human Pose Estimation from Visual Inputs: An alternative line of work is to use egocentric visual cues to estimate 3D full-body human pose. Many existing works utilize downward-facing cameras to capture device users, with several methods employing monocular approaches [1, 18, 24, 38–40]. Yuan and Kitani [48] proposed a reinforcement-learning method for pose estimation that receives egocentric video as input. Kinpoly [25] utilizes egocentric video evidence and simulation state to generate a target pose. STC-Former [37] models spatio-temporal correlation for 3D human pose estimation. Moreover, Wang *et. al* [40] introduced a convolutional variational autoencoder-based architecture with a reprojection energy term and a global pose optimizer for pose estimation using a head-mounted fisheye camera. You2Me [29] predicts the camera wearer's 3D body pose from egocentric video sequences through an LSTM model that leverages the wearer's pose from the preceding frame to predict the pose for the current frame. EgoEgo [21] addresses 3D pose estimation in a two-step manner. First, it extracts head motion as an intermediate representation and then uses it for full-body pose prediction. More recently, Wang *et. al* [41] proposed a scene-aware pose estimation network. Similarly, xr-EgoPose [38] proposed a convolutional two-step architecture for estimating human pose from synthetic fisheye egocentric videos. Recently, EgoHMR [49] estimates 3D human poses from egocentric views by accounting for body truncation, using a diffusion model informed by the surrounding 3D scene. Nevertheless, these methods are entirely based on the external awareness of egocentric images and ignore the critical internal proprioception in human motion.

Human Pose Forecasting: Some works have focused on trajectory forecasting from egocentric inputs [12, 31, 33], but this line of research only predicts future positions instead of complete human poses. EgoPose [48] estimates and forecasts 3D human poses from egocentric videos via reinforcement learning. HoloAssist [43] is an egocentric human interaction dataset primarily focused on 3D hand forecasting, omitting other body movements. Its predictive model forecasts pose for 1.5 seconds at 2 FPS from a 3-second input, indicating a need for extended prediction capabilities. Moreover, TEMPO [5] is a framework designed for pose estimation, tracking, and forecasting, leveraging spatiotemporal representations. However, this method is limited to third-person views, and forecasts pose only up to 0.33 seconds ahead, or three frames, without considering egocentric perspectives on human motion. Similarly, MotionMixer [2] and SiMLPe [14] predict 3D body pose using a multi-layer perceptron model based on previous 3D skeletons, an approach impractical for real-world applications. MotionMixer offers predictions up to 4 seconds short-term and 10 seconds long-term at 2.5 FPS, while SiMLPe extends forecasts up to 1 second at 25 FPS. EqMotion [44] introduces a graph-based, efficient equivariant motion prediction model that ensures motion equivariance, capable of predicting pedestrian trajectories up to 4.8 seconds at a rate of 4 FPS. Yan et al. proposed C^3HOST, a method for forecasting 3D whole-body human poses with a focus on grasping objects [45]. Moreover, GIMO [52] integrates eye gaze coordinates, 3D body poses, and contextual scene data through a cross-modal transformer using previous motion patterns. Nonetheless, GIMO's predictive capabilities are limited to forecasting future motion for 5 seconds, using an input sequence of 3 seconds, at a rate of 2 FPS. In contrast, our model, EgoCast, addresses these critical limitations by forecasting human motion over longer timeframes and integrating visual and proprioceptive inputs into the forecasting model.

3. Human Pose Forecasting Benchmark

Figure 2 shows an overview of our task formulation: predicting a set of 3D human poses in the future given visual and proprioceptive data from the past. We use the term *proprioceptive* to refer to the head pose. EgoCast focuses on a realistic forecasting setting since it does not assume that the models will have access to ground-truth poses from previous frames, and we evaluate on longer than usual timeframes.

