Supplementary Material for Active Object Detection with Knowledge Aggregation and Distillation from Large Models

Dejie Yang Yang Liu[*]
Wangxuan Institute of Computer Technology, Peking University
ydj@stu.pku.edu.cn yangliu@pku.edu.cn

Table 1. Variant detectors.

Dataset	Method	AP	AP50	AP75
Ego4D	DETR	15.5	32.8	13.0
	ours	**25.3**	**33.6**	**24.7**
Epic-Kitchens	DETR	10.4	15.7	10.1
	ours	**23.5**	**26.0**	**20.1**
Ego4D	DeformableDETR	18.7	33.3	17.5
	ours	**26.0**	**35.1**	**25.5**

Table 2. Results across datasets.

Target	Method	AP	AP50	AP75
Epic-Kitchens	InternVideo	11.2	14.8	9.8
	ours	**13.3**	**16.7**	**12.0**
MECCANO	InternVideo	6.4	10.2	5.3
	ours	**9.3**	**13.3**	**8.0**
100DOH	InternVideo	9.5	12.7	8.9
	ours	**13.0**	**14.2**	**9.5**

In this supplementary material, we provide more abalations, qualitative results and analysis, as well as additional implementation and experimental details. We add the generalization ability and performance across datasets of our KAD in additional abalations 1. We illustrate qualitative results, attention map visualizations and our generation results in Section 2. And we provide more details of datasets and experiment settings in Section 3.

1. Additional Abalations

Generalization Ability with Variant Detectors. we conduct an experiment with plain DETR detector to ensure fair comparisons with other approaches. In Table 1, ours surpasses the approach using DETR detector(rows 1-2, +9.8%@AP and +13.1%@AP on two datasets).

Performance across Datasets.Table 2 (trained on Ego4D and test on others) shows better generalization capability of ours. Compared with InternVideo, our method can achieve better performance cross datasets, which may be due to our incorporation of knowledge from spatial, vision, and semantic knowledge distillation, allowing the model to learn AOD priors that can be generalized.

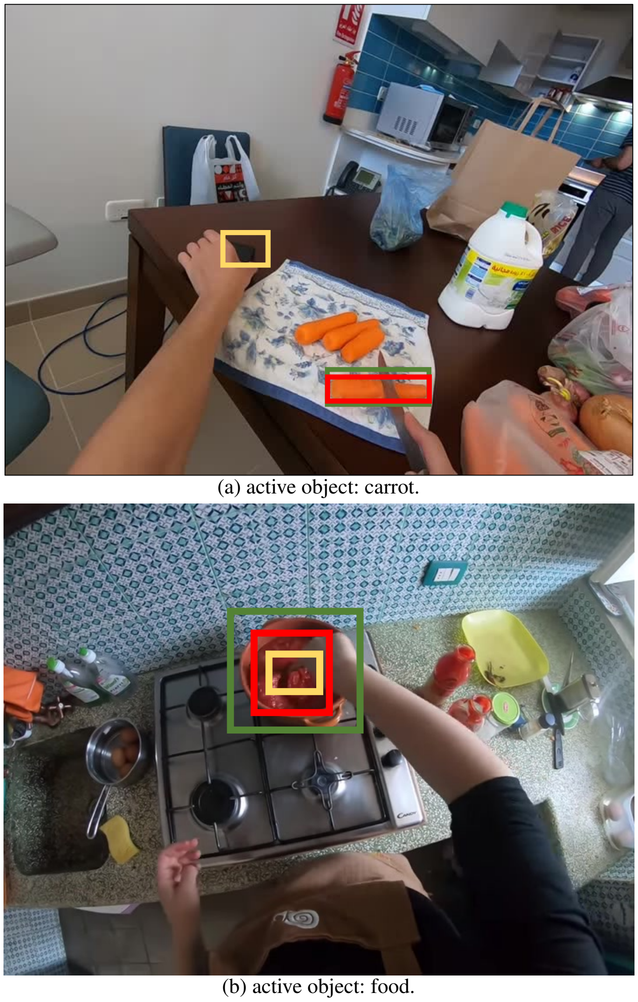
2. Qualitative Comparisons

Case Study. Figures 1a and 1b show our visual comparison results of active object detection and InterVideo[9](the best existing method), in which green box represents the ground truth, ours and InternVideo's predictions are colored with red and yellow. The Figure 1a illustrates our method's superior accuracy in detecting active objects. In Figure1a, we can distinguish the genuine active object 'carrot' as opposed to InterVideo's misidentification (a 'phone' in left hand). In Figure 1b, though our result is under low IoU(Intersection over Union) with the ground-truth, our approach accurately pays attention on the 'food' being stirred. Through the incorporation of related priors to active objects, the priors enriched cues function as enhanced indicators, effectively guiding the detection process towards active objects.

