

# The Invisible EgoHand: 3D Hand Forecasting through EgoBody Pose Estimation

Masashi Hatano<sup>1\*</sup> Zhifan Zhu<sup>2</sup> Hideo Saito<sup>1</sup> Dima Damen<sup>2</sup>

<sup>1</sup> Keio University <sup>2</sup> University of Bristol

<https://masashi-hatano.github.io/EgoH4>

## Abstract

Forecasting hand motion and pose from an egocentric perspective is essential for understanding human intention. However, existing methods focus solely on predicting positions without considering articulation, and only when the hands are visible in the field of view. This limitation overlooks the fact that approximate hand positions can still be inferred even when they are outside the camera’s view. In this paper, we propose a method to forecast the 3D trajectories and poses of both hands from an egocentric video, both in and out of the field of view.

We propose a diffusion-based transformer architecture for Egocentric Hand Forecasting, EgoH4, which takes as input the observation sequence and camera poses, then predicts future 3D motion and poses for both hands of the camera wearer. We leverage full-body pose information, allowing other joints to provide constraints on hand motion. We denoise the hand and body joints along with a visibility predictor for hand joints and a 3D-to-2D reprojection loss that minimizes the error when hands are in-view.

We evaluate EgoH4 on the Ego-Exo4D dataset, combining subsets with body and hand annotations. We train on 156K sequences and evaluate on 34K sequences, respectively. EgoH4 improves the performance by 3.4cm and 5.1cm over the baseline in terms of ADE for hand trajectory forecasting and MPJPE for hand pose forecasting.

## 1. Introduction

Understanding human motion from an egocentric video is critical for a variety of applications, including AR/VR, human-robot interaction, and assistive technology. Unlike category-level discrete predictions, such as action recognition [28, 54–56, 65] or action anticipation [19, 21, 60, 66, 71], motion provides fine-grained, continuous predictions.

Prior works predicting future hand positions are 2D-based, estimating the location within a moving camera frame. These works focus on hand-object interactions [40,

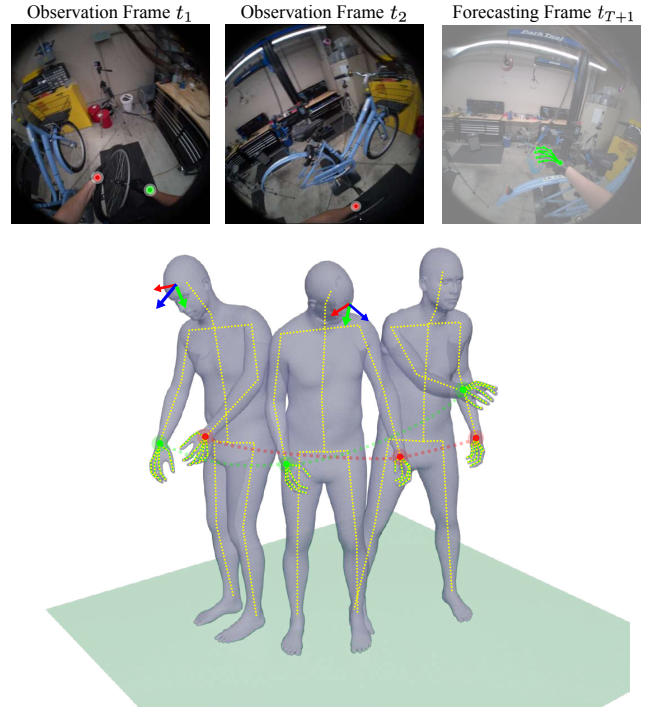


Figure 1. Given signals during observation: camera poses, images, and visible hand locations in 2D, our proposed method **EgoH4** forecasts future 3D hand pose. EgoH4 can forecast hand joints even when hands are out of view during observation. We show visible 2D hand positions overlaid on the observation frames  $t_1$  and  $t_2$ , and the corresponding camera poses attached on the heads. At  $t_2$ , the right hand is invisible. In the forecasting frame, the right hand is back in view while the left hand is now out of view.

41] or ego-motion awareness [29, 43, 44]. However, 2D methods are unable to give persistent predictions of the hand in the environment due to the camera motion. Instead, recent works have targeted predicting in a world coordinate frame, with successful methods targeting body pose [22, 35, 48, 58, 74], ego body pose [2, 5, 11, 37, 68, 77, 81] and recently 3D hand trajectories [3]. Works that focus on 3D forecasting body pose [9, 17, 78], hand location or pose [3, 64] remain scarce.

The current egocentric hand forecasting task has three

\*Work done during a research visit to the University of Bristol

significant limitations we address here:

**Out-of-view.** Previous works [3, 40, 41] have not considered scenarios where hands move out of the field of view during observations, yet this is a common occurrence in egocentric videos (as shown in Fig. 1). Hands are visible primarily when interacting with objects or just before intentional contact. For example, when reaching for distant objects, we initially move our body closer, only extending our hand once we are within reach. Consequently, relying on visible observations for forecasting can result in delayed predictions, limiting early-stage forecasting capabilities.

**Body Movements Awareness.** Given the natural coordination between the hands and body, incorporating body movements could enhance the accuracy of hand forecasting. Predicting body poses alongside hand poses helps prevent unrealistic hand motions as body joints serve as the constraints.

**Hand Articulation.** Prior egocentric hand forecasting works [3, 40, 41] predict a hand position without finger articulation. While HoloAssist [70] recently established a benchmark for egocentric hand pose forecasting without introducing a specific method, the task of hand pose forecasting in egocentric videos is still largely underexplored.

To address these limitations, we propose **EgoH4**, a 3D hand forecasting method that leverages body pose estimation to predict 3D hand motion and pose. To our knowledge, this is the first work to attempt 3D hand trajectory and pose forecasting when hands are out of the field of view, i.e., *invisible*. We achieve this by (1) jointly optimizing the hands and body joints. Leveraging the body pose knowledge helps locate the hand joints when they are out of frame by constraining hand joints relative to the body pose. Additionally, we (2) incorporate a classifier that estimates hand visibility. This improves the capability of dealing with invisible hands and enhancing hand forecasting accuracy.

We evaluate EgoH4 on the Ego-Exo4D [26] dataset, which offers 3D annotations even when hands are outside the camera’s field of view thanks to the multiple exocentric cameras. In summary, our contributions are:

- We are the first to address egocentric 3D hand forecasting when hands are in- or out-of-view, during both the observation and the forecasting timesteps.
- We also extend the task of egocentric hand trajectory forecasting to hand pose forecasting for a fine-grained understanding of human intention.
- We propose EgoH4, a diffusion-based transformer model jointly denoising body pose and hands along with a visibility predictor and 3D-to-2D reprojection regularization.
- We evaluate EgoH4 on the Ego-Exo4D dataset, a large-scale egocentric dataset. From available annotations, we curate a 3D hand forecasting task, resulting in 156K training sequences and 34K testing sequences.
- We improved the hand trajectory forecasting accuracy in ADE by 5.5cm, 1.9cm, and 3.4cm, and hand pose fore-

casting in MPJPE by 5.0 cm, 5.9 cm, and 5.1 cm for in-view, out-of-view, and overall sequences, respectively.

## 2. Related Work

We review related works on hands in egocentric videos, egocentric hand forecasting, motion forecasting, and ego-body pose estimation.