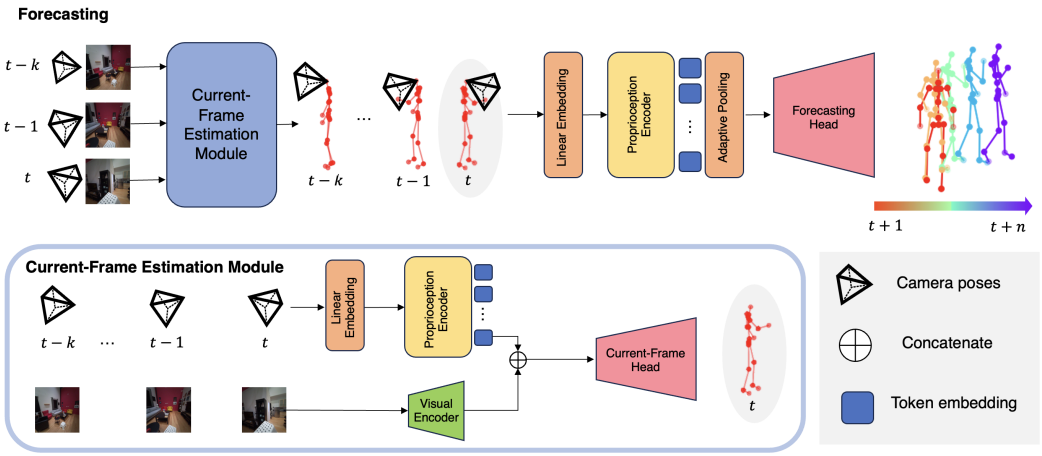

Figure 2. **EgoCast Overview.** Our method leverages proprioception and visual streams to estimate 3D human pose. (*Top*) For forecasting, we input previous camera poses and 3D full-body pose predictions through a forecasting head to estimate future 3D poses from $t+1$ to $t+n$. (*Bottom*) Since ground-truth 3D full-body poses are not available in real-case scenarios, we implement a current-frame estimation module that integrates camera poses and visual cues to estimate 3D pose at time t.

3.1. Task formulation

At a given point in time t, the input is an RGB video sequence $\mathcal{V}_{RGB} = \{v^{t-k}, \ldots, v^{t}\}$ and a proprioceptive sequence $\mathcal{P}_{headset} = \{p^{t-k}, \ldots, p^{t}\}$, $\mathcal{Y}_{headset} = \{y^{t-k}, \ldots, y^{t}\}$, where k defines the length of the time window in the past, $p^i \in \mathbb{R}^{1\times3}$ denotes the headset position and $y^i \in \phi^{1\times4}$ denotes the headset rotation in quaternion format. The objective is to predict the sequence of future human poses $\mathcal{Q}_{future} = q^{t+1}, \ldots, q^{n}$ where $q^i \in \mathbb{R}^{joints\times3}$ is the 3D position of the joints (17 joints in Ego-Exo4D [13], 21 joints in ADT [30]) for n future frames.

3.2. Evaluation metrics

We use the Mean Per Joint Position Error (MPJPE) as our primary metric following the standard in pose estimation [42]. The MPJPE measures the average euclidean distance between the ground truth and the predicted 3D joint positions without pre-alignment or root joint trajectory subtraction. We report the MPJPE in cm. We evaluate models for future predictions at {0.5, 1, 2, 3, 4, and 5} seconds. Moreover, since the MPJPE evaluates the performance for a fixed future time, we create a curve of the forecasting seconds vs MPJPE and compute the area under the curve (AUC) to measure the overall performance of each method. We perform a *minmax* normalization for AUC computation across time horizons. Since each curve represents errors, the smaller the AUC, the better the model's performance.

3.3. Datasets

We use the Ego-Exo4D dataset [13] as the base of our framework. Ego-Exo4D captures skilled human activities through both egocentric and exocentric perspectives. We focus on the egocentric data, which provides insights into close-range hand-object interactions and the wearer's focal points. Egocentric videos are recorded with Aria devices at 1404×1404 resolution and 30 FPS. The activities in Ego-Exo4D [13] include physical tasks (Soccer, Basketball, Dance and Music) or procedural tasks (Cooking, Bike Repair, and Health) across various locations. For human pose annotations, Ego-Exo4D [13] provides a collection of 17 3D joint positions representing the camera wearer's body for each time step. Please refer to Table 1 in the Supplementary Material for additional statistics of Ego-Exo4D [13].

Moreover, we utilize the ADT Dataset [30] for comparative analysis. This dataset comprises densely annotated sequences featuring ground-truth 3D body poses by Aria devices [34], including egocentric video recorded at 30 frames per second, along with the position and rotation of the camera at each frame. Furthermore, ADT [30] includes diverse activities (party, work, decorate, having a meal) showcasing a wide variety of human motions. Although all activities in ADT take place in the same space, the trajectories vary significantly. We selected ADT for further analysis due to its intentional design for capturing realistic, long-term human activities, which makes it a valuable resource for our research. Please refer to Table 3 in the Supplementary Material for additional statistics of ADT [30].