Attention Map Visualization. In Figures 2a and 2b, we show the attention map comparison results between InterVideo[9](the best existing method) and our method. We provide the final detection results in Figure2c. Compared with the ground-truth(colored with green, chain), our detection result(colored with red) obtain greater IoU than InterVideo [9] (colored with yellow). The attention of InterVideo[9](Figure 2a) is mainly distributed in the upper left part (possibly related to their incorrect detection results, a toolbox, also in the upper left corner). Our attention(Figure 2b) is associated with the surrounding objects and tools, as we introduce prior knowledge of the active object (including interactions and related objects) to guide the model in inferring and locating the active object by analyzing potential interactions. This also demonstrates the effectiveness of introducing prior knowledge of active objects.

Semantic Interaction Generation Results. In Figures 3a and 3b, we provide two examples of generating semantic

[*] Corresponding Author

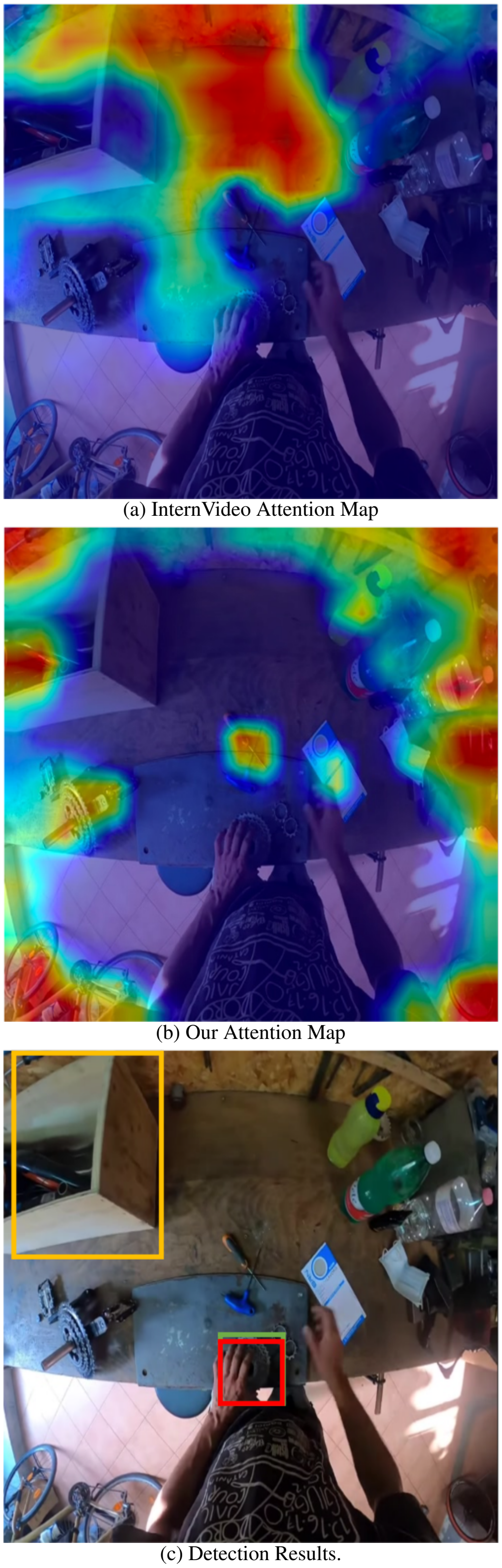

(a) active object: carrot.

(b) active object: food.

Figure 1. Qualitative Comparisons.The green box represents ground truth, the yellow box represents the detection results of InternVideo[9], and the red box represents our Knowledge Aggregation and Distillation(KAD) detection results.

interactions for 'carrot' and 'wood' respectively. Specifically, we used gpt-4[1] with prompt "describe 10 interaction descriptions of *[object]* undergoing state change (including tools)". It can be seen that these descriptions can indeed describe the scene, interaction, and description of object state changes. The effectiveness of this part of the text description can also be seen through the experimental results in the main paper.

Visual Image Generation Results. In Figures 4a and 4b, we show the generated image results by [10] with the interactions 'Carrot is being sliced using a knife' and 'Carrot is being juiced using a juicer' respectively. The images show the state and corresponding visual information of the 'carrot' under different interactions. Compared to abstract concepts in text, images can more intuitively display fine-grained


(a) InternVideo Attention Map

(b) Our Attention Map

(c) Detection Results.

Figure 2. Attention Map Visualization. In figure (c), the green box represents ground truth, the yellow box represents the detection results of InternVideo[9], and the red box represents our Knowledge Aggregation and Distillation(KAD) detection results.

