Supplementary material: AFF-ttention! Affordances and Attention models for Short-Term Object Interaction Anticipation

Lorenzo Mur-Labadia[1], Ruben Martinez-Cantin[1], Jose J.Guerrero[1], Giovanni Maria Farinella[2], and Antonino Furnari[2]

[1] University of Zaragoza, Spain
{lmur, rmcantin, jguerrer}@unizar.es
[2] University of Catania, Italy
{giovanni.farinella, antonino.furnari}@unict.it

1 Datasets

We validate our method on Ego4D [4] and EPIC-Kitchens [1], two large-scale datasets of egocentric videos with high diversity and long-tail distributions of verbs and nouns classes.

Ego4D We consider both the first and second versions of Ego4D STA annotations. Version 1 (v1) of the Ego4D STA split is composed of 27,801 training 17,217 validation and 19,870 test instances with 87 noun and 74 verb categories. Version 2 (v2) extends v1 with additional videos and annotations, for a total of 98,276 training, 47,385 validation and 19,870 test videos, with a taxonomy of 128 nouns and 81 verb classes. Ego4D STA contains a single test split which is compatible with v1 and v2, hence models trained on either versions can be compared on the same test split, with methods trained on v2 generally performing better thanks to the extended annotations.

EPIC-Kitchens Since Ego4D is the only dataset containing STA annotations to date, we contribute a second benchmark by extending over action and object annotations available in the popular EPIC-Kitchens dataset [1]. The extended annotations are obtained by post-processing and filtering existing active object and action segment annotations available in the EPIC-KITCHENS dataset. STA annotations are obtained in four stages. In the first stage, we process active object bounding box annotations to form object tracks. This is done by merging in the same track neighbouring annotations related to the same object class. Since annotators have been instructed to label all instances of the active object (e.g., all instances of “plate“), we removed all tracks containing frames with more than one labeled bounding box for the same object class. In the second stage, each object track is matched to one of the annotated action segments. Specifically, if an action segment including the same noun as the object track is found, this is matched to the object track. In the third stage, we truncate object tracks so that they do not include the action. This is done because we are interested in anticipating the future action, which hence should not be observed. In the fourth stage, we attach the following data to a given bounding box: the noun associated to the track as the object category, the of the associated action segment as the interaction verb, the distance from the timestep of the current frame to the beginning of the associated action segment as the time to contact. The final set of annotations contains 33,804 training and 7,055 validation instances with 104 noun and 51 verb categories. Annotations will be publicly released to encourage future research.

2 Implementation details

Training details We train the STAformer classifier with Adam as an optimizer, an initial learning rate of 10^{-4} with linear warm-up, and a weight decay of 10^{-6}, on 4 Tesla V100 GPUs. For both datasets, we sample 16 frames at 30 frames per second to form the input video clips, which are downscaled to a height of 320 pixels. We follow the standard Faster-RCNN resolutions and resize the short side in the range [640, 672, 704, 736, 768, 800] with a maximum long side of 1333. We use the DINOv2 "base" version composed of 12 blocks and an embedding space of 768 channels. Following previous work [7], we only fine-tune the last 3 blocks. We initialize the TimeSformer weights with the dual encoder version of EgoVLP-v2 [14] pre-trained on EgoClip [9]. We interpolate the spatial and temporal positional encoding of TimeSformer to adapt the video processing resolution, established in 224 × 336.

STA Prediction head We adopt the prediction head proposed in [15], which modifies the Faster-RCNN [3] head integrating components specialized for STA prediction. In short, the fused feature pyramid P_T (Figure 2.d) is passed to a Region Proposal Network (RPN), which computes object proposals. Such proposals are then used to extract local features from P_T with RoI Align [5], mapping bounding boxes to appropriate layers of the pyramid following [10]. Each extracted local feature vector is concatenated with the fused class token C_T (Figure 2.c) and passed through an MLP with a residual connection. Linear layers are used to compute noun probabilities $p(n)_i$, verb probabilities $p(v)_i$ and time-to-contact (ttc) predictions.