4. EgoCast

We propose a transformer-based approach for estimating 3D full-body human poses by leveraging two streams of egocentric information: (i) a proprioceptive stream, which includes the headset past positions, and (ii) a visual stream, which includes the RGB egocentric video. Both streams are

Method	Basketball	Soccer	Bike repair	Cooking	Health	Dance	Music	Avg.
EgoEgo [13, 21]	21.36	23.08	30.18	23.71	32.57	20.93	33.81	26,38
Kinpoly [13, 25]	24.98	19.09	25.19	20.80	39.23	18.03	30.30	24.36
Location-based [4, 13, 19]	19.89	16.62	20.61	12.65	11.63	21.15	15.00	18.51
EgoCast (Visual-only)	16.45	17.10	13.43	11.29	11.77	17.34	12.52	15.12
EgoCast (Full)	**16.31**	**14.35**	**13.42**	**10.24**	**10.61**	**16.58**	**10.58**	**14.36**

Table 1. **Ego-Exo4D Current Frame Pose Estimation comparison.** We present the MPJPE for the test split of Ego-Exo4D [13] in centimeters. EgoCast significantly outperforms the state-of-the-art body pose methods in both individual scenarios and overall performance in [13], as shown by the results available on the official leaderboard.

crucial for understanding human motion.

Figure 2 shows a schematic overview of our methodology. Given a sequence of camera poses and RGB images from the headset, we first estimate the 3D full-body pose at each timestamp via the current-frame estimation module. Then, the pose forecasting module uses these predicted body poses, together with the proprioceptive inputs, to estimate the 3D human poses in the following n frames in the future. Overall, Egocast proposes a whole pipeline for 3D human pose estimation in real-world scenarios where only the input streams given by the headset are available.

4.1. Current-Frame Estimation Module

Given the video sequence $\mathcal{V}_{RGB}$ and proprioceptive sequence $\mathcal{P}_{headset}$, this module estimates the 3D skeleton, q^t, in the current timestamp. We take as input tokens the three translation coordinates of the headset from times $t-k$ to t. Then, we compute deep embeddings using a linear layer and a transformer-based encoder. We keep the Transformer's output for the last token, corresponding to the encoded proprioceptive information e^t. In addition, we also include visual cues for current pose estimation. We use a visual encoder to extract a feature representation e^v from $\mathcal{V}_{RGB}$. Then, compute the final output q^t as

$$q^t = \mathcal{H}_c\left(e^t \oplus e^v\right) \tag{1}$$

where $\mathcal{H}_c$ is the current-pose estimation head composed of a 2-layer Multi-Layer Perceptron (MLP).

At inference time, if fewer frames than the required k are available, such as at the beginning of the sequence, the module adapts by using only the frames that are present. For example, for the first frame $t=0$, the estimation relies solely on the available information from that frame. For the second frame $t=1$, the input includes data from frames $t=0$ and $t=1$. This ensures that the method remains functional and accurate, even when complete temporal context is unavailable in the early stages of a video sequence.

4.2. Pose Forecasting Module

Once we have the predicted poses from the current-frame estimation module, we use them as input for the forecasting module. Given past sequences $\mathcal{P}_{headset}$, $\mathcal{Y}_{headset}$ and $\mathcal{Q}_{past}$, we compute input tokens

$$\mathcal{T}^j = (q^j \oplus p^j \oplus y^j), t-k \leq j \leq t \tag{2}$$

where $\mathcal{T}^j \in \mathbb{R}^{1\times m}$ and $m = (joints \times 3)+3+4$ represents the concatenation of (i) $(joints \times 3)$ for the predicted poses from the current-frame estimation module, (ii) 3 from the headset's translation and (iii) 4 from the headset's rotation in quaternion format. Each timeframe's data is treated as a distinct token within the transformer-based encoder, facilitating the extraction of long-range dependencies inherent in the sequence. We then extract deep forecasting embeddings e^f by

$$e^f = \mathcal{E}(l(\mathcal{T})) \tag{3}$$

where $e^f \in \mathbb{R}^{k\times d}$, being d the latent dimension of $\mathcal{E}$, l is a projection layer, and $\mathcal{E}$ is the proprioception encoder that implements a self-attention mechanism to codify each input token and its relations across time and location. Our proprioception encoder includes an adaptive pooling layer to learn a fused representation of all the input tokens. Finally, we pass the fused representation through a 2-layer MLP forecasting head $\mathcal{H}_f$ to compute the final output $\mathcal{Q}_{future}$ as well as a prediction for the translation $\mathcal{P}_{future}$ and rotation of the headset $\mathcal{Y}_{future}$.