[1] https://chat.openai.com/

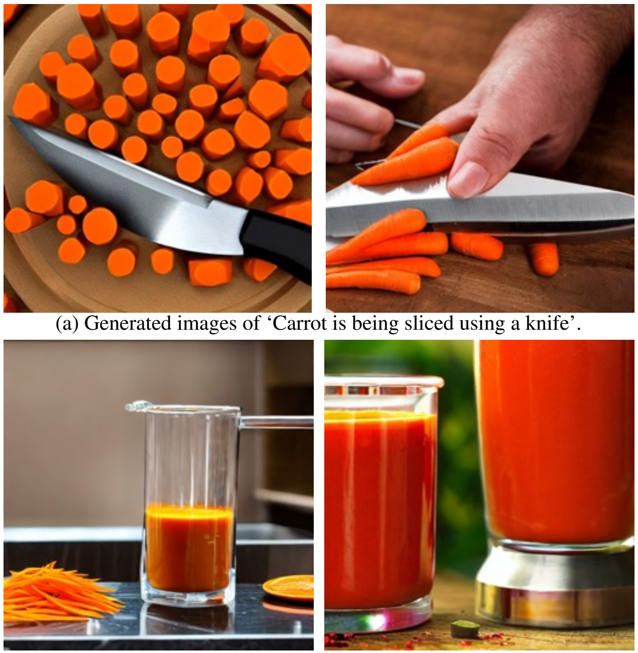
1. Carrot is being washed using a faucet.
2. Carrot is being peeled using a peeler.
3. Carrot is being sliced using a knife.
4. Carrot is being grated using a grater.
5. Carrot is being boiled using a pot.
6. Carrot is being steamed using a steamer.
7. Carrot is being roasted using an oven.
8. Carrot is being pureed using a blender.
9. Carrot is being juiced using a juicer.
10. Carrot is being fermented using a fermentation jar and salt.

(a) active object: carrot.

1. Wood is being cut using a saw.
2. Wood is being sanded using sandpaper.
3. Wood is being planed using a plane.
4. Wood is being drilled using a drill.
5. Wood is being carved using chisels.
6. Wood is being hammered using a hammer.
7. Wood is being painted using a paintbrush.
8. Wood is being stained using a staining brush.
9. Wood is being varnished using a varnish brush.
10. Wood is being burned using a wood burner.

(b) active object: wood.

Figure 3. Semantic Interaction Generation Results for 'carrot' and 'wood'.



(a) Generated images of 'Carrot is being sliced using a knife'.

(b) Generated images of 'Carrot is being juiced using a juicer'.

Figure 4. Visual Image Generation Results.

visual information about object interactions.

3. Implementation Details

3.1. Dataset

Ego4D [5] stands as one of the latest expansive egocentric video datasets. We focus on subsets of this dataset for our state-change object detection (SCOD) tasks. The original train and validation sets encompass 19,070 and 12,800 annotated frames, respectively, marking the point of no return, or the initiation of a state change. For active object, we only detect the state-change object (active object) on the PNR frame in Ego4D [5] to make fair comparisons with other methods[1, 4, 7–9].

Epic-Kitchens [2] is another prominent and extensively utilized large-scale dataset in the domain of egocentric vision. In our context, we convert the segmentation annotations of action-related objects within the VISOR subset [3] into bounding boxes, specifically tailored for the active object detection task. For Epic-Kitchen, we treat these objects and bounding boxes as state-changing object detection (SCOD) annotations, convert the segmentation annotations in VISOR [3] into bounding boxes, and filter out non-action-related active objects, akin to Ego4D [5]. We consider the frames annotated in VISOR as keyframes for state changes and select the center annotated frame if multiple annotations exist in the same video. And we adopt the average precision (AP) as the metric following [5]. Notably, we employ a total of 67,217 and 9,668 annotated frames for our train and validation splits, respectively.

3.2. Implementation Details

We use GPT-4 to generate the interaction descriptions and a stable diffusion model[10]. The embeddings of semantic and visual features are extracted through CLIP[6]. The spatial prior is the normalized bounding box of active object in the input image. The dimension d of encoded features and queries is 2048. And the dimensions of the fused semantic prior and visual prior d_t and d_v are both 510.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3

[2] 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. *IJCV*, 130:33–55, 2022. 3

[3] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2022. 3

[4] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6569–6578, 2019. 3

[5] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, pages 18995–19012, 2022. 3

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3

[7] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 3

[8] Dandan Shan, Jiaqi Geng, Michelle Shu, and David Fouhey. Understanding human hands in contact at internet scale. In *CVPR*, 2020.

[9] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 1, 2, 3

[10] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7452–7461, 2023. 2, 3