Interaction-Aware Prompting for Zero-Shot Spatio-Temporal Action Detection Supplementary Material

Wei-Jhe Huang[1] Jheng-Hsien Yeh[1] Min-Hung Chen[2] Gueter Josmy Faure[3] Shang-Hong Lai[1]
[1]National Tsing Hua University, Taiwan [2]NVIDIA [3]National Taiwan University
{weijhie, goodcharlie1018, vitec6, josmyfaure}@gmail.com lai@cs.nthu.edu.tw

In this supplementary material, we aim to provide further analysis on our experiments, and show more experimental results for comparison with other methods.

1. The efficacy of baseline

To further showcase the CLIP baseline's capability in recognizing unseen actions, we directly utilize the baseline to infer all labels of the J-HMDB dataset. We compare with the methods [2, 3] that have done the same experiment on J-HMDB. To hava a fair comparison, we use the same perosn detector as theirs(i.e. Faster R-CNN, pre-trained on MS-COCO), and use video mAP for evaluation as they also provided. Among the settings of IoU threshold from 0.1 to 0.5, our CLIP baseline all register significant gains compared to them. When the threshold is set to 0.1, we can achieve 56.82 mAP score, while these two methods obtain 27.5 and 32.1, respectively. Note that the influence of localization error can be almost ignored when the threshold is very small, thus it can be more focused in the comparison of classification.

2. Analysis on 50% vs 50% experiment

In our study, we perform a zero-shot experiment in 50% vs 50% labels split and compare the performance of our model with the baseline model on the UCF101-24 dataset [5]. The results in Table 1 show that with prompting, the baseline model achieved a higher mAP score than our model on UCF101-24. Further analysis revealed that due to the lower resolution of UCF101-24 videos, it will lead to noisy results in the process of generating interaction features. Moreover, the 50% vs 50% labels split reduces the training data, which may amplify the noise and favor the baseline model that relies solely on image features. Nevertheless, the prompting mechanism enhances the performance of both our model and the baseline. These findings suggest that prompts are beneficial for this task. Additionally, our model outperforms the baseline for the 75% vs 25% experiment on UCF101-24, indicating that with an appropriate amount of training data, we can still generate representative interaction features for detecting unseen actions even if the video has lower resolution.

Dataset	model		+IAP
J-HMDB	Baseline	42.31	44.55
	iCLIP	**44.29**	**45.18**
UCF101-24	Baseline	58.90	**61.86**
	iCLIP	**59.78**	60.30

Table 1: **Zero-shot inference results in 50% vs 50% labels split.** The baseline uses the image feature from the whole frame for inference. +IAP: Complete model that contains Interaction-Aware Prompting.

3. Average precision (AP) of each unseen class

For a more detailed comparison of the results, we present the average precision (AP) of each unseen class. Table 2 presents the result on J-HMDB, our model performs better on half of the classes. In addition, in these worse classes, we are only 12% lower than the baseline at most, while the others are almost the same. On the other hand, we have made great progress in better classes, with a minimum improvement of 12% and a maximum of almost 21%. From Table 3, we can see that on UCF101-24, our model has progressed in most classes, especially for the challenging class where the baseline has only 3.55% AP.

4. More details of bounding box

For person boxes, we take groundtruth boxes at training time, and we use the boxes detected from [1] at inference time, which is a single-stage framework for action localization and classification. Besides, in order to avoid wrongly detected person boxes from causing noise in the interaction module, we only take boxes whose confidence scores are greater than 0.2 at inference time. Regarding object detection, we employ Faster-RCNN to detect object boxes during both training and inference. We select objects that intersect

model	catch	clap	pullup	sit	throw	wave	mAP
Baseline	69.70	45.95	**99.98**	**41.87**	33.57	**65.71**	59.46
iCLIP	**81.66**	**66.74**	99.94	38.46	**52.22**	53.42	**65.41**

Table 2: **Frame AP of J-HMDB per unseen class in 75%v.s.25% labels split.** The baseline uses the image feature of whole frame for inference. Both baseline and iCLIP are without prompting.

model	FloorGymnastics	IceDancing	SalsaSpin	SkateBoarding	SoccerJuggling	VolleyballSpiking	mAP
Baseline	74.69	65.98	**63.92**	91.25	**98.64**	3.55	66.34
iCLIP	**87.29**	**67.26**	58.79	**92.68**	98.62	**21.37**	**71.00**

Table 3: **Frame AP of UCF101-24 per unseen class in 75%v.s.25% labels split.** The baseline uses the image feature of whole frame for inference. Both baseline and iCLIP are without prompting.

with any person (i.e. IoU > 0) to capture relevant contextual information.

For the action detection framework, since we can only use part of the training data in the zero-shot setting, it is more challenging to train a localization network from scratch. Instead, we use extra person detector for localization and let our model focus on recognizing unseen actions. Notably, even in fully-supervised settings, several SOTA methods [6, 4, 7] exploit human detector and only focus on classification.

5. Capability for full supervision

For fully-supervised setting, our approach achieves frame mAP of 73.70 on J-HMDB and 78.19 on UCF101-24 respectively. Note that the CLIP encoders are frozen during training, which indicates that our tunable parameters(11.6M) are much less than other SOTA methods. In addition, we also conduct the experiment with our baseline, which obtains 48.70 on J-HMDB and 58.08 mAP on UCF101-24. No matter in fully-supervised or zero-shot setting, our method can make more effective use of visual-language features to improve performance.

6. Advantage

Our model has a significant advantage over other models in its class, as it requires only **11.6M** parameters for training. Compared to other models, which usually require larger numbers of parameters in the training, our model can achieve high performance without incurring as much computational cost. In particular, this advantage is especially pronounced for zero-shot scenarios, where models must be able to learn quickly and adapt to unseen data. By utilizing fewer parameters, our model is able to learn faster and more efficiently, enabling it to outperform other models for zero-shot action detection tasks.

References

[1] Okan Köpüklü, Xiangyu Wei, and Gerhard Rigoll. You only watch once: A unified cnn architecture for real-time spatiotemporal action localization. *arXiv preprint arXiv:1911.06644*, 2019. 1

[2] Pascal Mettes and Cees G. M. Snoek. Spatial-aware object embeddings for zero-shot localization and classification of actions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1

[3] Pascal Mettes, William Thong, and Cees GM Snoek. Object priors for classifying and localizing unseen actions. *International Journal of Computer Vision*, 129:1954–1971, 2021. 1

[4] Junting Pan, Siyu Chen, Mike Zheng Shou, Yu Liu, Jing Shao, and Hongsheng Li. Actor-context-actor relation network for spatio-temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 464–474, June 2021. 2

[5] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1

[6] Jiajun Tang, Jin Xia, Xinzhi Mu, Bo Pang, and Cewu Lu. Asynchronous interaction aggregation for action detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 71–87. Springer, 2020. 2

[7] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 2