### 2.1. Understanding Hands in Egocentric Videos

Hand-object interactions are best studied in egocentric videos. Prior works have addressed 2D hand detection and side classification [7, 10, 63], hand segmentation [12, 32, 79], grasp type classification [24] and hand pose estimation [51–53, 57, 59]. Other works also involve modeling hand-object interactions [14, 18, 20, 23, 49, 62] and object affordance [24, 75]. These works provide robust methods to solve 2D hand-related tasks for egocentric videos.

### 2.2. Egocentric Hand Forecasting

Initial efforts [40, 41] in 2D hand trajectory forecasting from egocentric videos aim to understand human intention, often combined with interaction hotspot prediction and action anticipation, as hand motion is a key cue for anticipation. Recent works [29, 43, 44] consider the ego-motion, the head motion of the human, to improve the 2D hand trajectory. However, predicting hand locations only in the 2D image plane limits the range of forecasting, as hands move in a much wider range in 3D space.

USST [3] is the first work to address 3D hand forecasting and propose a pipeline to lift up from a manually annotated 2D hand landmark into 3D to acquire training data. Also, they propose the uncertainty-aware state space transformer model that takes the 3D hand trajectory and egocentric videos as input and forecasts the 3D hand trajectory. However, their scope is limited to position without pose and only when the hands are visible.

We are the first to address 3D hand forecasting even when hands are out-of-view during observation, and the first to explore forecasting hand poses (not only positions).

### 2.3. Body Motion Forecasting

Forecasting human motion has gained significant attention from various perspectives [6, 13, 31, 47]. While most studies focus on body pose, a few address hand articulation [64] or integrate it with body pose [72]. Among these, GCN-based methods [38, 45] are widely used and effective, where joints serve as nodes and edges capture spatial relationships. Other works [4, 73] first predict 2D keypoints from images and estimate 3D poses, subsequently forecasting future motions based on the estimated past motions. However, these methods make strong assumptions: past motions are accessible or body joints are visible, limiting their applicability in egocentric scenarios.

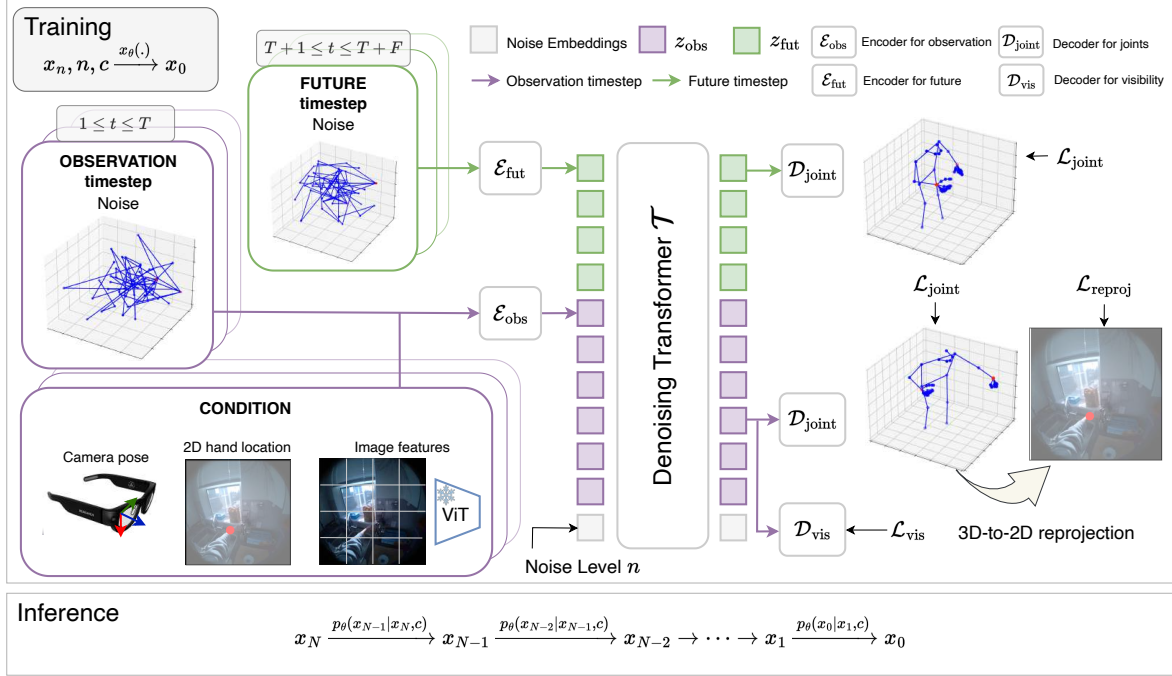


Figure 2. **The framework of our proposed method, EgoH4.** We show the denoising network in a single denoising step. During training, we estimate the original data  $x_0$  from an arbitrary noise level  $n$  to learn the denoising network. During inference, we iteratively denoise the noisy joints over the maximum diffusion step  $N$  from  $N$  to 0.

On the other hand, several works [17, 78] address body pose forecasting from an egocentric perspective. Ego-Cast [17] proposes a two-stage approach: first estimate the body pose of the camera wearer, and then forecast motion from the estimated body poses. These works [17, 78] do not predict hand articulation or report performance on hand position specifically.

## 2.4. Egocentric 3D Body Pose Estimation

There has been a growing interest in 3D human body pose estimation from egocentric videos. Several works [2, 68, 69] utilize the fisheye camera, which provides a large field of view to ensure body joints are visible. EgoPoser [33] uses a wearable device with hand joint locations as input. You2Me [50] focuses on estimation via interaction with other people [80]. Other works use deep RL-based approach [78] or physically plausible predictions [42].

EgoEgo [37] is the first work that proposed a diffusion-based body generation model conditioned on head poses. Other works [5, 8, 33, 76] investigate this task using ground-truth head pose as a condition. This can be provided by smart glasses devices like Project Aria [16].

In this work, we also estimate the egocentric human body poses from head poses, and we leverage recent advancement in egocentric body pose estimation using diffusion models; however, our focus is not on improving the body estimation but leveraging it for 3D hand forecasting.

## 3. Our Method, EgoH4

We propose EgoH4, a diffusion-based transformer architecture for 3D hand forecasting, which takes head poses, 2D hand locations, and image features as input during the observation period and aims to predict future 3D hand poses.

We first introduce a novel egocentric 3D hand pose forecasting problem setup (Sec. 3.1). Then, we introduce our proposed method, EgoH4, with full body motion estimation (Sec. 3.2), and training objective (Sec. 3.3). Fig. 2 provides an overview of our approach.

### 3.1. Problem Setup

Given an input egocentric video and the corresponding camera poses for  $T$  observation frames, the goal is to forecast the 3D hand poses  $Y_{fut} = \{y^{T+1}, \dots, y^{T+F}\}$  for the future time horizons  $F$ . At each timestep  $T+1 \leq t \leq T+F$ ,  $y^t$  consists of left/right hand joints. Importantly, we employ sequence canonicalization [37, 76], where 3D points are expressed relative to a canonicalized coordinate system from the first camera coordinate.

### 3.2. EgoH4 Architecture

Leveraging knowledge of 3D body movements helps enhance hand forecasting when the hand is out of view, during observation, or forecasting. Since body joints are mostly invisible and hand visibility frequently changes, there is no direct, deterministic mapping function for estimating these

joints from these observations. Therefore, we adopt a generative approach based on the denoising diffusion probabilistic model (DDPM) [30] for estimating and forecasting 3D hand and body motion.