Extracting environment affordances zones To build the affordance database we follow Algorithm 1 from [13] to define zones, but omit adding edges between zones. This algorithm compares the similarity between two frames with a pre-trained Siamesse neural network composed by a Resnet-18 [6] followed by a 5 layer multi-layer perceptron (MLP). Previously, we select similar/dissimilar frames for the training labels. We consider two similar frames if they were separated less than 15 frames or they share 10 inlier homography key-points. We extract SuperPoint keypoint descriptors [2] for the homography estimation. We measure the visual similarity from pre-trained ResNet-152 features [6] to select dissimilar frames.

Table 1: Ablation study on different configurations of the environment affordances performance on the Ego4D-STA v1, using a StillFast architecture as base model.

K(text)	K(video)	Dist.	N	N+V	N+δ	All
No affordances			16.20	7.47	4.94	2.48
Ego-Topo [13]			14.92	6.45	4.01	2.14
top-4	-	U	17.64	8.05	5.13	2.58
-	top-4	U	17.98	7.95	5.19	2.67
top-4	top-4	U	18.27	8.24	5.35	2.71
top-2	top-2	W	17.90	7.77	5.34	2.65
top-4	top-4	W	18.44	**8.52**	5.46	**2.85**
top-6	top-6	W	**18.75**	8.46	**5.47**	2.81
top-8	top-8	W	18.53	8.40	5.45	2.76

Leveraging interaction hotspots: On the interaction hotspots model, we fine-tune the hands detector by Shan et al. [16] with the annotations present at the Object State Change detection (SCOD) split of Ego4D. We remove from the SCOD training split those scenes present at the STA validation splits.

3 Extra results

Ablation study on Environment Affordances Table 1 compares the performance of StillFast when using different environment affordance configurations as priors. Compared with training an NN using the affordances as ground-truth for supervision as in [13] does not improve the final performance after fusing the measurements. Our intuition is that the NN learns almost the same interactions as the STA classifier and therefore it does not improve when both are fused and loses the generalist quality of our approach. First, we evaluate the influence of the visual and cross similarity when selecting the K-Nearest Neighbours (NN). As the results show, the combined version obtains the best performance (18.27 N, 8.24 N+V, 5.35 N+δ and 2.71 All). This indicates that visual similarity is not enough in a cultural-diverse dataset like Ego4D, where the same action can be performed in multiple different environments. Second, we weigh the influence of each zone affordance distribution according to its respective similarity. This version obtains a minor improvement up to 18.44 N, 8.52 N+V, 5.47 N+δ and 2.85 Overall. Finally, we analyze the influence of the number of considered zones at the K-NN. Keeping a low number of zones reduces the affordance richness and makes it more difficult for the next interaction to falls inside, while increasing the number of considered zones creates a wider affordance distribution but dilutes the STA value.

Visualization of Environment affordances nodes We illustrate in Figures 1 and 2 some environmental affordances zones extracted from Ego4D and EPIC-Kitchens, respectively. Each row shows different frames that compose the zone captured at different moments of the sequence, representing an activity-centric region like *washing the dishes, ironing, unboxing a package, taking fresh food....*