For training, we compute the L_1 distance between the ground truth and predicted 3D full-body poses, headset translations, and rotations in the future as:

$$L_f = \lambda_{\mathcal{Q}} L_1\left(\mathcal{Q}, \hat{\mathcal{Q}}\right) + \lambda_R L_1\left(\mathcal{Y}, \hat{\mathcal{Y}}\right) + \lambda_T L_1\left(\mathcal{P}, \hat{\mathcal{P}}\right) \tag{4}$$

where $\mathcal{Q}, \mathcal{R}$, and $\mathcal{P}$ correspond to the ground-truth sequences and $\hat{\mathcal{Q}}, \hat{\mathcal{Y}}$ and $\hat{\mathcal{P}}$ are the predictions for 3D full-body poses and rotation and translation of the headset.

5. Experiments

We perform extensive experimentation to assess the performance of our proposed method. In Section 5.1, we detail the performance outcomes of the Current-Frame Estimation Module in estimating the 3D skeleton. Subsequently, Section 5.2 presents our results on forecasting the future sequence of human poses, using the predicted pseudo-groundtruth poses created in the Current-Frame Estimation Module.

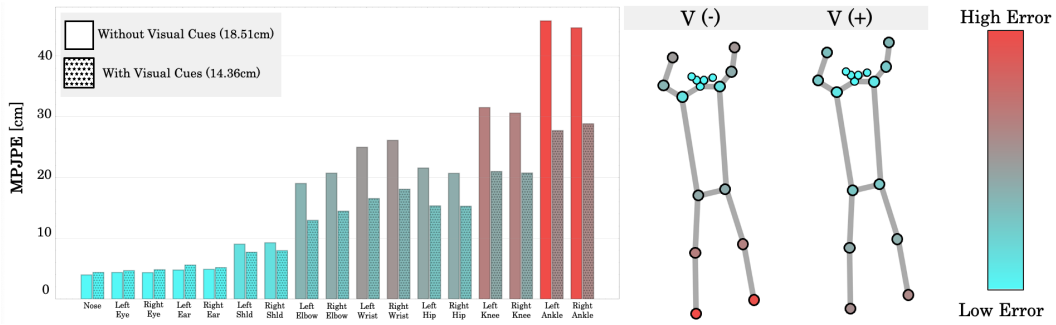

Figure 3. **Effect of visual cues on MPJPE for the current-frame estimation module.** For each joint, we present the Mean Per-Joint Position Error (MPJPE) variation, contrasting conditions without visual cues against those with visual cues, through a color scale from blue (low error) to red (high error). Visual egocentric data significantly reduces errors, especially in the lower body.

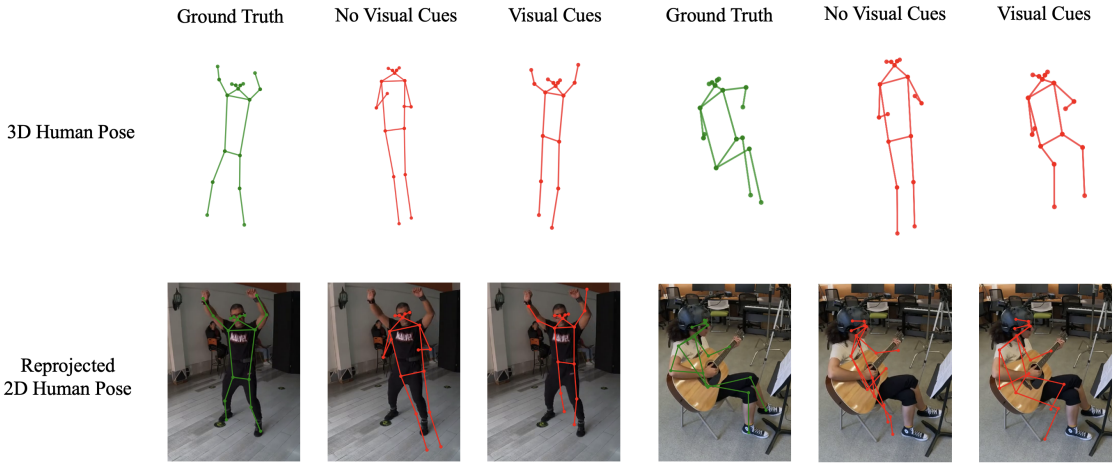

Figure 4. **3D Human Pose Estimation with and without Visual Inputs.** A common assumption in human pose estimation is that individuals always stand with their hands by their sides. However, integrating visual information into our Current-Frame Estimation module challenges this notion, accurately predicting when a person sits down or raises their hands.

