

EgoCast: Forecasting Egocentric Human Pose in the Wild

–Supplementary material–

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

1. Video

We highly recommend viewing the accompanying video in the supplementary material to further appreciate the predictions generated by our method. This video showcases the effectiveness of our approach in accurately forecasting poses and generating trajectories with high plausibility.

2. Ego-Exo4D

2.1. Dataset statistics

Ego-Exo4D dataset [1] encompasses seven activities: Soccer, Basketball, Cooking, Dance, Bike Repair, Music and Health. Each set of activities provides a rich set of human motion patterns and poses, which allows one to understand 3D human motion in real-case scenarios, making Ego-Exo4D [1] a rich and realistic framework for 3D human pose forecasting. Table 1 shows the statistics for the EgoExo4D dataset portion compatible with our EgoCast framework.

The license for using the Ego-Exo4D dataset can be found here <https://ego4d-data.org/pdfs/Ego-Exo4D-Model-License.pdf>.

2.2. Implementation details

For the current-frame estimation module, we use Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 24. We train this module for 2×10^5 iterations. Then, we perform a second-stage finetuning that incorporates the visual stream. We use EgoVLP [2] as a visual encoder and start from pretrained weights. Then, we finetune for an additional 5×10^4 iterations. Similarly, we train the forecasting branch for 3×10^4 iterations. The past time window k is set to 20 frames for both current-pose estimation and forecasting. For our experimental setup, we utilized the standard train/test splits as proposed by EgoExo-4D [1]. We

Activity	Takes	Annotated frames	Seconds	Keypoints
Soccer	160	596,109	20,170	9,232,969
Basketball	612	1,005,790	34,044	16,744,792
Cooking	334	2,307,796	82,230	32,512,567
Dance	581	790,148	27,720	13,134,050
Bike Repair	275	1,163,466	42,320	17,558,150
Music	170	222,179	10,856	3,186,588
Health	240	1,284,327	45,515	15,664,947
Total	2372	7,369,815	262,855	108,034,063

Table 1. **Ego-Exo4d [1] 3D human pose statistics.** Summary of the statistics for the annotated 3D human poses of the Ego-Exo4D [1] dataset. Overall, the dataset includes more than 100 million annotated keypoints distributed across 2372 takes.

train our model using PyTorch on 2 NVIDIA Quadro RTX 8000 GPUs over four hours.

2.3. Architecture details

In our transformer-based forecasting module, we begin by applying a linear embedding map the 70 input features into a 256-dimensional space. Subsequently, the data is processed through a transformer composed of three self-attention layers, each with eight heads. To maintain a consistent output dimensionality, we use adaptive average pooling, ensuring an ending dimension size of 256, regardless of the input length. The process concludes with another linear embedding to obtain the final forecasting output.

2.4. Ablation experiments

Table 2 shows the performance of EgoCast for 1 second forecasting under different past window sizes. We find that optimizing the past window size has a notable impact, with a window size of 20 frames achieving the lowest error (19.66) in comparison to using fewer or more frames. Thus, 20 frames is the sweet spot for getting the most accurate predictions. These results suggest that having just the right amount

Window size	5	10	20	40	80
MPJPE	29.34	26.57	19.66	32.80	33.11

Table 2. **Forecasting ablation experiments.** We evaluate the impact of the past window size on forecasting 1 second in the future (30 frames). We achieved the best performance by using a window size of 20 frames.

of past information is key; insufficient data leads to inadequate forecasting, while excessive information potentially deteriorates the model’s effectiveness, possibly due to the introduction of contradictory context.

3. Aria Digital Twin

We present results in the Aria Digital Twin (ADT) [3] dataset to showcase the versatility and generalization capacity of EgoCast.

3.1. Dataset description

The Aria Digital Twin (ADT) [3] dataset includes densely annotated sequences with ground-truth 3D body poses and the measurements collected by the Aria devices [4]: egocentric video at 30 frames per second, and position and rotation of the camera at every frame.