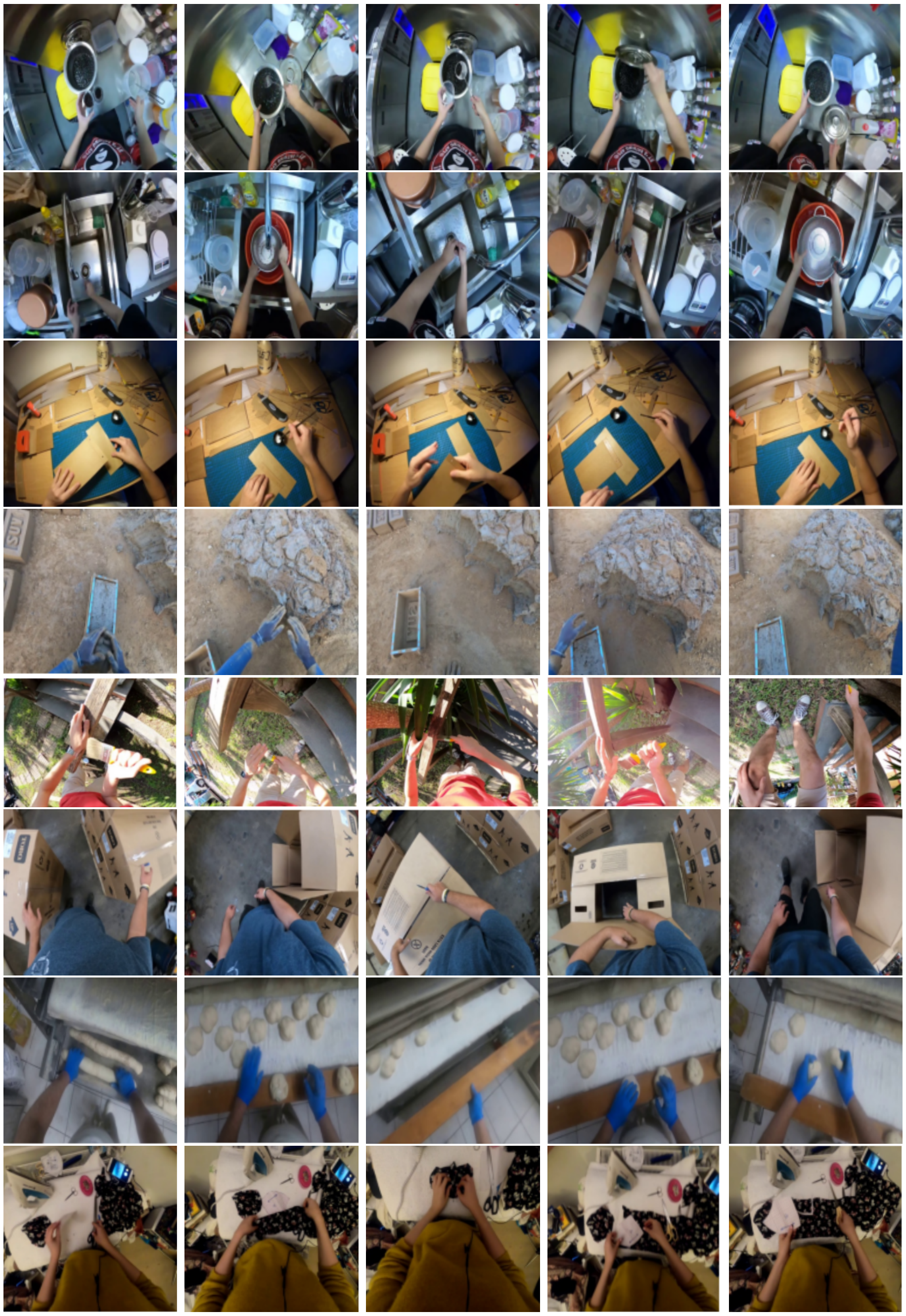

Fig. 1: Zones of the environment affordances database extracted in Ego4D following [13]. Each row corresponds to a single node representing an activity-centric region.

