Supplementary Material: Interaction Region Visual Transformer for Egocentric Action Anticipation

Debaditya Roy[1], Ramanathan Rajendiran[1], and Basura Fernando[1,2]

[1]Institute of High-Performance Computing, Agency for Science, Technology and Research, Singapore
[2]Centre for Frontier AI Research, Agency for Science, Technology and Research, Singapore

1. Cross-Attention and Self-Attention

In cross-attention, queries are obtained from the target that needs to be refined. The key and values are obtained from the source that needs to be queried to obtain the refined target tokens. Let $\mathbf{a}_m \in A$ be a target token from all target tokens A and $\mathbf{b}_n \in B$ be a source token from all the source tokens B. The queries, keys, and values are constructed as

$$\mathbf{q} = \mathbf{a}_m \mathbf{W}_q, \mathbf{k} = \mathbf{b}_n \mathbf{W}_k, \mathbf{v} = \mathbf{b}_n \mathbf{W}_v \tag{1}$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v$ are learnable weights. In the rest of the paper, we use $\mathbf{W}_q,\ \mathbf{W}_k,$ and $\mathbf{W}_v$, as learnable weights for query, key, and value, respectively, for all attention operations. We do this for clarity in notation but all the weights are different and learned independently. In self-attention, query, key, and value are all obtained from target tokens $\mathbf{a}_m \in A$.

$$\mathbf{q} = \mathbf{a}_m \mathbf{W}_q, \mathbf{k} = \mathbf{a}_m \mathbf{W}_k, \mathbf{v} = \mathbf{a}_m \mathbf{W}_v \tag{2}$$

The output of both cross-attention and self-attention is the refined target token $\tilde{\mathbf{a}}_m$.

$$\tilde{\mathbf{a}}_m = \sum \mathbf{v} \frac{\exp \langle \mathbf{q}, \mathbf{k} \rangle}{\sum \exp \langle \mathbf{q}, \mathbf{k} \rangle} \tag{3}$$

We omit the factor $d^{1/2}$ for clarity and assume that both queries and keys have been scaled by $d^{1/4}$ [6].

2. Does failed hand detection affect performance?

No, this is not an issue. The below Table 1 shows the number of frames with no hands is less than 0.5% of all frames. Hence, the influence of missing hands is negligible. We use the same hand-object detector used by the EK100 dataset. We use a threshold of 0.05 instead of 0.1 to obtain more hand detections. Object occlusion (50% or more) due to hands happens only in 8.6% of the frames in EK100. In Fig. 1, we show that reducing the threshold for hand detection results results in noisy hand detections but InAViT is still able to predict method is able to predict the future action correctly.

Split	# of hands present per frame(%)					
	EK100			EGTEA		
	0	1	2	0	1	2
train	0.18	05.20	94.62	0.11	03.67	96.21
val	0.41	03.77	95.82	0.01	04.01	95.98
test	0.40	03.69	95.91	0.05	02.13	97.82

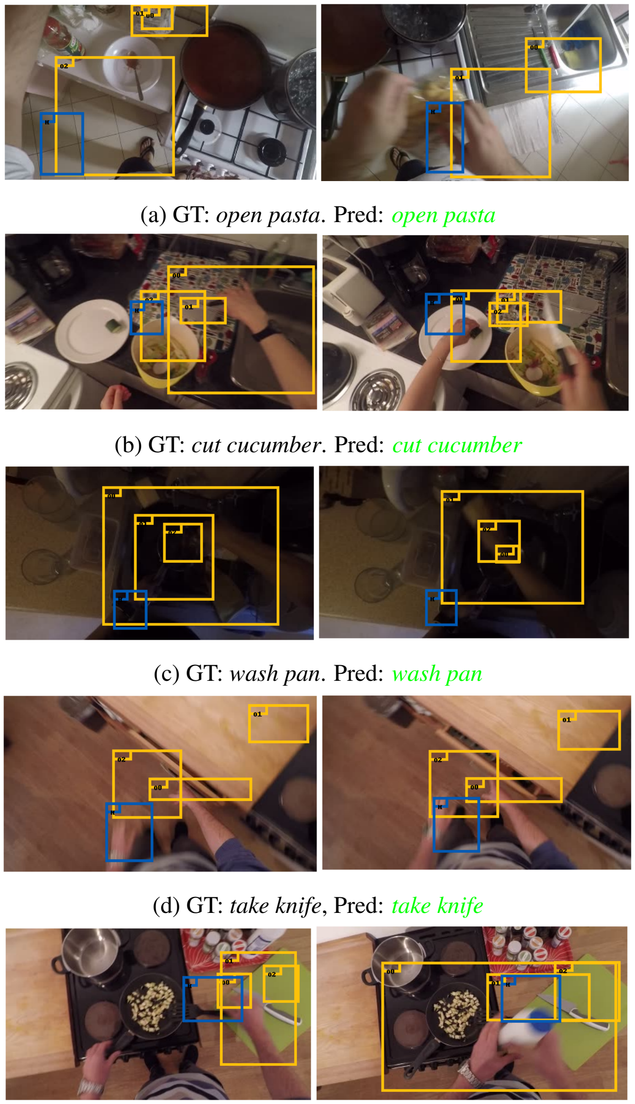
Table 1: Comparing the number of hands present per frame in train, test, and val splits of EK100 and EGTEA.

