Leveraging Temporal Context in Low Representational Power Regimes (Supplementary Material)

Camilo L. Fosco camilolu@mit.edu
SouYoung Jin souyoung@mit.edu
Emilie Josephs ejosephs@mit.edu
Aude Oliva oliva@mit.edu
MIT CSAIL

A. Qualitative Prediction Examples

We analyze the qualitative performance of our model in action recognition and action anticipation tasks. We present both correct and incorrect detections, shedding light on failure modes. Figure S1 shows some qualitative examples from our best MoViNet A0, trained on EK100 with a 2562×2562 exponentially decayed ETM. We observe that our model correctly estimates nouns and verbs, and fails in some cases where the action portrayed is ambiguous.

Figure S2 shows qualitative prediction performance on action anticipation. Our model manages to correctly estimate the imminence of actions like "wipe sink" after "wipe counter", and struggles at inferring the end of an action that typically spans multiple timesteps like "roll dough".

Figure S3 sheds light on a few examples indicating how MoViNet A0 models trained with and without ETM supervision perform. Our ETM-based model tends to be less affected by distractors and distinguishes well between visually similar actions that have different temporal contexts.

B. More Ablation Study Results

We repeat the ablation studies with the ConvNext architecture and show the results in Table S1. We observe very similar results to our MoViNets A0 experiments: including the ETM during training increases action recognition performance.

C. ETM with Rare Events

To further understand how our framework behaves, we analyze performance on rare events. We attempt to provide evidence to answer the question: is action recognition performance hindered when attempting to predict a rare class? We perform this analysis over the EK100 dataset.

We analyzed the performance of our model under rare events by computing performance metrics for two sets of classes: tail classes identified by [2] (Table S3 (a)), and classes from EK100's validation corresponding to the 11k actions not included in the dimensionality-reduced ETM. We show the results in the Table S3.

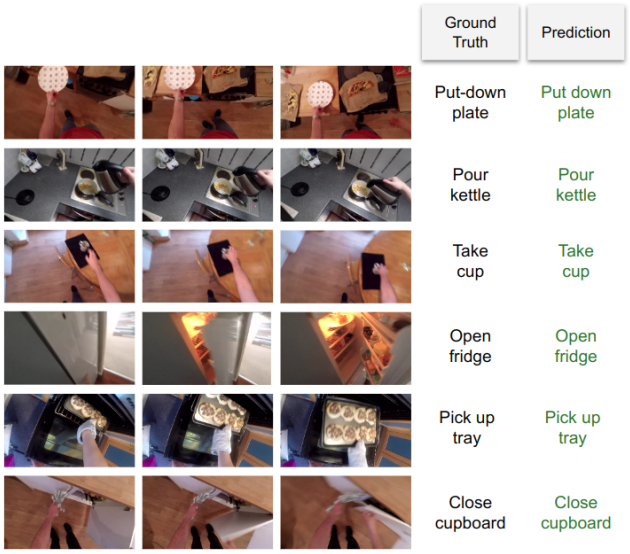
We observe our models with ETM show very similar or higher performance than the baselines. This indicates that training with ETM does not hinder performance on rare events, while improving performance on common ones. Our best models with ETM training still achieve competitive performance when compared to baselines and previous work. We hypothesize that the strength of the embeddings generated by our protocol yields better performance across classes, although improvements are mostly visible on the common classes encompassed by our dimensionality-reduced ETM.

References

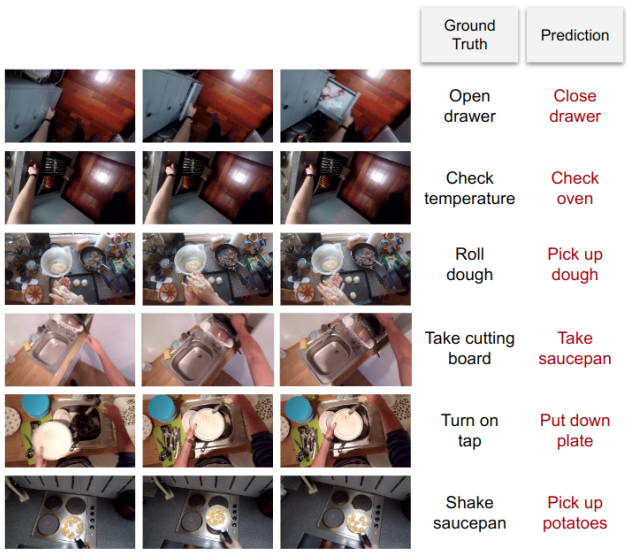
[1] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 2021. 2