5.1. Current-Frame Pose Estimation

Table 1 presents the comparative evaluation results for the state-of-the-art methods detailed in Grauman *et al*. [13] from the Ego-Exo4D [13] EgoPose task. We also compare our performance when using only visual streams as input (Visual-only) versus mixing proprioceptive and visual streams (Full). EgoCast significantly improves over all other state-of-the-art approaches, reducing the error by 22%. As shown in Table 1, EgoCast performs better in all evaluated scenarios, significantly reducing the MPJPE. Current state-of-the-art approaches utilize only the proprioceptive information (Location-based) or have sequential stages for using the proprioceptive and visual information (EgoEgo and Kinpoly). Using only visual inputs for EgoCast results in an improvement of 4.15 cm in comparison to the best-performing method in Ego-Exo4D [13]. Furthermore, the additional improvement of EgoCast (Full) shows that merging both information streams since the beginning is a strategy for exploiting the interaction between the internal and external cues for human pose estimation.

Notably, EgoCast achieves significant MPJPE reduction in the *Music* and *Health* scenarios. This can be attributed to the conventional assumption that subjects are standing; however, incorporating visual cues through EgoCast enables the accurate recognition of a seated posture, particularly affecting performance in the lower body joints (Figure 4). An in-depth view of this enhancement is shown in Figure 3, which represents the average error for each joint across all test sequences, using a spectrum from blue (low error) to red (high error). While the accuracy for facial key points (eyes, ears, nose) remains essentially the same with the addition of visual cues, there is a notable improvement in the accuracy for extremities. Through visual data integration, EgoCast improves the accuracy of lower-body joints, notably in the knees and ankles. Remarkably, the error in the ankles is reduced by 39%, emphasizing EgoCast's capability

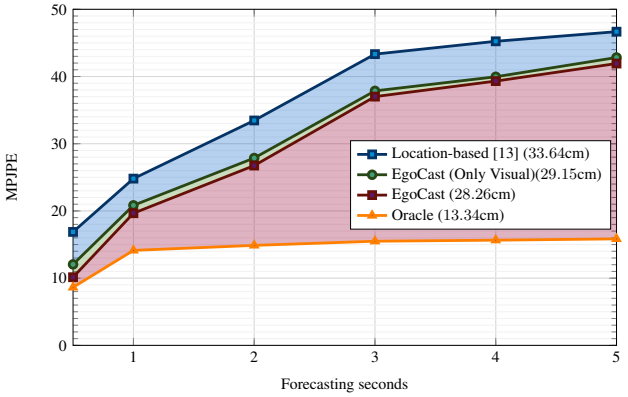

Figure 5. **Ego-Exo4D Forecasting at different timeframes.** We show performance curves for forecasting at {0.5, 1, 2, 3, 4, and 5} seconds in the future. We compare our final EgoCast approach against a forecasting extension of the current state-of-the-art method for current-frame pose estimation and an Oracle approach that aligns the trajectories with the ground truth. Note that since the graph shows MPJPE, lower curves represent better performance.

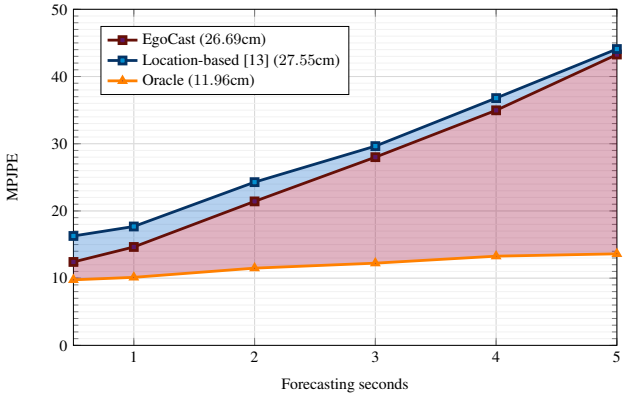

Figure 6. **ADT Overall Forecasting performance.** We compare our final EgoCast approach against a forecasting extension of the current state-of-the-art method for current-frame pose estimation and an Oracle approach that aligns the trajectories with the ground truth.

in accurately predicting the position of lower-body joints, which is challenging without the context provided by visual cues. Furthermore, the *Soccer* scenario sees the most significant improvement when transitioning from the visual-only approach to the full approach. This improvement arises because soccer involves extensive movement across the playing area, making the proprioceptive input highly relevant for capturing precise motion dynamics. The analysis of activities requiring rapid upper body movements, such as dance and basketball, shows that integrating visual data significantly reduces inaccuracies in predicting arm joint positions. Figure 4 demonstrates that, without visual information, the model inaccurately predicts the individual's arms as lowered instead of raised. In contrast, incorporating egocentric visual inputs leads to accurate predictions of the arms in a raised position. Figure 3 shows this enhancement, especially in the shoulders, elbows, and wrists, highlighted by blue tones, thanks to the addition of visual information.