3. Changing number of objects per frame

We compare the performance of varying the number of objects per frame N when training InAViT. For EK100 (Tab. 2), the performance improves when we increase the number of objects per frame to 4 but then deteriorates when we further increase it to 5. So, we use N=4 in all our experiments on EK100 in the main paper. Similarly, for EGTEA (Tab. 3), we find that InAViT performs the best when we set N=2 objects per frame.

# Objects per frame	VERB	NOUN	ACTION
1	37.62	41.56	20.67
2	38.54	42.44	21.89
3	40.14	43.66	23.65
4	**40.68**	**44.13**	**24.65**
5	40.25	44.04	24.45

Table 2: Effect of changing the number of objects per frame N on anticipation performance when training InAViT. Metric is Mean Recall@5 evaluated on EK100 validation set.



(a) GT: *open pasta*. Pred: *open pasta*

(b) GT: *cut cucumber*. Pred: *cut cucumber*

(c) GT: *wash pan*. Pred: *wash pan*

(d) GT: *take knife*, Pred: *take knife*

(e) GT: *pour salt*, Predicted: *pour salt*

Figure 1: More examples of detections that cover areas around the hand after lowering threshold of Faster RCNN to 0.05 on EK100. The hand detections (in blue) along with the objects (in yellow) cover the region around the hand that is useful for interaction. InAViT still anticipates correctly with noisy hand detection.

4. More attention qualitative results

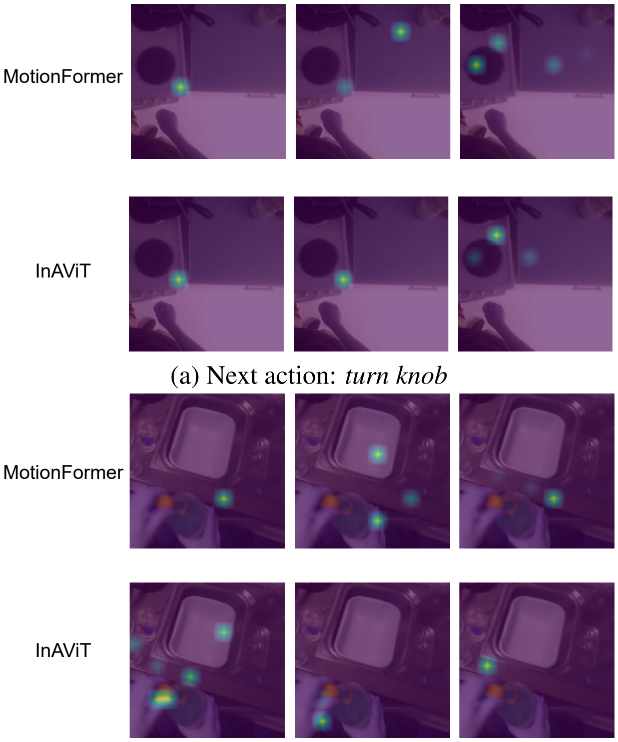
In Fig. 2, we show more qualitative results of attention outputs comparing MotionFormer and InAViT on two more actions - *turn knob* and *open bottle*. InAViT is able to focus on the important regions relevant for the next action in both cases better than MotionFormer.

