




EMAG: Ego-motion Aware and Generalizable 2D Hand Forecasting from Egocentric Videos

– Supplementary Materials –

Masashi Hatano¹, Ryo Hachiuma², and Hideo Saito¹

¹ Keio University
² NVIDIA

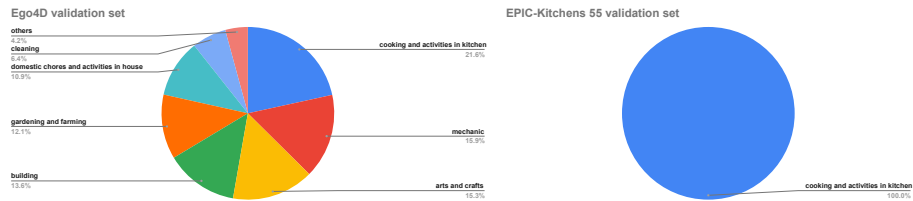


Fig. 5: Scenario breakdown. The left pie chart represents the scenario breakdown on the validation set of the Ego4D dataset. There are eight categories in total, including inside/outside scenes. The right pie chart represents the scenario breakdown on the validation set of the EPIC-Kitchens 55 dataset. The EPIC-Kitchens 55 dataset contains only one category, cooking and activities in the kitchen.

A Dataset

A.1 Statistics

This section provides statistics on two large-scale egocentric video datasets, Ego4D [2] and EPIC-Kitchens 55 [1]. Fig. 5 presents pie charts illustrating the proportional distribution, categorized by action types or situations, of camera wearers within each validation set of the dataset. The categories are summarized as follows:

- **Cooking and activities in kitchen** contains videos where the camera wearer performs tasks in the kitchen, such as cutting vegetables, washing a pan, and putting dishes away on the shelf.
- **Mechanic** contains situations where the camera wearer uses specific mechanical tools to repair vehicles such as cars or bikes.
- **Arts and crafts** consist of indoor and outdoor scenarios, including activities such as painting and trimming excess materials.
- **Building** category contains a construction scene and a scene depicting brick fabrication.
- **Gardening and farming** consist of both small-scale and large-scale plant caring scenes.

Table 6: Input modality ablation study. Ablation study on the input modalities on Ego4D and EPIC-Kitchens 55. We evaluate the model in the intra and cross-dataset settings to verify the contribution of each input modality to the hand forecasting performance and the generalizability against novel scenes. In the last two rows, we summarize the results of two scenarios, intra and cross-dataset. The last column is the result of the proposed method, which uses all the modal information.

Object RGB Flow Ego				Ego4D → Ego4D		EPIC → Ego4D		EPIC → EPIC		Ego4D → EPIC		Intra		Cross	
				ADE ↓	FDE ↓	ADE ↓	FDE ↓	ADE ↓	FDE ↓	ADE ↓	FDE ↓	ADE ↓	FDE ↓	ADE ↓	FDE ↓
	✓	✓	✓	49.02	53.00	54.25	56.79	48.50	54.57	51.31	57.25	48.76	53.79	52.78	57.02
✓		✓	✓	51.02	54.30	54.09	57.15	49.14	55.35	52.90	57.93	50.08	54.83	53.30	57.54
✓	✓		✓	50.82	53.77	55.57	57.70	51.17	55.78	53.90	58.15	51.00	54.78	54.74	57.93
✓	✓	✓		49.04	52.69	54.22	57.01	47.66	53.79	51.55	57.02	48.35	53.24	52.89	57.02
✓	✓	✓	✓	48.99	52.83	53.67	56.36	48.78	54.03	51.03	56.78	48.89	53.43	52.35	56.57

- **Domestic chores and activities in house** contain activities in the house except for the situation in the kitchen, such as laundering, knitting, ironing, and playing cards.
- **Cleaning** category contains cleaning activities such as sweeping with a broom, mopping the floor, and washing a car.
- **Others** consist of various scenarios such as sports (playing basketball or working out at the gym), driving, walking a dog, and activities in the laboratory.

While all videos in the EPIC-Kitchens 55 dataset are categorized as cooking and activities in the kitchen, the Ego4D dataset contains various categories described above. More than three-quarters of the videos in the validation set of Ego4D are composed of cooking and activities in the kitchen (21.6%), mechanic (15.9%), arts/crafts (15.3%), building (13.6%), and gardening/farming (12.1%).

B Further Results

B.1 Input Modality Ablation

Tab. 6 shows all four intra/cross-dataset scenarios using two datasets, trained and evaluated on either Ego4D or EPIC-Kitchens 55, and the aggregated results for intra/cross-dataset scenarios.

Analysis. As shown in Tab. 6, our proposed method is outperformed by the model that omits object or ego-motion information in the scenario, where models are trained and tested on EPIC-Kitchens 55. This occurs due to the overfit to the context of the cooking category. Methods lacking object or ego-motion information tend to rely more on RGB information to predict future hand positions than the proposed method that leverages all modalities.