Furthermore, Table 4 in the Supplementary Material presents the ablation performance of the current-frame estimation module on ADT [30], examining the use of visual streams and variations in window size. Similar to the findings of EgoCast on EgoExo-4D [13], the best improvement occurs when visual streams are integrated, highlighting the crucial role of combining proprioceptive inputs with visual data to achieve greater accuracy in pose estimation.

5.2. Pose Forecasting

Once we have the predictions from the current-frame estimation module, we use them as inputs for our forecasting module. We extend the previous best-performing method of Ego-Exo4D (Location-based in Tab. 1) to forecasting for a comprehensive state-of-the-art comparison. Furthermore, we establish an Oracle approach to understand the limitations of our method. Our Oracle is designed to align the translation of predicted trajectories with the ground truth, focusing primarily on the precision of the predicted 3D poses.

As shown in Figure 5, the error curves reveal a notable increase in prediction error over extended time frames. This trend aligns with expectations, considering the increasing unpredictability of human motion over time. The Oracle approach, serving as an upper bound for performance, evaluates solely the quality of pose prediction, excluding translational aspects. The considerable difference between EgoCast and the Oracle suggests that, while EgoCast effectively predicts poses, further refinement is needed in correctly estimating the trajectory. When comparing the performance between EgoCast and the Location-based approach, we find that EgoCast consistently performs better across all timeframes. This finding is consistent with the improvements shown for current-frame pose estimation.

Figure 6 illustrates the forecasting performance on the ADT dataset [30]. Notably, the performance aligns with the results observed for the Ego-Exo4D dataset [13], where our EgoCast approach significantly reduces the forecasting error compared to the previous state-of-the-art method.

Figure 7 qualitatively shows the performance of our EgoCast baseline against the ground truth for a temporal horizon of 4 seconds on EgoExo-4D [13] and ADT [30]. On the one hand, our method can predict accurate poses that resemble the ground truth. On the other hand, the predicted trajectories suffer from irregularities, indicating high translation changes between frames. Since forecasting in a large temporal window is a problem with several degrees of freedom, the method struggles to learn smooth transitions between

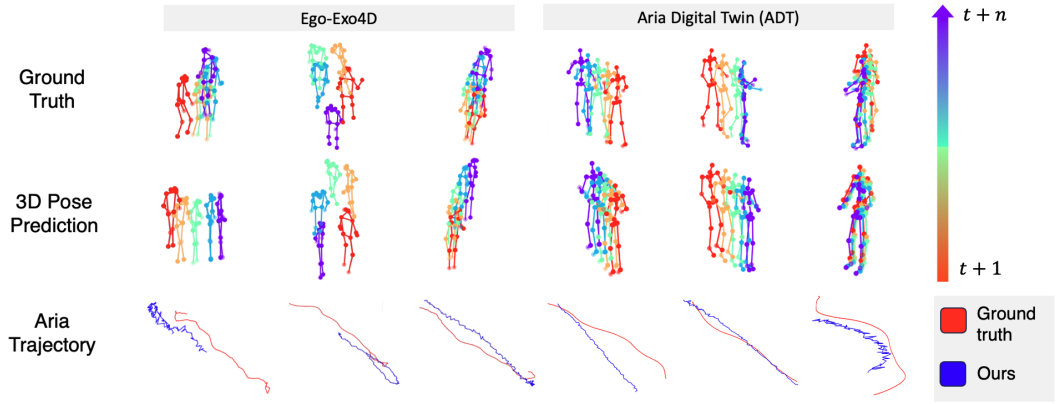

Figure 7. **Forecasting qualitative results**. We compare qualitatively the predicted 3D human poses against the ground-truth sequences. We also show a comparison of the trajectory prediction. We use an input window size of 20 and forecast 4 seconds (120 frames). Note that we subsampled the results for visualization. Our results show a high resemblance of the 3D poses and realistic trajectory estimation.