**Conditional Diffusion Model.** The diffusion model takes random noise sampled from Gaussian as initial motion  $x_N^t \in \mathbb{R}^{J \times 3}$ <sup>1</sup> at temporal timestep  $t$  with the maximum diffusion step  $N$  and the number of joints  $J$ . It iteratively removes noise at each diffusion step  $n$  with a learned mean and fixed variance:

$$p_\theta(x_{n-1} | x_n, c) := \mathcal{N}(x_{n-1}; \mu_\theta(x_n, n, c), \sigma_n^2 I), \quad (1)$$

where  $\mu_\theta(\cdot)$  can be computed by a neural network, and  $c$  is any given conditions (we detail our conditions later), to generate the 3D positions.  $\mu_\theta(\cdot)$  is parameterized as follows:

$$\mu_\theta(x_n, n, c) = \frac{\sqrt{\alpha_n(1-\bar{\alpha}_{n-1})}}{1-\bar{\alpha}_n} x_n + \frac{\sqrt{\bar{\alpha}_{n-1}(1-\alpha_n)}}{1-\bar{\alpha}_n} x_\theta(x_n, n, c), \quad (2)$$

where  $\bar{\alpha}_n = \prod_{s=0}^n \alpha_s$ ,  $\alpha_n$  is a fixed parameter. We train the network  $x_\theta(x_n, n, c)$  to directly predict  $x_0$  during training time, following [37]. The training loss for the 3D hand and body joints prediction is defined as a reconstruction loss of the original data  $x_0$ :

$$\mathcal{L} = \mathbb{E}_{x_0, n} \|x_\theta(x_n, n, c) - x_0\|_1. \quad (3)$$

At the inference time, we apply  $N$  steps of denoising with Eq. (1) to obtain the final denoised output  $\hat{x}_0$ .

**Conditioning Cues.** We use three conditioning cues: i) camera pose, ii) 2D hand locations (if visible), and iii) image features. Humans have a remarkable ability to stabilize their heads, keeping them aligned with the body’s center of mass [37]. The camera pose, which captures the head pose given this is a head-mounted camera, provides crucial information for estimating body joints, including hand locations. The camera pose  $c_{\text{cam}}^t \in \mathbb{R}^9$  at timestep  $t$  consists of the 6D represented rotation [83] and translation vector, canonicalized w.r.t the first frame without losing gravity information. The left/right hand locations  $(x, y)$  coordinates in 2D image space  $c_{\text{left}}^t, c_{\text{right}}^t \in \mathbb{R}^2$  are utilized when they are visible to help improve forecasting accuracy. When one hand is not visible, we replace the location of this hand side  $c_{\text{side}}^t$  by  $(-1, -1)$  to indicate it is invisible. We also leverage the image features  $c_{\text{img}}^t \in \mathbb{R}^{d_{\text{img}}}$ , extracted from a visual encoder [15], to provide visual context. This complements camera poses but, importantly, allows observing any body parts in the camera view.

**Noise Encoder.** During the observation, the diffusion model is conditioned on the above three cues, while no conditioning is used for future timesteps. Thus, we use

two types of noise encoders: conditional noise encoder  $\mathcal{E}_{\text{obs}}$  shared across observation timesteps and unconditional noise encoder  $\mathcal{E}_{\text{fut}}$  shared for forecasting timesteps. The noise tokens for observation  $z_{\text{obs}}^t \in \mathbb{R}^{d_z}$  at time  $t$  are encoded as follows:

$$z_{\text{obs}}^t = \mathcal{E}_{\text{obs}}(x_n^t, c_{\text{cam}}^t, c_{\text{left}}^t, c_{\text{right}}^t, c_{\text{img}}^t). \quad (4)$$

As for future timesteps, these conditions cannot be obtained. We adopt a linear layer as the noise encoder  $\mathcal{E}_{\text{fut}}$  to directly tokenize the noise:

$$z_{\text{fut}}^t = \mathcal{E}_{\text{fut}}(x_n^t), \quad (5)$$

where  $t > T$  and  $z_{\text{fut}}^t$  has the same dimension as the tokens for the observation encoder.

**Denoising Transformer.** We adopt the Transformer architecture [67] as the denoising function to deal with our sequential input. The encoded noise tokens  $z_{\text{obs}}$  and  $z_{\text{fut}}$  are concatenated and combined with the noise level information (embedding), followed by adding positional embeddings. The denoising function  $\mathcal{T}$  takes these combined tokens as input. The final output  $\hat{x}_0^t$  for each temporal timestep is obtained after passing through the decoder  $\mathcal{D}_{\text{joint}}$ , a linear layer, for hand and body joints:

$$\hat{x}_0^{1:T+F} = \mathcal{D}_{\text{joint}}(\mathcal{T}([z_{\text{obs}}^{1:T}, z_{\text{fut}}^{T+1:T+F}], n)). \quad (6)$$

### 3.3. Training Objective

We train our model using three losses: (1) 3D joint reconstruction loss  $\mathcal{L}_{\text{joint}}$ , (2) visibility loss  $\mathcal{L}_{\text{vis}}$ , and (3) 2D reprojection loss for visible hands  $\mathcal{L}_{\text{reproj}}$ . We detail these next.

**3D Joint Loss.** As shown in several egocentric human body pose estimation works [37, 42, 76, 78], the body pose can be estimated given the gravity-aligned camera pose. In our work, the conditional diffusion model reconstructs 3D body joints in addition to the 3D hand poses for all timesteps: observation and forecasting. The 3D joint loss  $\mathcal{L}_{\text{joint}}$  is computed from the error between the prediction  $\hat{x}_0$  and the ground-truth data  $x_0$  using Eq. (3).

**Visibility Loss.** Furthermore, we incorporate a visibility loss so the model correctly perceives when hands are in- or out-of the field of view. This is analogous to the visibility loss used in point tracking to address occlusion, which is one of the significant issues that lead to tracking errors. Inspired by point tracking methods [27, 34], we incorporate the visibility loss to increase the model’s ability to position hands in/out-of-view, and, importantly, regulate the learning. We predict the visibility score  $\hat{v}^{1:T} \in \mathbb{R}^{2T}$  for both hands in each observation timestep using the decoder  $\mathcal{D}_{\text{vis}}$ :

$$\hat{v}^{1:T} = \mathcal{D}_{\text{vis}}(\mathcal{T}([z_{\text{obs}}^{1:T}, z_{\text{fut}}^{T+1:T+F}], n)). \quad (7)$$

We employ the binary cross-entropy (CE) loss as visibility loss  $\mathcal{L}_{\text{vis}}$ ,

$$\mathcal{L}_{\text{vis}} = \text{CE}(v^{1:T}, \hat{v}^{1:T}), \quad (8)$$

<sup>1</sup> We use superscript  $t$  for time horizon step and subscript  $n$  for diffusion denoising step.



where the target  $v^{1:T}$  can be obtained from the ground-truth. **2D Reprojection Loss.** We use the 2D hand location as input when the hand is visible to help improve the 3D hand pose estimation and forecasting. However, merely adding 2D locations as input does not maintain the consistency between 2D hand input and 3D hand output. We use the reprojection loss to penalize the error between input 2D coordinates and reprojected 2D location from the 3D output using extrinsic  $P$  and intrinsic  $K$  camera parameters. We only use the wrist position for each hand side in this reprojection loss, defined as:

$$\mathcal{L}_{\text{reproj,side}} = \sum_{t=1}^T v_{\text{side}}^t \|c_{\text{side}}^t - \Pi_K(P(\hat{x}_{\text{side}}^t))\|_1, \quad (9)$$

where side is left/right,  $\Pi_K$  denotes the projection onto 2D image space and  $\hat{x}_{\text{side}}$  represents the reconstructed hand locations. We have  $\mathcal{L}_{\text{reproj}} = \mathcal{L}_{\text{reproj,left}} + \mathcal{L}_{\text{reproj,right}}$ .