[2] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13505–13515, 2021. 1, 3

[3] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *arXiv preprint arXiv:2201.03545*, 2022. 3



(a) Accurate action recognition predictions.



(b) Inaccurate action recognition predictions.

Figure S1. Action recognition predictions from our MoViNet A0 trained with ETM supervision on the validation set of EPIC-KITCHENS-100 [1].

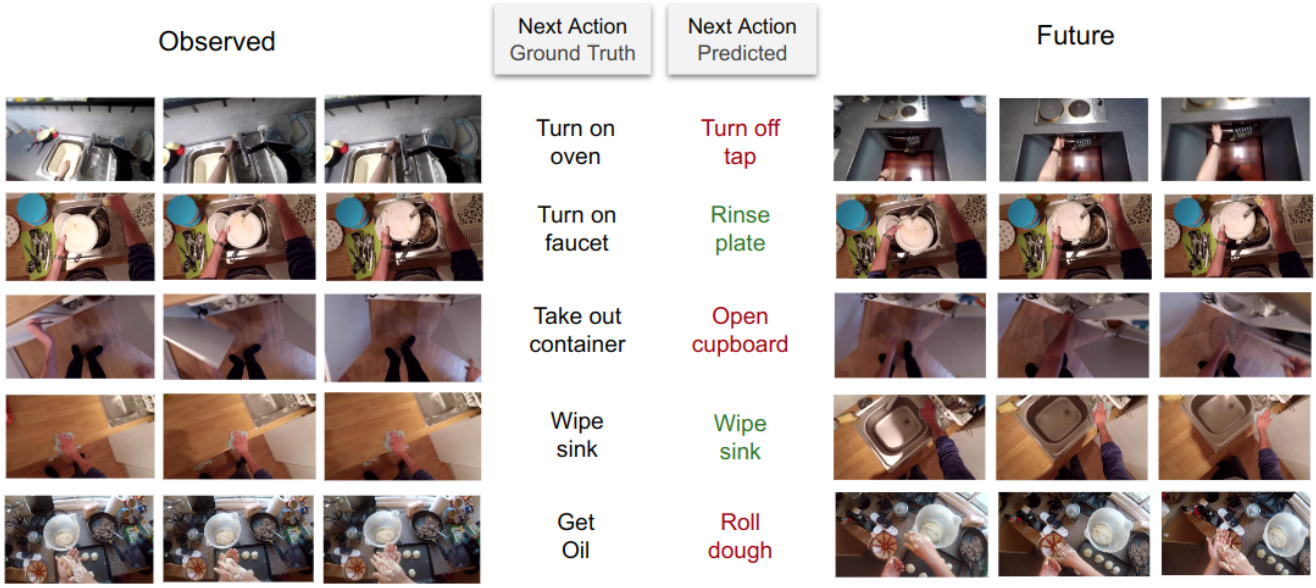


Figure S2. Action anticipation predictions from our MoViNet A0 trained with ETM supervision on the validation set of EPIC-KITCHENS-100. We show accurate and inaccurate predictions.