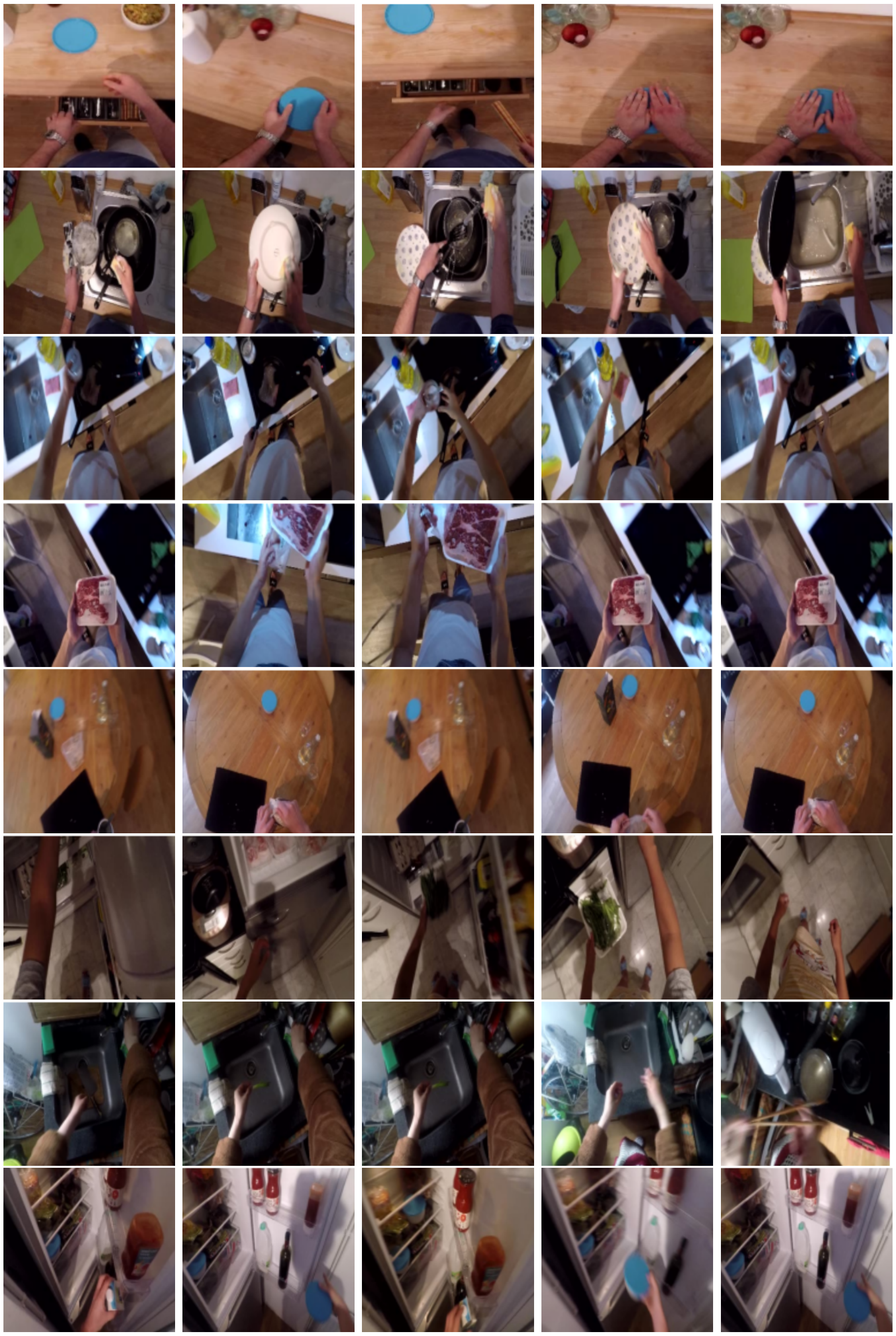

Fig. 2: Zones of the environment affordances database extracted in EK following [13]. Each row corresponds to a single node representing an activity-centric region.

Table 2: Object interaction hotspots performance on the Ego4D-STA v1.

	SIM	AUC-Judd	NSS
Hotspots [12]	0.309	0.665	0.145
LOCATE [8]	0.324	0.641	0.258
Join hands [11]	0.652	0.806	1.305
Join hands (ours)	**0.672**	**0.832**	**1.421**

Table 3: Hands detection Average Precision (AP) score on the Ego4D SCOD v1.