**Training Loss.** We train our encoders, decoders, and denoising function with a linear combination of these losses with balancing hyperparameters for the final training loss:

$$\mathcal{L} = \mathcal{L}_{\text{joint}} + \lambda_{\text{vis}} \mathcal{L}_{\text{vis}} + \lambda_{\text{reproj}} \mathcal{L}_{\text{reproj}}. \quad (10)$$

## 4. Experiments

In this section, we elaborate on the dataset used to train and evaluate EgoH4 (Sec. 4.1), implementation details (Sec. 4.2), methods we compare to (Sec. 4.3), quantitative results (Sec. 4.4), ablation study of our proposed method (Sec. 4.5), and qualitative results (Sec. 4.6).

### 4.1. Dataset

We use the recently released Ego-Exo4D dataset, a diverse and large-scale multimodal multiview video dataset. The dataset is released with two *separate* sets of manual annotations: one for body pose (including wrist but without hand pose) and the second for hand pose. As no prior work has targeted hand forecasting using body pose, we curate our dataset from these annotations as follows:

- Ego-Exo4D Body Pose: We use the manual annotations providing 17 joints of body and wrist (without hand pose).
- Ego-Exo4D Hand Pose: We use the manual annotations providing  $21 \times 2$  hand joints, along with automatic body annotations from exocentric cameras. We use the manual wrist annotations with the hand pose, overwriting those from the automatic body annotations. We only use the automatic body annotations for training (not evaluation).

We use the same train/val splits for both sets of annotations from [26], combining these pool of annotations to form a dataset with 3D hand and body poses even when hands are not visible from the egocentric camera. During training with this heterogeneous data, we only backpropagate the loss on joints we have annotations for.

Table 1. **Dataset Comparison of Train/Test Sequences.** We report the number of training and testing sequences for each dataset, H2O, EgoPAT3D, and Ego-Exo4D, categorized into in-view, out-of-view scenarios, and total sequences. The sequence counts are provided separately for each hand side. Moreover, we report the availability of body pose annotation.

Dataset	Body Pose	In-view		Out-of-view		All	
		train	test	train	test	train	test
H2O [36]		9.9k	3.7k	-	-	9.9k	3.7k
EgoPAT3D [39]		7.2k	3.8k	-	-	7.2k	3.8k
Ego-Exo4D [26] (Body)	✓	52.4k	11.6k	85.2k	18.9k	138k	30.5k
Ego-Exo4D [26] (Hand)	✓	14.2k	3.4k	4.5k	0.1k	18.7k	3.5k
Ego-Exo4D [26] (Ours)	✓	66.6k	15.0k	89.7k	19.0k	156k	34.0k

Aligned with prior forecasting works [25, 29], we define the task so that observation is a two-second temporal duration, followed by one second of forecasting<sup>2</sup>.

Tab. 1 showcases the annotations we combine to form our train/test splits and those of previous datasets used to evaluate the hand forecasting. We evaluate 15x and 9x more sequences for training and testing, respectively, compared to the H2O [36] dataset. We separate sequences into those where all observation frames have in-view hands and those where one or more observation frames have out-of-view hands. Note that these previous datasets are not only significantly smaller but also do not have 3D hand annotation when hands are out-of-view, so not suitable for our experiments.

### 4.2. Implementation Details

**Experimental Setup.** We sample at 10 FPS (frames per second) for observation and forecasting. As a result, we have  $T = 20$  sampled frames for input observation and  $F = 10$  sampled frames for forecasting. Visual features are extracted from a pre-trained on ImageNet and frozen ViT-S. We reconstruct the hand and body joints, with the number of joints  $J = 57 - (15 \text{ body joints exc. wrist} + 21 \times 2 \text{ hand joints})$ . We use the ground-truth 2D hand locations for input, normalized to the range of  $[0, 1]$ .

**Training.** We train the model from random weights for 40K iterations with a constant learning rate of  $1e - 4$ . Regarding the parameters for the objective function, we empirically choose each balancing hyperparameter  $\lambda_{\text{vis}}$  and  $\lambda_{\text{reproj}}$  to  $1e - 1$  and  $5e - 2$ , respectively. (See suppl. for ablation).

**Evaluation Metrics.** We report the Average Displacement Error (ADE) and Final Displacement Error (FDE) in global 3D space for wrist trajectory forecasting, often used in trajectory forecasting works. Regarding hand pose forecasting, we adopt the Mean Per Joint Position Error (MPJPE), which averages all future timesteps, and MPJPE-F, which averages the performance at the last forecasting frame. Furthermore, we report MPJPE and Mean Per Joint Velocity Error (MPJVE) to evaluate the accuracy of body pose es-

<sup>2</sup>Note that [3] adopts a different protocol (0.8s obs and 0.53s forecast). We re-train this method with the standard protocol for direct comparison.

Table 2. **Hand Forecasting Accuracy.** We report the hand trajectory and pose forecasting results on in-view, out-of-view, and all scenarios on the Ego-Exo4D dataset.

Method	Hand Trajectory Forecasting						Hand Pose Forecasting					
	In-view		Out-of-view		All		In-view		Out-of-view		All	
	ADE	FDE	ADE	FDE	ADE	FDE	MPJPE	MPJPE-F	MPJPE	MPJPE-F	MPJPE	MPJPE-F
Static	0.199	0.209	0.434	0.546	0.335	0.405	0.163	0.176	0.297	0.325	0.166	0.179
CVM [61]	0.201	0.217	0.451	0.648	0.346	0.467	0.162	0.177	0.352	0.427	0.166	0.183
EgoEgoForecast	0.171	0.185	0.385	0.472	0.295	0.352	0.162	0.173	0.299	0.345	0.166	0.177
USST [3]	0.277	0.280	0.763	0.792	0.562	0.581	-	-	-	-	-	-
Ours	<b>0.116</b>	<b>0.152</b>	<b>0.366</b>	<b>0.459</b>	<b>0.261</b>	<b>0.324</b>	<b>0.112</b>	<b>0.140</b>	<b>0.240</b>	<b>0.280</b>	<b>0.115</b>	<b>0.143</b>

timation and forecasting. All 3D evaluation metrics are reported in meters. We generate one sample during evaluation for fair comparisons with the existing deterministic models.