# Objects per frame	VERB	NOUN	ACTION
1	78.9	75.8	65.7
2	**79.3**	**77.6**	**67.8**
3	78.5	76.0	64.8

Table 3: Changing number of objects per frame N on EGTEA dataset. Metric is Top-1 Accuracy.

Method	Top-1 Accuracy (%)			Mean Class Accuracy (%)		
	VERB	NOUN	ACT.	VERB	NOUN	ACT.
I3D-Res50 [1]	48.0	42.1	34.8	31.3	30.0	23.2
FHOI [4]	49.0	45.5	36.6	32.5	32.7	25.3
RU-LSTM [2]	50.3	48.1	38.6	-	-	-
AFFT [7]	53.4	50.4	42.5	42.4	44.5	35.2
AVT [3]	54.9	52.2	43.0	49.9	48.3	35.2
Abs. Goal [5]	64.8	65.3	49.8	63.4	55.6	37.4
MF*	77.8	75.6	66.6	77.5	72.1	56.9
ORViT-MF*	78.8	76.3	67.3	78.8	75.8	57.2
InAViT (Ours)	**79.3**	**77.6**	**67.8**	**79.2**	**76.9**	**58.2**

Table 4: Comparison of anticipation performance on EGTEA Gaze+.



(a) Next action: *turn knob*

(b) Next action: *open bottle*

Figure 2: (a) InAViT focuses near the knob when predicting the next action *turn knob*. (b) InAViT focuses on the bottle cap when predicting the next action of *open bottle*.

5. Full ablation table

In Tab. 5, we present the full ablation table as an addendum to Tab. 1 in the main paper. We include the verb and noun anticipation performance and observe similar trends as action anticipation. SCA+CI+ICV performs the best on both verb and noun anticipation.

Method	Overall (%)			Unseen (%)			Tail (%)		
	Verb	Noun	Action	Verb	Noun	Action	Verb	Noun	Action
SCA	28.06	28.01	12.66	28.73	34.51	15.49	19.14	13.79	6.03
SCA + CI	37.63	38.71	14.21	34.93	38.89	14.26	30.68	29.10	9.12
SCA+ICV	48.14	47.71	22.21	43.61	46.44	20.85	42.49	37.83	17.07
SCA + CI + TA	**49.14**	**49.97**	**23.75**	**44.36**	**49.28**	**23.49**	**43.17**	**39.91**	**18.11**
(a) Component-wise validation of InAViT									
UB+CI+ICV	45.16	47.99	22.75	42.9	48.05	22.14	38.54	36.94	17.04
SOT+CI+ICV	44.39	47.44	22.48	42.74	47.07	20.56	37.83	36.86	17.46
(b) Comparing interaction modeling methods									
SCA-(Hand)+CI+ICV	47.44	48.91	23.27	43.42	48.07	23.21	41.08	38.27	17.57
SCA-(Obj)+CI+ICV	46.03	47.75	22.49	42.44	47.44	22.23	39.76	37.04	16.73
(c) Comparing refined hand vs. object as interaction tokens									
SCA+CI(Mask FG)+ICV	29.67	25.38	8.05	24.02	23.19	5.92	24.01	16.57	5.84
SCA+ Concat +ICV	46.32	48.71	22.14	42.47	49.01	23.47	39.78	37.91	17.24

(d) Effect of context infusion

Table 5: Full ablation of InAViT including verb and noun results on Action anticipation on EK100 evaluation server. CI=Context infusion, CI(Mask FG) = Context Infusion with foreground (hands and objects) masked out, Concat = Context infusion by concatenating context tokens with interaction tokens

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2

[2] Antonino Furnari and Giovanni Maria Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4021–4036, 2020. 2

[3] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13505–13515, 2021. 2

[4] 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 *European Conference on Computer Vision*, pages 704–721. Springer, 2020. 2

[5] Debaditya Roy and Basura Fernando. Predicting the next action by modeling the abstract goal. *arXiv preprint arXiv:2209.05044*, 2022. 2

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[7] Zeyun Zhong, David Schneider, Michael Voit, Rainer Stiefelhagen, and Jürgen Beyerer. Anticipative feature fusion transformer for multi-modal action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6068–6077, 2023. 2