	AP25	AP50	AP75
Zisserman [17]	0.781	0.567	0.487
Sha et al. [16] ❄	0.874	0.855	0.751
Sha et al. [16] 🔧	**0.931**	**0.906**	**0.783**

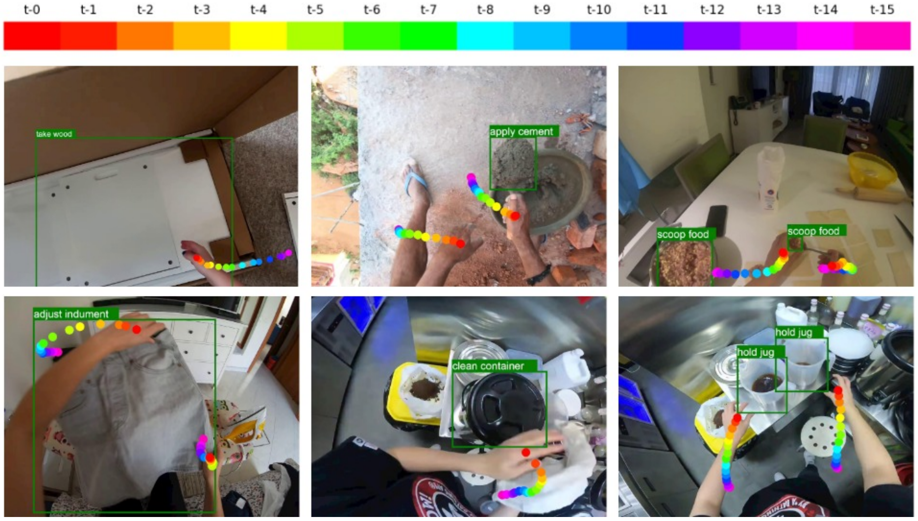

Fig. 3: Hands trajectory along the seen video, obtained by finetunned a hand object detector on Ego4D sequences

Interaction hotspots results We compare in Table 2 the performance of different baselines on the interaction hotspots prediction. We obtain minor improvements (0.672 vs 0.652 SIM, 0.832 vs 0.806 AUC-Judd, and 1.421 vs 1.305) on the model by Liu et al. [11] with features pre-extracted with EgoVLP and refined bounding boxes obtained with a fine-tunned version of Sha et al. [16] hand objects detector. We report in Table 3 the effect of fine-tuning the hands detector on the AP on the Ego4D - State Change Object Detection (SCOD) validation split hand bounding box labels.

4 Qualitative results

We show in Figures 4 qualitative examples of the STA predictions on Ego4D. Figure 5 illustrates our curated STA annotations and predictions on EPIC-Kitchens. The model predicts several next-active objects, showing multiple plausible scenarios and the challenges associated to the task.