### 4.3. Baselines

We use two naive baselines and three previous works (one for body pose estimation and one adapted to forecasting):

- **Static** is a naive baseline that keeps the average whole-body pose of training data at the last observable timestep. It showcases the difficulty of the dataset.
- **CVM [61]** is another naive baseline that is often used in trajectory forecasting. The Constant Velocity Model (CVM) assumes that the most recent relative velocity is the most relevant predictor for future trajectory.
- **EgoAllo [76]** is a diffusion-based model for body pose estimation. The method is not designed for forecasting and does not attempt it. We evaluate the guidance-free version of EgoAllo, pre-trained on AMASS [46] dataset<sup>3</sup>.
- **EgoEgoForecast (Baseline)** We extend [37] to a forecasting method. Similar to our proposed method, there are two noise encoders: one is for encoding with head poses during observation, and the other is for encoding random noise. This model can be seen as an architectural baseline for our EgoH4, as it is a vanilla diffusion model without the additional conditioning losses we introduce. The model is trained from scratch on Ego-Exo4D dataset.
- **USST [3]** is the only prior work that evaluates egocentric 3D hand trajectory forecasting. We retrained USST on our dataset using the official implementation. When the hand is out-of-view during observation, we use masking for both training and inference.

### 4.4. Quantitative Results

**Hand Trajectory and Pose Forecasting.** We compare the performance of egocentric 3D hand trajectory forecasting with the noted baselines above on the Ego-Exo4D dataset. We report ADE and FDE in two different scenarios: 1) *in-view* scenario where a hand is in-view in all observation timesteps, same as the previous evaluation setup, and 2) *out-of-view* scenario where a hand is out of the field of view at

<sup>3</sup>We use the SMPL’s internal joint regressor to convert into MS COCO 17 body joints.

Table 3. **Architecture and Losses Ablations.**  $\mathcal{L}_{\text{body}}$  represents the reconstruction loss for body joints. The  $\mathcal{L}_{\text{obs}}$  represents all losses in the observation timesteps, including 3D joint loss during observation, reprojection loss, and visibility loss.

Method	Hand Trajectory Forecasting			Hand Pose Forecasting		
	In-view	Out-of-view	All	In-view	Out-of-view	All
EgoEgoForecast	0.171	0.385	0.295	0.162	0.299	0.166
Ours w/o. 2D joint	0.151	0.377	0.282	0.139	0.269	0.142
Ours w/o. image	<b>0.116</b>	0.367	<b>0.261</b>	0.117	<b>0.234</b>	0.120
Ours w/o. $\mathcal{L}_{\text{reproj}}$	0.132	0.368	0.269	0.125	0.250	0.128
Ours w/o. $\mathcal{L}_{\text{vis}}$	0.127	0.377	0.272	0.121	0.240	0.124
Ours w/o. $\mathcal{L}_{\text{body}}$	0.129	0.385	0.277	0.120	0.258	0.123
Ours w/o. $\mathcal{L}_{\text{obs}}$	0.149	0.390	0.289	0.139	0.250	0.142
Ours	<b>0.116</b>	<b>0.366</b>	<b>0.261</b>	<b>0.112</b>	0.240	<b>0.115</b>

Table 4. **Evaluation on Different Out-of-view Ratio Intervals.** We report the ADE and FDE results across five equally divided out-of-view ratio intervals  $\gamma_{(i,j)}$  ranging from zero to one.

Method	$\gamma_{(0.0,0.2]}$		$\gamma_{(0.2,0.4]}$		$\gamma_{(0.4,0.6]}$		$\gamma_{(0.6,0.8]}$		$\gamma_{(0.8,1.0]}$	
	ADF	FDE	ADF	FDE	ADF	FDE	ADF	FDE	ADF	FDE
EgoEgoForecast	0.284	0.329	0.393	0.475	0.424	0.519	0.415	0.504	0.363	0.459
Ours w/o. $\mathcal{L}_{\text{body}}$	0.250	0.300	0.398	0.477	0.432	0.519	0.430	0.514	0.353	0.442
Ours w/o. $\mathcal{L}_{\text{vis}}$	0.254	0.303	0.394	0.487	0.423	0.521	0.412	<b>0.498</b>	0.349	0.447
Ours	<b>0.236</b>	<b>0.284</b>	<b>0.379</b>	<b>0.476</b>	<b>0.417</b>	<b>0.517</b>	<b>0.404</b>	0.505	<b>0.335</b>	<b>0.434</b>

Table 5. **Body Pose Estimation/Forecasting Accuracy.** Comparison with the body pose estimation/forecasting in terms of MPJPE and MPJVE. The location-based model, used as a baseline of the body pose estimation in the Ego-Exo4D, is a transformer that takes head poses as input to output the body joints. \* denotes the method is not trained on the Ego-Exo4D.

Method	Observation		Forecasting	
	MPJPE	MPJVE	MPJPE	MPJVE
Static	0.357	0.778	0.286	0.778
Location-based [26]	0.148	<b>0.583</b>	-	-
EgoAllo [76]*	0.219	-	-	-
EgoEgoForecast	0.173	0.651	0.245	0.771
Ours	<b>0.142</b>	0.697	<b>0.221</b>	<b>0.763</b>

Table 6. **Hand Pose Estimation Accuracy.** Comparison with the hand pose estimation in terms of MPJPE.

Method	In-view	Out-of-view	All
THOR-net [1]	0.051	-	-
POTTER [82]	<b>0.031</b>	-	-
EgoEgoForecast	0.158	0.289	0.161
Ours	0.067	<b>0.206</b>	<b>0.081</b>

least one timestep during observation. We consider the out-of-view sample for each hand side: the same sequence can be in-view for left hand but out-of-view for right hand. We

show the results in Tab. 2.

Naive baselines that assume no motion (i.e., static position) or constant velocity fail to capture the complex dynamics of future hand movements. This shows the challenging aspect of the 3D hand forecasting task and this dataset. USST, poorly performs especially when hands are out-of-view. It indicates that 2D hand + 3D camera pose is more suitable as input than 3D hand alone, as the latter is unavailable when hands are out-of-view.

Our proposed method, EgoH4, outperforms the baseline EgoEgoForecast, which forecasts hands and body joints, across both in-view and out-of-view scenarios. Notably, significant improvements are observed in the in-view scenario, as the proposed model incorporates visual cues during observation, such as image features and 2D hand locations, which the baseline lacks. Additionally, while EgoEgoForecast leverages body movement information to enable forecasting for out-of-view scenarios, our model further improves on this by introducing awareness of in-view and out-of-view status for each hand side during observation, resulting in superior forecasting accuracy.

We also evaluate the hand pose forecasting performance except for USST as the method cannot predict multiple joints. EgoH4 outperforms baselines on in- and out-of-view scenarios. The improvements over EgoEgoForecast support the utilization of visible cues for pose forecasting.

#### 4.5. Ablation Analysis

**Architecture and Losses Ablation.** This ablation study focuses on the conditional input and loss components to verify the contribution of each module. We evaluate the hand trajectory and pose forecasting in Tab. 3. We experiment with removing each component individually and compare with the baseline EgoEgoForecast and the full model of EgoH4, including 1) visible 2D hand joints, 2) image features, 3) 2D reprojection loss  $\mathcal{L}_{\text{reproj}}$ , 4) visibility loss  $\mathcal{L}_{\text{vis}}$ , 5) body pose loss  $\mathcal{L}_{\text{body}}$ , and 6) losses during observation  $\mathcal{L}_{\text{obs}}$ .