Table 3 shows the statistics for the ADT dataset portion compatible with our EgoCast framework. Overall, the dataset contains 211,824 frames with 3D human pose annotations of 21 joints distributed over 74 captures, with an average length of 2876 frames (96s) per capture. Furthermore, ADT includes diverse activities showcasing a vast variety of human motions; the most common activity is partying, followed by working, decorating, and having a meal. We divide the ADT dataset into a training and testing set for our experiments, using a 50:50 proportion. We ensure a uniform distribution of activities for both folds. ADT presents a high diversity of trajectories and motion of the camera wearer for each activity in the dataset. While for some activities, such as meal and work, trajectories are densely located in certain rooms of the scenario (*e.g.*, kitchen and dinner room), other activities like decoration are distributed across the scene. Additionally, each set of activities provides a rich set of human motion patterns and poses, which allows one to understand 3D human motion in real-case scenarios, making the ADT dataset a rich and realistic framework for studying 3D human pose forecasting.

The licence for using the ADT dataset can be found here <https://www.projectaria.com/datasets/adtl/advertise/>.

Activity	Captures	Annotated frames	Seconds	Keypoints
Party	40	118,398	3947	2,486,358
Work	15	41,332	1378	867,972
Decoration	10	27,497	917	577,437
Meal	9	24,597	820	516,537
Total	74	211,824	7061	4,448,304

Table 3. **ADT 3D human pose statistics.** Summary of the statistics for the annotated 3D human poses of the ADT dataset. Overall, the dataset includes more than 4 million annotated keypoints distributed across 74 captures.

Window size	Visual	MPJPE
5	✗	9.76
10	✗	9.57
20	✗	9.27
40	✗	9.36
20	✓	7.43

Table 4. **Ablation experiments for the current-frame estimation module.** We perform ablation experiments that include the rotation of the head-mounted device, the past window size, and the impact of using the visual cues. We achieved the best performance by using proprioception inputs (rotation and translation), a window size of 20 frames, and integrating visual streams for prediction.

3.2. Ablation experiments

Table 4 shows the performance of the current-frame estimation module in the EgoCast baseline. This analysis shows that optimizing the past window size has a notable impact, with a window size of 20 frames achieving the lowest error (9.27) without visual cues. The most substantial improvement is observed when integrating visual streams, which reduces the MPJPE to 7.43, underscoring the importance of combining proprioceptive inputs and visual data for enhanced pose estimation accuracy.

3.3. Results

The per-category analysis in Figure 1 show that most activities within the ADT dataset have similar difficulty with the exception of decorating. The added difficulty in decorating comes from having large displacements throughout the environment. Note that our EgoCast approach is able to predict plausible poses and trajectory movements.

References

- [1] 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.

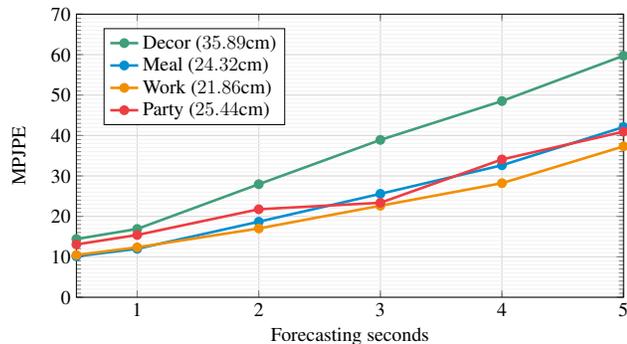


Figure 1. **ADT Forecasting Performance by Category** For each activity in the ADT dataset, we show the performance of our method when forecasting $\{0.5, 1, 2, 3, 4, \text{ and } 5\}$ seconds into future. Note that since the graph shows MPJPE, lower curves represent better performance.

- [2] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z XU, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pre-training. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022.
- [3] 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.
- [4] 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.