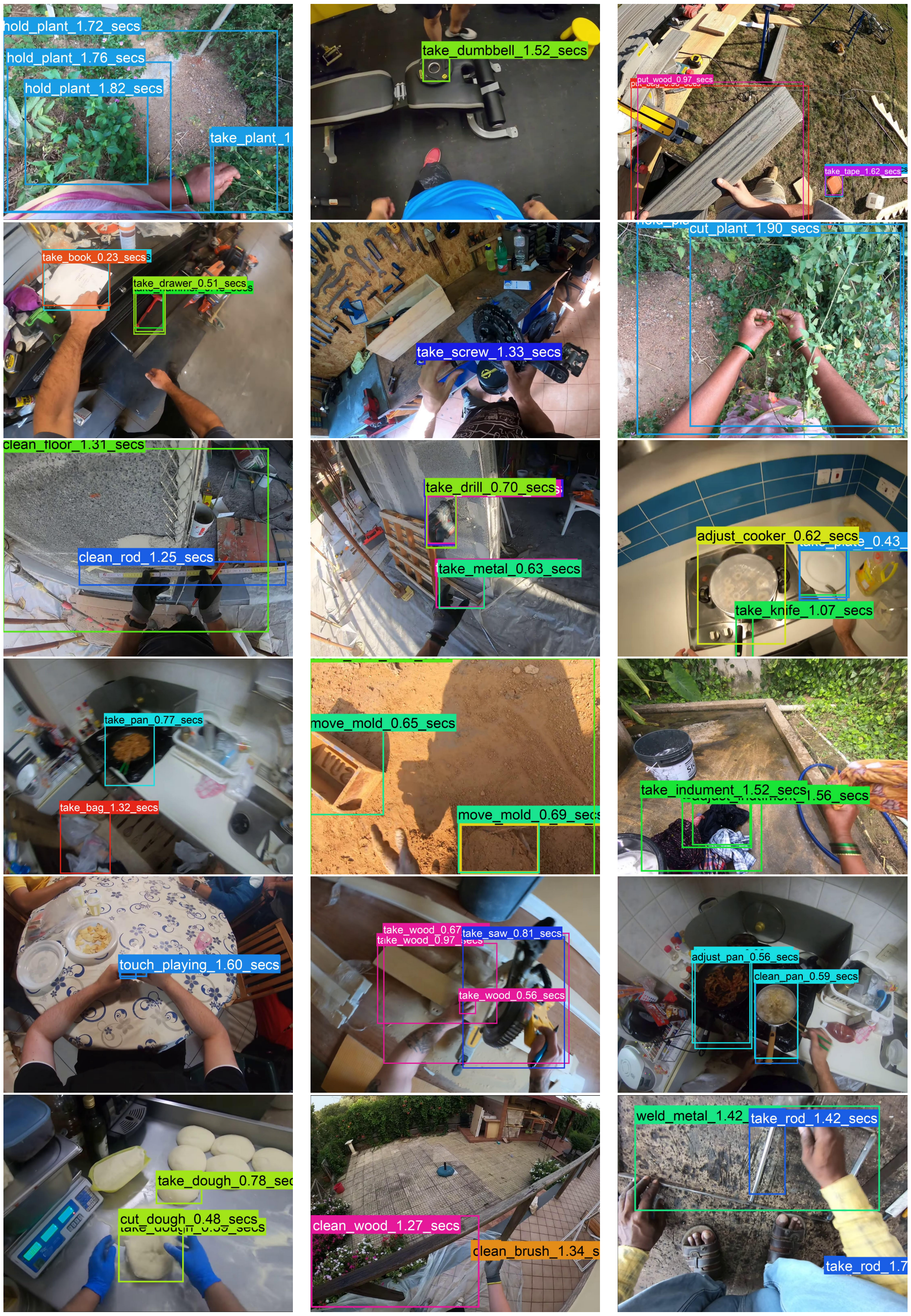

Fig. 4: Qualitative results on Ego4D

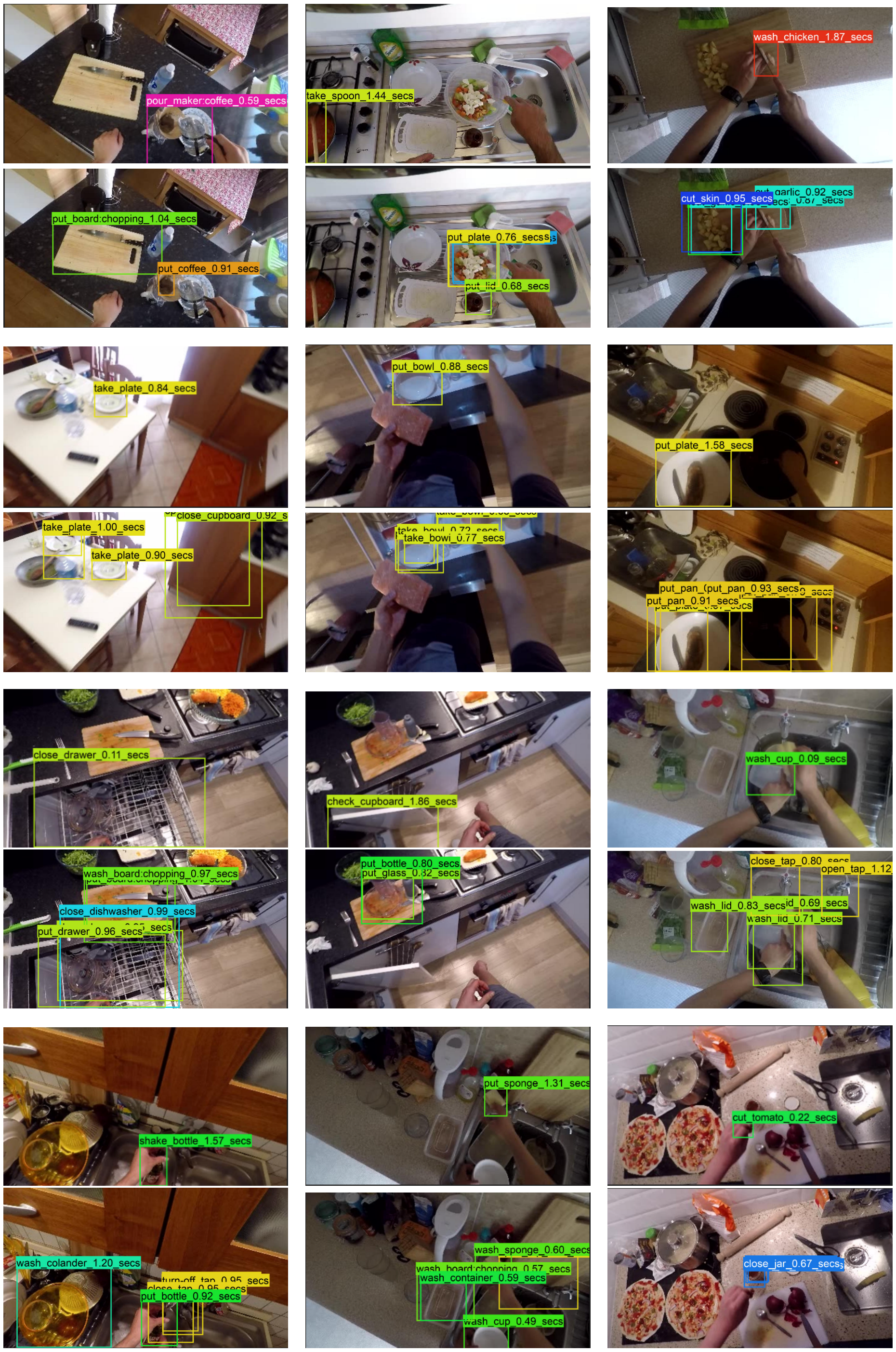

Fig. 5: STA curated labels (top) and qualitative results (bottom) on EK

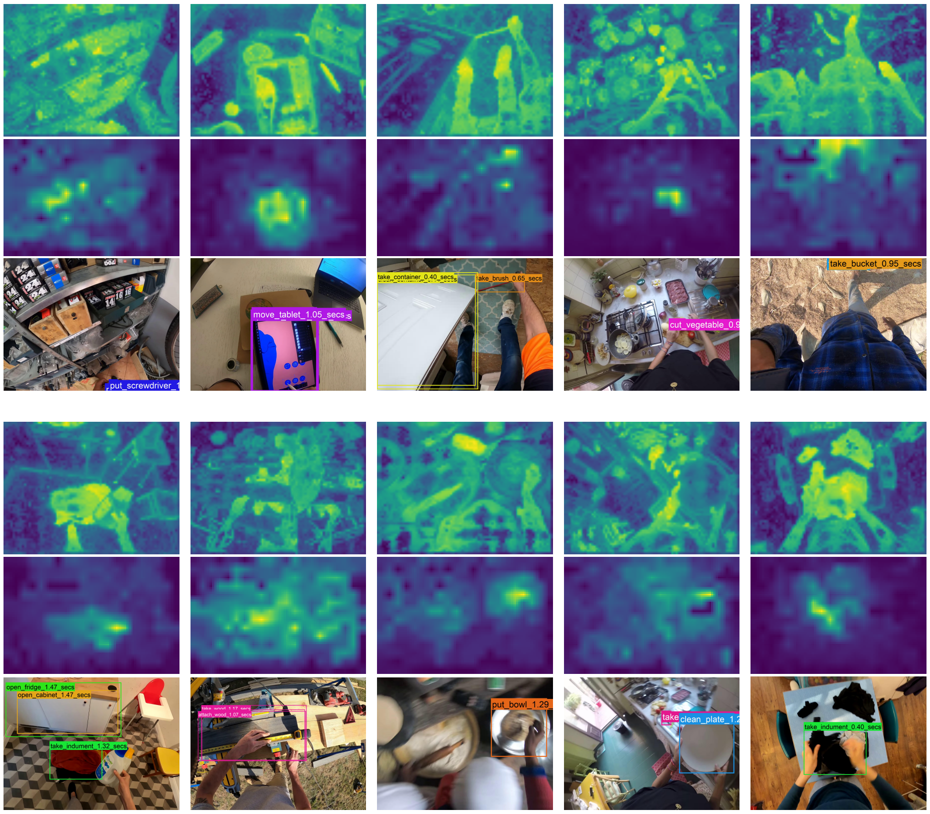

Fig. 6: Qualitative results and dual cross-attention maps. We show the attention map of the video on the image features (top), the attention map of the image on the video features (middle) and the result of the final predictions.

Attention maps The attention maps reflects the different parts of the scene attended by at the Dual Cross Attention mechanism. On top, we appreciate how the video features are refined with fine-grained object information in the high-resolution image, like the small vegetables on the fourth column. In the middle, we show the parts of the video more relevant for the static image, which correspond to dynamic regions associated with the hands motions very informative for the upcoming interaction, like the third column example. In the last row, we show the result of the predictions.

References

1. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Scaling egocentric vision: The epic-kitchens dataset. In: Proceedings of the European conference on computer vision (ECCV). pp. 720–736 (2018) 1
2. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 224–236 (2018) 2
3. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 1440–1448 (2015) 2
4. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18995–19012 (2022) 1
5. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017) 2
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 2
7. Izquierdo, S., Civera, J.: Optimal transport aggregation for visual place recognition (2024) 2
8. Li, G., Jampani, V., Sun, D., Sevilla-Lara, L.: Locate: Localize and transfer object parts for weakly supervised affordance grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10922–10931 (2023) 6
9. Lin, K.Q., Wang, J., Soldan, M., Wray, M., Yan, R., XU, E.Z., Gao, D., Tu, R.C., Zhao, W., Kong, W., et al.: Egocentric video-language pretraining. Advances in Neural Information Processing Systems **35**, 7575–7586 (2022) 2
10. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017) 2
11. Liu, S., Tripathi, S., Majumdar, S., Wang, X.: Joint hand motion and interaction hotspots prediction from egocentric videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3282–3292 (2022) 6
12. Nagarajan, T., Feichtenhofer, C., Grauman, K.: Grounded human-object interaction hotspots from video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8688–8697 (2019) 6

13. Nagarajan, T., Li, Y., Feichtenhofer, C., Grauman, K.: Ego-topo: Environment affordances from egocentric video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 163–172 (2020) 2, 3, 4, 5
14. Pramanick, S., Song, Y., Nag, S., Lin, K.Q., Shah, H., Shou, M.Z., Chellappa, R., Zhang, P.: Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5285–5297 (2023) 2
15. Ragusa, F., Farinella, G.M., Furnari, A.: Stillfast: An end-to-end approach for short-term object interaction anticipation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3635–3644 (2023) 2
16. Shan, D., Geng, J., Shu, M., Fouhey, D.F.: Understanding human hands in contact at internet scale. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9869–9878 (2020) 3, 6
17. Zhang, C., Gupta, A., Zisserman, A.: Helping hands: An object-aware ego-centric video recognition model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13901–13912 (2023) 6