Without 2D hand joint coordinates, significant performance drops can be seen in both in-view and out-of-view sequences. In contrast, the image features only marginally improve performance and marginally harm performance for out-of-view hand pose forecasting. The 2D reprojection loss serves as an effective regularization, further boosting hand forecasting performance and underscoring the importance of maintaining spatial consistency between observed and predicted hand positions in 2D image space. Notably, without our proposed visibility loss or body pose loss, hand trajectory forecasting performance degrades significantly in out-of-view scenarios. Lastly, estimating and optimizing joints for the observation timestamps  $\mathcal{L}_{\text{obs}}$ , instead of solely optimizing 3D joints during the forecasting period, also improves results for trajectory and pose forecasting.

Overall, our full model performs the best among the vari-

ants of our model in the entire evaluation set.

**Impact of Out-of-view Ratio During Observation.** We conduct an in-depth analysis to better understand performance as the ratio of hands out-of-view varies during observation. Let  $\gamma$  represent the out-of-view ratio, calculated as  $\frac{h}{T}$ , where  $h$  is the number of out-of-view frames, and  $T$  is the total number of observation frames. To systematically analyze the performance of models across varying out-of-view ratios, we divide  $\gamma$  into discrete intervals. We define the interval  $\gamma_{(i,j)}$  as follows:  $\gamma_{(i,j)} : 0 \leq i < \gamma \leq j \leq 1$ . We compare the forecasting accuracy with the baseline and variants of our models: without the visibility loss  $\mathcal{L}_{\text{vis}}$  and the body pose loss  $\mathcal{L}_{\text{body}}$  since these loss components are effective for out-of-view cases.

As shown in Tab. 4, the proposed method consistently outperforms the baseline and the variants of our models across all out-of-view ratio intervals in both ADE and FDE. Without jointly optimizing with body poses or visibility loss, hand trajectory forecasting performance drops by an average of 1.8/1.2 cm regardless of the out-of-view ratio, respectively. These results suggest that both jointly optimizing body pose and visibility awareness are crucial for enhancing hand forecasting accuracy.

**Body Pose Estimation/Forecasting.** To assist 3D hand forecasting, our method predicts body pose jointly. Here we compare our body pose estimation with other relevant body estimation methods during both observation and forecasting. As shown in Tab. 5, our proposed model surpasses the baseline. MPJPE is improved by 3cm over the EgoEgoForecast when estimating body pose from given head poses, and by 2.4cm in forecasting future body pose. The performance of EgoAllo [76], pretrained on AMASS, suggests that the training on body motion data from the in-the-wild dataset is necessary.

**Hand Pose Estimation.** We report the hand pose estimation performance during observation<sup>4</sup> in Tab. 6. We compare with EgoEgoForecast as well as the baseline used as the hand pose estimation task in the Ego-Exo4D paper: THORnet [1] and POTTER [82]. Note that these methods only optimize for observed frames, and thus a drop in performance of EgoH4 when attempting forecasting would be expected. EgoH4 outperforms EgoEgoForecast by a large margin.

#### 4.6. Qualitative Results

The qualitative results for hand trajectory forecasting on the Ego-Exo4D dataset in different scenarios, in- or out-of-view, are visualized in Fig. 3. The in-view results demonstrate that EgoH4 accurately forecasts the hand locations compared to the baseline as we leverage visible cues as conditions. In out-of-view scenarios, our model effectively predicts future hand trajectories. In Fig. 3, the failure case

<sup>4</sup>We report the wrist-relative MPJPE for in-view sequences, following Ego-Exo4D evaluation.

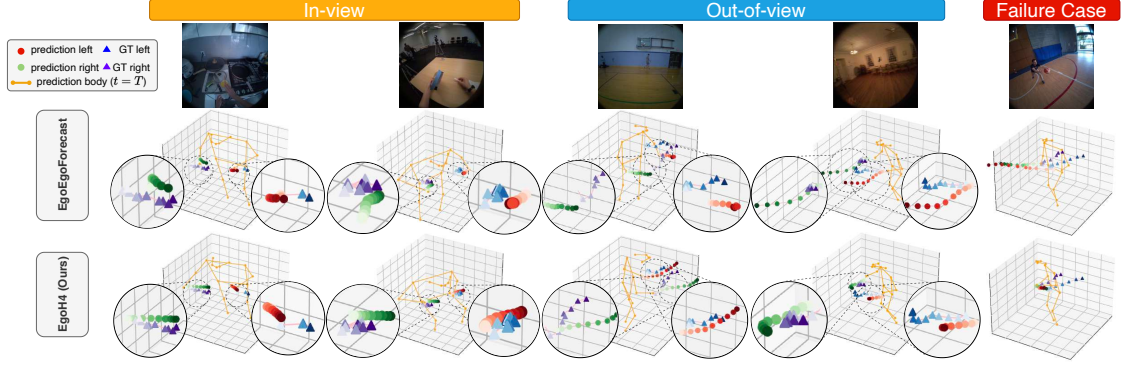


Figure 3. **Qualitative results for hand trajectory forecasting.** We show sample qualitative results compared to our best-performing baseline across activities: cooking, covid testing, basketball, and dance exercises. Dots in red, green, blue, purple, and orange represent the prediction of left/right future hands, ground-truth of left/right hands, and the prediction of body joints at the last observable frame, respectively. For each track, darker colors indicate later times.

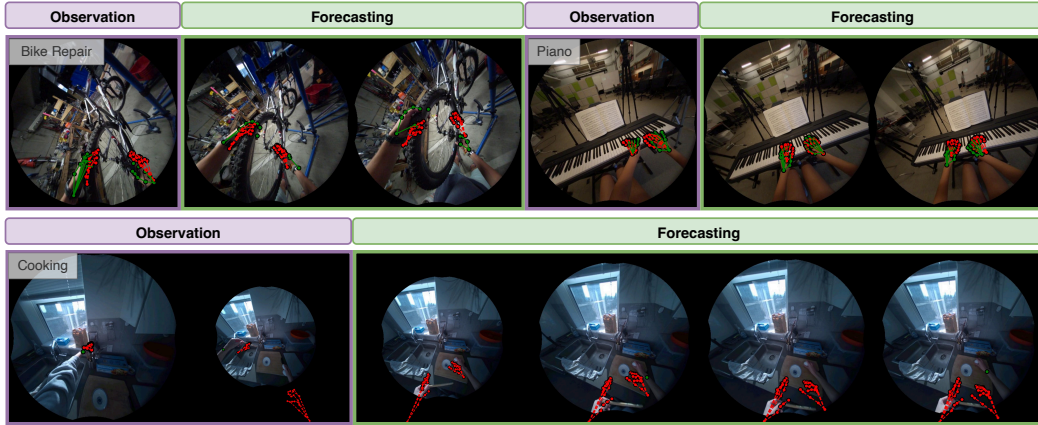


Figure 4. **Qualitative results for hand pose forecasting.** We show two in-view cases, a bike repair and piano playing scene, in the first row and one out-of-view case, a cooking scene, in the second row. Dots in red and green denote the prediction and ground truth, respectively. Note that we expand the image plane so that we can also show the out-of-view hands.

shows the ground truth body pose moving backward, while the body poses are forecasted to stay around. This reveals a limitation: significant errors in body pose forecasting can negatively impact hand forecasting accuracy.

We visualize hand pose forecasting results, reprojected on the 2D image, in Fig. 4. The first row demonstrates our method’s ability to estimate and forecast the hand joints correctly. The second row shows an interesting case where the right hand comes into view during observation. The right hand is forecasted to take a bowl and manipulate it with two hands. This shows the capability of our approach to deal with out-of-view scenarios by leveraging the body motion.