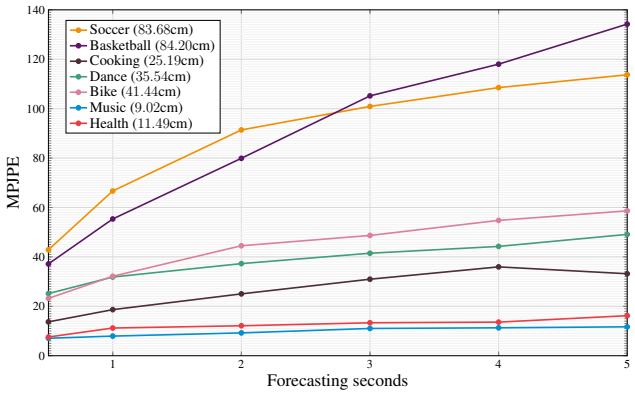

Figure 8. **Forecasting performance by activity.** For each activity in the Ego-Exo4D dataset, we show the performance of our method when forecasting up to 5 seconds into the future. Note that since the graph shows MPJPE, lower curves represent better performance.

frames. However, given the complex setup of predicting motion 4 seconds in the future, our method can accurately create plausible poses and trajectories.

Figure 8 presents the forecasting error for the eight activities within the Ego-Exo4D [13] dataset. Activities like music and health show minimal variations in both trajectory and pose, resulting in the lowest AUC values (9.02cm and 11.49cm). In contrast, bike repairing and cooking involve limited pose adjustments but require large trajectory alterations due to spatial movements associated with these tasks (*e.g.* going to the other side of the room to pick up a tool and returning to use it on a bike). Dance is characterized by minor trajectory shifts but significant pose variations. Although the requirements of the activity are the opposite of the cooking and bike repair, the error rates are similar. The most challenging activities for accurate forecasting are soccer and basketball, which demand frequent pose and trajectory changes reflected in the highest MPJPE values as the prediction time increases. These results highlight that complexity in translations and pose inherent to each activity varying over time significantly impacts the forecasting performance. Moreover, the per-category analysis of the ADT Database [30], presented in Figure 1 of the Supplementary Material, supports the ability of our EgoCast approach to predict plausible poses and trajectory movements. These findings underscore the applicability of our framework across different datasets for human pose estimation and trajectory forecasting in real-world scenarios.

6. Conclusion

In this paper, we present EgoCast, a new experimental framework aimed at human pose forecasting from an egocentric perspective. Our approach introduces a novel method for combining visual and proprioceptive data. We also devise a current-frame estimation module to generate pseudo-groundtruths to avoid using groundtruth poses as input at inference time. We believe that our work lays the groundwork for subsequent studies in human pose forecasting with egocentric inputs, potentially driving significant advancements in the field of computer vision and interactive technologies.

References

[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*, pages 767–776, 2024.

[2] Arij Bouazizi, Adrian Holzbock, Ulrich Kressel, Klaus Dietmayer, and Vasileios Belagiannis. Motionmixer: Mlp-based 3d human body pose forecasting. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence,*

IJCAI-22, pages 791–798. International Joint Conferences on Artificial Intelligence Organization, 7 2022.

[3] Robert Briscoe. Egocentric spatial representation in action and perception. *Philosophy and Phenomenological Research*, 79(2):423–460, 2009.

[4] Angela Castillo, Maria Escobar, Guillaume Jeanneret, Albert Pumarola, Pablo Arbeláez, Ali Thabet, and Artsiom Sanakoyeu. Bodiffusion: Diffusing sparse observations for full-body human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[5] 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*, pages 14750–14760, 2023.

[6] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018.

[7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(11):4125–4141, 2021.

[8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022.

[9] 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*, pages 481–490, 2023.

[10] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE international conference on computer vision*, pages 4346–4354, 2015.

[11] Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Diffpose: Toward more reliable 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13041–13051, 2023.

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

[13] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. *arXiv preprint arXiv:2311.18259*, 2023.

[14] Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Back to mlp: A simple baseline for human motion prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4809–4819, 2023.

[15] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics (TOG)*, 37(6):1–15, 2018.

[16] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 5308–5317, 2016.

[17] Hao Jiang and Kristen Grauman. Seeing invisible poses: Estimating 3d body pose from egocentric video. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3501–3509. IEEE, 2017.

[18] Hao Jiang and Vamsi Krishna Ithapu. Egocentric pose estimation from human vision span. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10986–10994. IEEE, 2021.

[19] 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 European Conference on Computer Vision*. Springer, 2022.

[20] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1346–1353. IEEE, 2012.

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

[22] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European conference on computer vision (ECCV)*, pages 619–635, 2018.

[23] Mei Liu, Bo Peng, and Mingsheng Shang. Lower limb movement intention recognition for rehabilitation robot aided with projected recurrent neural network. *Complex & Intelligent Systems*, 8(4):2813–2824, 2022.