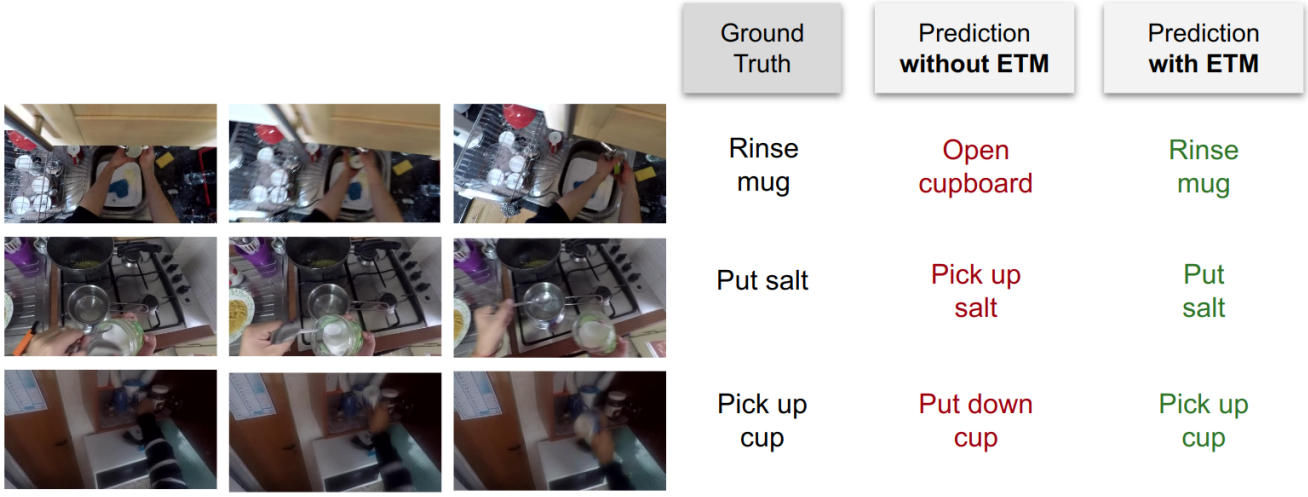


Figure S3. Qualitative comparison of our action recognition results with MoViNet A0 trained with and without ETM supervision. Our ETM-supervised model tends to be less affected by visual distractors (e.g. the cupboard), and generally distinguishes correctly between similar actions that occur at different points in time (e.g. pick up and put down an object).

Model	Present				MAE on **Past** ↓	MAE on **Future** ↓
	Verb ↑	Noun ↑	Action ↑			
			top-1	top-5		
Baseline	52.6	39.6	20.1	45.3	-	-
Full shuffle	51.1	40.4	19.9	46.7	4.515	4.112
Columns/rows shuffle	51.4	39.1	19.8	45.4	3.111	3.755
Co-occurrence	55.1	45.3	24.3	48.1	1.198	1.121
Only past vector	56.6	43.3	22.9	43.3	1.011	-
Only future vector	56.1	42.1	22.1	44.4	-	0.988
ETM (Ours)	**60.3**	**50.3**	**32.4**	**51.1**	**0.901**	**0.885**

Table S1. Action recognition results on various baseline models. We train the models on the EPIC-KITCHENS-100 dataset [2] with the ConvNext architecture [3].

Dataset	Frozen?	**Baseline**			**ETM (Ours)**		
		Verb ↑	Noun ↑	Action ↑	Verb ↑	Noun ↑	Action ↑
EK100	✓	9.9	9.7	4.11	**12.1**	**11.1**	**4.51**
		10.5	10.3	4.05	**12.9**	**12.2**	**5.01**

Table S2. Action anticipation results, shown for a ConvNext-S architecture with encoder weights frozen or with parameter updating allowed (check marks), both with and without the ETM.

Model	Verb ↑	Noun ↑	Action ↑
MoViNet A6	39.9	26.8	19.9
MoViNet A6 + ETM	40.3	27.1	20.1
X3D-M	39.2	25.9	19.2
X3D-M + ETM	39.4	25.8	19.2

(a) Performance on Tail Classes (Action Recognition on EK100, tail classes, Top-1 Accuracy)

Model	Verb ↑	Noun ↑	Action ↑
MoViNet A6	61.4	46.6	28.9
MoViNet A6 + ETM	61.3	46.9	29.2
X3D-M	59.8	45.7	26.8
X3D-M + ETM	60.1	45.4	26.5

(b) Performance on our custom validation subset (Action Recognition on EK100, infrequent classes, Top-1 Accuracy)

Table S3. How does the model handle rare events?