## 5. Conclusion

**Conclusion.** We are the first to explore egocentric 3D hand position and pose forecasting even when hands are partially or completely invisible during observation. We propose EgoH4, a diffusion-based transformer model that denoises the body and hand joints, given head pose, 2D hand locations (if visible), and image features. Leverag-

ing knowledge of body motion enhances our method of estimating/forecasting the invisible hand and improves hand forecasting accuracy. Moreover, we employ the 3D-to-2D reprojection loss for prediction consistency and visibility loss to acquire out-of-view awareness. We evaluate our proposed method on the Ego-Exo4D dataset, showing significantly improved forecasting accuracy for the in- and out-of-view sequences.

**Limitations and Future Work.** When the hands are invisible, our model primarily relies on body joint information to estimate and forecast the 3D hand position. This can be ambiguous – with the same input, the hands can be in different positions - e.g., by the side of the body or behind one’s back. While the diffusion model is capable of generating both options, performance degrades when the distribution of body movements during evaluation differs from those of training. This is a known limitation in evaluating forecasting, and we leave its exploration to future work.



## Acknowledgements

M Hatano is supported by JSPS Overseas Challenge Program for Young Researchers, JST BOOST, Japan Grant Number JPMJBS2409, and Amano Institute of Technology. Z Zhu is supported by University of Bristol - Chinese Scholarship Council studentship. D Damen is supported by EPSRC Fellowship UMPIRE (EP/T004991/1).

## References

- [1] Ahmed Tawfik Aboukhadra, Jameel Malik, Ahmed Elhayek, Nadia Robertini, and Didier Stricker. Thor-net: End-to-end graformer-based realistic two hands and object reconstruction with self-supervision. In *WACV*, 2023. 6, 7
- [2] 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 *ECCV*, 2022. 1, 3
- [3] Wentao Bao, Lele Chen, Libing Zeng, Zhong Li, Yi Xu, Junsong Yuan, and Yu Kong. Uncertainty-aware state space transformer for egocentric 3d hand trajectory forecasting. In *ICCV*, 2023. 1, 2, 5, 6
- [4] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *ECCV*, 2020. 2
- [5] Angela Castillo, Maria Escobar, Guillaume Jeanneret, Albert Pumarola, Pablo Arbeláez, Ali Thabet, and Arsiom Sanakoyeu. Bodiffusion: Diffusing sparse observations for full-body human motion synthesis. In *ICCV*, 2023. 1, 3
- [6] Ling-Hao Chen, Jiawei Zhang, Yewen Li, Yiren Pang, Xiaobo Xia, and Tongliang Liu. Humanmac: Masked motion completion for human motion prediction. In *ICCV*, 2023. 2
- [7] Tianyi Cheng, Dandan Shan, Ayda Hassen, Richard Higgins, and David Fouhey. Towards a richer 2d understanding of hands at scale. In *NeurIPS*, 2023. 2
- [8] Seunggeun Chi, Pin-Hao Huang, Enna Sachdeva, Hengbo Ma, Karthik Ramani, and Kwonjoon Lee. Estimating egobody pose from doubly sparse egocentric video data. In *NeurIPS*, 2024. 3
- [9] Rohan Choudhury, Kris M Kitani, and László A Jeni. Tempo: Efficient multi-view pose estimation, tracking, and forecasting. In *ICCV*, 2023. 1
- [10] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. <https://github.com/open-mmlab/mmpose>, 2020. 2
- [11] Hanz Cuevas-Velasquez, Charlie Hewitt, Sadegh Aliakbarian, and Tadas Baltrušaitis. Simpleego: Predicting probabilistic body pose from egocentric cameras. In *3DV*, 2024. 1
- [12] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. In *NeurIPS*, 2022. 2
- [13] Christian Diller, Thomas Funkhouser, and Angela Dai. Futurehuman3d: Forecasting complex long-term 3d human behavior from video observations. In *CVPR*, 2024. 2
- [14] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David Crandall. Hope-net: A graph-based model for hand-object pose estimation. In *CVPR*, 2020. 2
- [15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 4
- [16] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brigid Meredith, Cheng Peng, Chris Sweeney, Cole Wilson, and et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. 3
- [17] Maria Escobar, Juanita Puentes, Cristhian Forigua, Jordi Pont-Tuset, Kevis-Kokitsi Maninis, and Pablo Arbeláez. Egocast: Forecasting egocentric human pose in the wild. In *WACV*, 2025. 1, 3
- [18] Zicong Fan, Takehiko Ohkawa, Linlin Yang, Nie Lin, Zhihan Zhou, Shihao Zhou, Jiajun Liang, Zhong Gao, Xuanyang Zhang, Xue Zhang, et al. Benchmarks and challenges in pose estimation for egocentric hand interactions with objects. In *ECCV*, 2024. 2
- [19] Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *ICCV*, 2019. 1
- [20] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *CVPR*, 2018. 2
- [21] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *ICCV*, 2021. 1
- [22] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 1
- [23] Gabriele Goletto, Tushar Nagarajan, Giuseppe Averta, and Dima Damen. Amego: Active memory from long egocentric videos. In *ECCV*, 2024. 2
- [24] Mohit Goyal, Sahil Modi, Rishabh Goyal, and Saurabh Gupta. Human hands as probes for interactive object understanding. In *CVPR*, 2022. 2
- [25] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abraham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kotur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will