[24] Yuxuan Liu, Jianxin Yang, Xiao Gu, Yijun Chen, Yao Guo, and Guang-Zhong Yang. Egofish3d: Egocentric 3d pose estimation from a fisheye camera via self-supervised learning. *IEEE Transactions on Multimedia*, 25:8880–8891, 2023.

[25] Zhengyi Luo, Ryo Hachiuma, Ye Yuan, and Kris Kitani. Dynamics-regulated kinematic policy for egocentric pose estimation. *Advances in Neural Information Processing Systems*, 34:25019–25032, 2021.

[26] Zhengyi Luo, Shun Iwase, Ye Yuan, and Kris Kitani. Embodied scene-aware human pose estimation. *Advances in Neural Information Processing Systems*, 35:6815–6828, 2022.

[27] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2891–2900, 2017.

[28] Tassos Natsakis and Lucian Busoniu. Predicting intention of motion during rehabilitation tasks of the upper-extremity. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 6037–6040. IEEE, 2021.

[29] 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*, pages 9890–9900, 2020.

[30] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20133–20143, 2023.

[31] Hyun Soo Park, Jyh-Jing Hwang, Yedong Niu, and Jianbo Shi. Egocentric future localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4697–4705, 2016.

[32] Thammathip Piumsomboon, Gun A Lee, Jonathon D Hart, Barrett Ens, Robert W Lindeman, Bruce H Thomas, and Mark Billinghurst. Mini-me: An adaptive avatar for mixed reality remote collaboration. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–13, 2018.

[33] Jianing Qiu, Lipeng Chen, Xiao Gu, Frank P-W Lo, Ya-Yen Tsai, Jiankai Sun, Jiaqi Liu, and Benny Lo. Egocentric human trajectory forecasting with a wearable camera and multi-modal fusion. *IEEE Robotics and Automation Letters*, 7(4):8799–8806, 2022.

[34] Kiran Somasundaram, Jing Dong, Huixuan Tang, Julian Straub, Mingfei Yan, Michael Goesele, Jakob Julian Engel, Renzo De Nardi, and Richard Newcombe. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023.

[35] Dongnan Su, Zhigang Hu, Jipeng Wu, Peng Shang, and Zhaohui Luo. Review of adaptive control for stroke lower limb exoskeleton rehabilitation robot based on motion intention recognition. *Frontiers in Neurorobotics*, 17, 2023.

[36] Yu-Chuan Su and Kristen Grauman. Detecting engagement in egocentric video. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 454–471. Springer, 2016.

[37] Zhenhua Tang, Zhaofan Qiu, Yanbin Hao, Richang Hong, and Ting Yao. 3d human pose estimation with spatio-temporal criss-cross attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4790–4799, 2023.

[38] 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*, pages 7728–7738, 2019.

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

[40] Jian Wang, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, and Christian Theobalt. Estimating egocentric 3d human pose in global space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11500–11509, 2021.

[41] 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*, pages 13031–13040, 2023.

[42] Jinbao Wang, Shujie Tan, Xiantong Zhen, Shuo Xu, Feng Zheng, Zhenyu He, and Ling Shao. Deep 3d human pose estimation: A review. *Computer Vision and Image Understanding*, 210:103225, 2021.

[43] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, et al. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20270–20281, 2023.

[44] Chenxin Xu, Robby T Tan, Yuhong Tan, Siheng Chen, Yu Guang Wang, Xinchao Wang, and Yanfeng Wang. Eqmotion: Equivariant multi-agent motion prediction with invariant interaction reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1410–1420, 2023.

[45] Haitao Yan, Qiongjie Cui, Jiexin Xie, and Shijie Guo. Forecasting of 3d whole-body human poses with grasping objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1726–1736, 2024.

[46] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13167–13178, 2022.

[47] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021.

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

[49] Siwei Zhang, Qianli Ma, Yan Zhang, Sadegh Aliakbarian, Darren Cosker, and Siyu Tang. Probabilistic human mesh recovery in 3d scenes from egocentric views. *arXiv preprint arXiv:2304.06024*, 2023.

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

[51] Yu Zhang, Songpengcheng Xia, Lei Chu, Jiarui Yang, Qi Wu, and Ling Pei. Dynamic inertial poser (dynaip): Part-based motion dynamics learning for enhanced human pose estimation with sparse inertial sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1889–1899, 2024.

[52] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, C Karen Liu, and Leonidas J Guibas. Gimo: Gaze-informed human motion prediction in context. In *European Conference on Computer Vision*, pages 676–694. Springer, 2022.