Generalizability of each input modality. We further analyze the generalizability of each input modality: the trajectory of bounding boxes of objects, RGB, optical flow, and ego-motion information. Fig. 6 shows the drop in performances

Table 7: Loss component ablation study. Ablation study on ego-motion estimation loss on two datasets in intra and cross-dataset scenarios to verify the effectiveness of estimating future ego-motion as an auxiliary task.

Method	Ego4D \rightarrow Ego4D		EPIC \rightarrow Ego4D		EPIC \rightarrow EPIC		Ego4D \rightarrow EPIC	
	ADE \downarrow	FDE \downarrow	ADE \downarrow	FDE \downarrow	ADE \downarrow	FDE \downarrow	ADE \downarrow	FDE \downarrow
w/o \mathcal{L}_{ego}	49.59	53.15	53.85	56.57	49.72	55.37	51.83	57.59
w/ \mathcal{L}_{ego} (Ours)	48.99	52.83	53.67	56.36	48.78	54.03	51.03	56.78

for each model that is missing one of the four input modalities, from the intra-dataset scenario to the cross-dataset scenario in the average of two datasets. The smaller the performance drop is, the more the leveraged modalities (the other three modalities other than the lacking modality) contribute to the generalizability against unseen data. The performance drops of the method without object, RGB, optical flow, and ego-motion, are 8.24%, 6.43%, 7.33%, and 9.39%, respectively. This confirms that RGB is the most susceptible to unseen data, as RGB depends on appearance, which leads to the overfit to backgrounds or the contexts, and the ego-motion information (homography) is the most generalizable input modality among the four modalities against novel scenes.

Average of two datasets

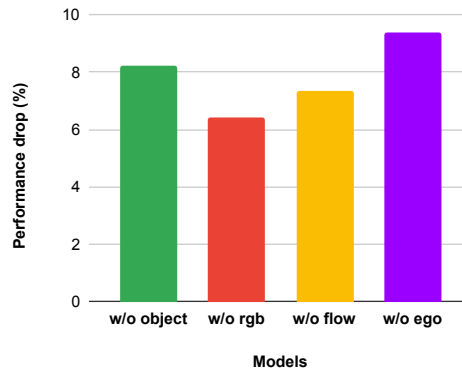


Fig. 6: The performance drop of each model that lacks one of the input modalities. The green, red, yellow, and purple bar charts represent models without objects, RGB, optical flow, and ego-motion information.

B.2 Loss Component Ablation

Tab. 7 shows the hand forecasting performance of whether adopting the ego-motion estimation loss \mathcal{L}_{ego} in all four intra/cross-dataset scenarios. The method without using \mathcal{L}_{ego} deteriorates the performance in all intra/cross-dataset scenarios, verifying the effectiveness of estimating future ego-motion as an auxiliary task for both intra and cross-dataset settings.

B.3 Ego-motion Representation

We conducted an additional ablation study on ego-motion representation, considering the homography matrix and background optical flow (Tab. 8). The proposed homography matrix representation outperformed the background optical

Table 8: Ablation study of ego-motion representation.

Method	Ego4D→EPIC	
	ADE ↓	FDE ↓
Background flow	52.08	58.03
Ours	51.03	56.78

flow representation in cross-scenarios, underscoring the effectiveness of the proposed ego-motion representation.

References

1. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling egocentric vision: The epic-kitchens dataset. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
2. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., Martin, M., Nagarajan, T., Radosavovic, I., Ramakrishnan, S.K., Ryan, F., Sharma, J., Wray, M., Xu, M., Xu, E.Z., Zhao, C., Bansal, S., Batra, D., Cartillier, V., Crane, S., Do, T., Doulaty, M., Erapalli, A., Feichtenhofer, C., Fragomeni, A., Fu, Q., Gebreselasie, A., González, C., Hillis, J., Huang, X., Huang, Y., Jia, W., Khoo, W., Kolář, J., Kottur, S., Kumar, A., Landini, F., Li, C., Li, Y., Li, Z., Mangalam, K., Modhugu, R., Munro, J., Murrell, T., Nishiyasu, T., Price, W., Ruiz, P., Ramazanov, M., Sari, L., Somasundaram, K., Southerland, A., Sugano, Y., Tao, R., Vo, M., Wang, Y., Wu, X., Yagi, T., Zhao, Z., Zhu, Y., Arbeláez, P., Crandall, D., Damen, D., Farinella, G.M., Fuegen, C., Ghanem, B., Ithapu, V.K., Jawahar, C.V., Joo, H., Kitani, K., Li, H., Newcombe, R., Oliva, A., Park, H.S., Rehg, J.M., Sato, Y., Shi, J., Shou, M.Z., Torralba, A., Torresani, L., Yan, M., Malik, J.: Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)