- Price, Paola Ruiz, Merey Ramazanov, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022. 5
- [26] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, and et al. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *CVPR*, 2024. 2, 5, 6
- [27] Adam W. Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle video revisited: Tracking through occlusions using point trajectories. In *ECCV*, 2022. 4
- [28] Masashi Hatano, Ryo Hachiuma, Ryo Fujii, and Hideo Saito. Multimodal cross-domain few-shot learning for egocentric action recognition. In *ECCV*, 2024. 1
- [29] Masashi Hatano, Ryo Hachiuma, and Hideo Saito. Emag: Ego-motion aware and generalizable 2d hand forecasting from egocentric videos. In *ECCV Workshop*, 2024. 1, 2, 5
- [30] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 4
- [31] Zhiming Hu, Zheming Yin, Daniel Haeufle, Syn Schmitt, and Andreas Bulling. Hoimotion: Forecasting human motion during human-object interactions using egocentric 3d object bounding boxes. *IEEE Transactions on Visualization and Computer Graphics*, 30(11):7375–7385, 2024. 2
- [32] Wenqi Jia, Miao Liu, and James M. Rehg. Generative adversarial network for future hand segmentation from egocentric video. In *ECCV*, 2022. 2
- [33] Jiayi Jiang, Paul Strel, Manuel Meier, and Christian Holz. Egoposer: Robust real-time egocentric pose estimation from sparse and intermittent observations everywhere. In *ECCV*, 2024. 3
- [34] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Co-tracker: It is better to track together. In *ECCV*, 2024. 4
- [35] Rawal Khrodar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, and Kris Kitani. Ego-humans: An egocentric 3d multi-human benchmark. In *ICCV*, 2023. 1
- [36] Taein Kwon, Bugra Tekin, Jan Stuhmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *ICCV*, 2021. 5
- [37] Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *CVPR*, 2023. 1, 3, 4, 6
- [38] Maosen Li, Siheng Chen, Zijing Zhang, Lingxi Xie, Qi Tian, and Ya Zhang. Skeleton-parted graph scattering networks for 3d human motion prediction. In *ECCV*, 2022. 2
- [39] Yiming Li, Ziang Cao, Andrew Liang, Benjamin Liang, Luoyao Chen, Hang Zhao, and Chen Feng. Egocentric prediction of action target in 3d. In *CVPR*, 2022. 5
- [40] Miao Liu, Siyu Tang, Yin Li, and James M Rehg. Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In *ECCV*, 2020. 1, 2
- [41] Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *CVPR*, 2022. 1, 2
- [42] Zhengyi Luo, Ryo Hachiuma, Ye Yuan, and Kris Kitani. Dynamics-regulated kinematic policy for egocentric pose estimation. In *NeurIPS*, 2021. 3, 4
- [43] Junyi Ma, Xieyuanli Chen, Wentao Bao, Jingyi Xu, and Hesheng Wang. Madiff: Motion-aware mamba diffusion models for hand trajectory prediction on egocentric videos. *arXiv preprint arXiv:2409.02638*, 2024. 1, 2
- [44] Junyi Ma, Jingyi Xu, Xieyuanli Chen, and Hesheng Wang. Diff-ip2d: Diffusion-based hand-object interaction prediction on egocentric videos. *arXiv preprint arXiv:2405.04370*, 2024. 1, 2
- [45] Tiezheng Ma, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In *CVPR*, 2022. 2
- [46] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *ICCV*, 2019. 6
- [47] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. Generating smooth pose sequences for diverse human motion prediction. In *ICCV*, 2021. 2
- [48] Lea Müller, Vickie Ye, Georgios Pavlakos, Michael Black, and Angjoo Kanazawa. Generative proxemics: A prior for 3d social interaction from images. In *CVPR*, 2024. 1
- [49] Tushar Nagarajan, Santhosh Kumar Ramakrishnan, Ruta Desai, James Hillis, and Kristen Grauman. Egoenv: Human-centric environment representations from egocentric video. In *NeurIPS*, 2023. 2
- [50] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. You2me: Inferring body pose in egocentric video via first and second person interactions. In *CVPR*, 2020. 3
- [51] Yeounguk Oh, JoonKyu Park, Jaeha Kim, Gyeongsik Moon, and Kyoung Mu Lee. Recovering 3d hand mesh sequence from a single blurry image: A new dataset and temporal unfolding. In *CVPR*, 2023. 2
- [52] JoonKyu Park, Yeounguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handocnet: Occlusion-robust 3d hand mesh estimation network. In *CVPR*, 2022.
- [53] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024. 2
- [54] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012. 1
- [55] Chiara Plizzari, Toby Perrett, Barbara Caputo, and Dima Damen. What can a cook in Italy teach a mechanic in India? action recognition generalisation over scenarios and locations. In *ICCV*, 2023.

- [56] Yair Poleg, Ariel Ephrat, Shmuel Peleg, and Chetan Arora. Compact cnn for indexing egocentric videos. In *WACV*, 2016. 1
- [57] Aditya Prakash, Ruisen Tu, Matthew Chang, and Saurabh Gupta. 3d hand pose estimation in everyday egocentric images. In *ECCV*, 2024. 2
- [58] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *ICCV*, 2021. 1
- [59] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *ICCV Workshop*, 2021. 2
- [60] Debaditya Roy, Ramanathan Rajendiran, and Basura Fernando. Interaction region visual transformer for egocentric action anticipation. In *WACV*, 2024. 1
- [61] Christoph Schöller, Vincent Aravatinos, Florian Lay, and Alois Knoll. What the constant velocity model can teach us about pedestrian motion prediction. *IEEE Robotics and Automation Letters (RA-L)*, 5(2):1696–1703, 2020. 6
- [62] F. Sener, D. Chatterjee, D. Shelepov, K. He, D. Singhania, R. Wang, and A. Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *CVPR*, 2022. 2
- [63] Dandan Shan, Jiaqi Geng, Michelle Shu, and David Fouhey. Understanding human hands in contact at internet scale. In *CVPR*, 2020. 2
- [64] Bowen Tang, Kaihao Zhang, Wenhan Luo, Wei Liu, and Hongdong Li. Prompting future driven diffusion model for hand motion prediction. In *ECCV*, 2024. 1, 2
- [65] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+o: Unified egocentric recognition of 3d hand-object poses and interactions. In *CVPR*, 2019. 1
- [66] Sanket Thakur, Cigdem Beyan, Pietro Morerio, Vittorio Murino, and Alessio Del Bue. Leveraging next-active objects for context-aware anticipation in egocentric videos. In *WACV*, 2024. 1
- [67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 4
- [68] Jian Wang, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, and Christian Theobalt. Estimating egocentric 3d human pose in global space. In *ICCV*, 2021. 1, 3
- [69] Jian Wang, Zhe Cao, Diogo Luvizon, Lingjie Liu, Kripasindhu Sarkar, Danhang Tang, Thabo Beeler, and Christian Theobalt. Egocentric whole-body motion capture with fisheye and diffusion-based motion refinement. In *CVPR*, 2024. 3
- [70] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *ICCV*, 2023. 2
- [71] Yu Wu, Linchao Zhu, Xiaohan Wang, Yi Yang, and Fei Wu. Learning to anticipate egocentric actions by imagination. *IEEE Transactions on Image Processing*, 30:1143–1152, 2020. 1
- [72] Haitao Yan, Qiongjie Cui, Jiexin Xie, and Shijie Guo. Forecasting of 3d whole-body human poses with grasping objects. In *CVPR*, 2024. 2
- [73] Ji Yang, Youdong Ma, Xinxin Zuo, Sen Wang, Minglun Gong, and Li Cheng. 3d pose estimation and future motion prediction from 2d images. *Pattern Recognition*, 124:108439, 2022. 2
- [74] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *CVPR*, 2023. 1
- [75] Yufei Ye, Xueting Li, Abhinav Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance diffusion: Synthesizing hand-object interactions. In *CVPR*, 2023. 2
- [76] Brent Yi, Vickie Ye, Maya Zheng, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. Estimating body and hand motion in an ego-sensed world. *arXiv preprint arXiv:2410.03665*, 2024. 3, 4, 6, 7
- [77] Ye Yuan and Kris Kitani. 3d ego-pose estimation via imitation learning. In *ECCV*, 2018. 1
- [78] Ye Yuan and Kris Kitani. Ego-pose estimation and forecasting as real-time pd control. In *ICCV*, 2019. 1, 3, 4
- [79] Lingzhi Zhang, Shenghao Zhou, Simon Stent, and Jianbo Shi. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *ECCV*, 2022. 2
- [80] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Ego-body: Human body shape and motion of interacting people from head-mounted devices. In *ECCV*, 2022. 3
- [81] Siwei Zhang, Qianli Ma, Yan Zhang, Sadegh Aliakbarian, Darren Cosker, and Siyu Tang. Probabilistic human mesh recovery in 3d scenes from egocentric views. In *ICCV*, 2023. 1
- [82] Ce Zheng, Xianpeng Liu, Guo-Jun Qi, and Chen Chen. Potter: Pooling attention transformer for efficient human mesh recovery. In *CVPR*, 2023. 6, 7
- [83] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 4