Exploring Scalability of Self-Training for Open-Vocabulary Temporal Action Localization

Jeongseok Hyun[1] Su Ho Han[1] Hyolim Kang[1] Joon-Young Lee[2] Seon Joo Kim[1]
[1]Yonsei University [2]Adobe Research

Appendix

In this section, we provide additional experimental results and details not presented in the main paper.

A. Additional Implementation Details

We primarily adopt the hyperparameters from ActionFormer [20] and ViFi-CLIP [14] since our architecture is based on them. As addressed in Sec. 4.1, the ActivityNet [2] dataset differs significantly from the THUMOS14 [4] and FineAction [9] datasets. It consists of only 1 – 2 long action instances, limiting the evaluation of the capability in action localization. Accordingly, we adjust hyperparameters for the ActivityNet dataset, following the TAL literature.

A.1. Video Feature Extraction Details

Here, we detail the video feature extraction process using VLMs. As stated in Sec. 4.3, we extract the video snippet features ($\mathbf{F}_{\mathcal{V}}$) using the conventional sliding window manner with a window size of 16 frames and a stride size of 4 frames, after interpolating videos into 30 fps. For ActivityNet, we interpolate $\mathbf{F}_{\mathcal{V}}$ to a fixed length following the widely used convention [7, 8, 18, 20]. Specifically, each video is interpolated to a length of 192 feature vectors, as employed in ActionFormer [20]. For THUMOS14 and FineAction, we retain the video snippet features in their original length. When we conduct experiments with ViCLIP [17] VLM model, we interpolate its learned temporal embedding from 8 to 16 to use the same window size for video feature extraction. Note that such details are often absent in existing OV-TAL methods [5, 11], making it challenging to reproduce their results. We open-source the extracted features to promote further development in the OV-TAL community.

A.2. STOV-TAL Inference Details

Compared to ActionFormer [20], we use the class-agnostic action localizer and assign the action classes from the VLM. As a result of this change, we delay the Soft-NMS operation until after the action classes are assigned, rather than immediately following the output of the action localizer. On the other hand, when we compute pseudo labels on unlabeled videos, we directly perform Soft-NMS on the class-agnostic action instances, based on its actionness score (s_a). We employ the same Soft-NMS configurations for both cases, with slight variations based on the dataset. We choose the top 100 scoring action instances for ActivityNet and 200 for THUMOS14 and FineAction, applying a minimum confidence score threshold of 0.001 and a tIoU threshold of 0.1.

A.3. Gemini Inference Details

We provide the instruction template used for Gemini to perform the TAL task in Fig. 1. To incorporate temporal information effectively, we adopt an interleaved format, as shown in Fig. 1 {time_instructed_video_data}, where RGB frame data alternates with its corresponding temporal information data throughout the entire video sequence. This ensures that both the visual content and temporal details are presented simultaneously. We also experimented with another format, which provides temporal information after all RGB frame data, following this structure: "These frames are located at {frame_time_list}." However, Gemini was unable to effectively perform TAL with this instruction format. The maximum length of the Gemini output is set to 4096. For parsing Gemini's response to TAL format, we employ the regular expression of $(\mathbb{R}, \mathbb{R}, \mathbb{Z}, \mathbb{R})$, where $\mathbb{R}$ represents real numbers and $\mathbb{Z}$ represents integers. Additionally, we filter out results where the class index falls outside the valid range, which is between 0 and $|\mathcal{C}| - 1$.

Recently, Wake et al. [15] introduced the T-PIVOT method for TAL using GPT-4o. Due to the limited context length of GPT-4o, which cannot accommodate the densely sampled frames of long videos, T-PIVOT progressively narrows the search window over time. In contrast, our Gemini-based method can detect all action instances for the target categories in a single iteration, avoiding the need for multiple API calls.

A.4. Selection of Previous Methods for OV-TAL

Including the results of existing OV-TAL methods [5, 11, 12] in our proposed OV-TAL benchmarks would be a

You are an intelligent question answering agent. I will ask you question about temporal action localization and you must detect action instances in a video.
You will be shown a set of frames that have been extracted from a single video.
Given the video frames, you must answer the tuple of (start_time_of_action, end_time_of_action, category_index_of_action, confidence_score_of_action).
Note that the time is in second and rounded to the second decimal place, confidence score is the number between 0 and 1 and is rounded to the fourth decimal place, and the list of target action categories are provided below.
Here is the list of target action category and its correponding index.
[Target Action Category]
{action_cls_map}

You must output `text` to answer the detected action instances.
For your reference, we provide the example of output format with five detected action instances.
[Output]
(0.00, 6.50, 5, 0.7515), (5.88, 9.50, 3, 0.8805), (10.00, 15.25, 0, 0.9523), (12.50, 40.00, 11, 0.6789), (80.35, 123.50, 15, 0.8792)
Now frames of a video and its extracted temporal information are provided below.
{time_instructed_video_data}

Please detect all action instances of actions listed in [Target Action Category].
[Output]

{action_cls_map}

Action_classname_0: 0
Action_classname_1: 1
Action_classname_2: 2
…..

{time_instructed_video_data}

This video lasts for {video_time:.2f} second. | This frame is sampled at {t:.2f} second. |  Frame #1 | This frame is sampled at {t:.2f} second. |  Frame #2 | This frame is sampled at {t:.2f} second. |  Frame #3

Figure 1. **Gemini instruction template for TAL.**

good practice to ensure a sufficient comparison. However, due to the difficulty of reproducing their results[1], we instead choose OpenTAL [1], which focuses on localizing actions unseen during training through uncertainty modeling. We train it on the base categories defined in our OV-TAL benchmark and use it as the class-agnostic action localizer in our decoupled architecture. Although OpenTAL utilizes the I3D [3] backbone for action localization, we assign action classes to output action instances using ViFi-CLIP [14], as same as our model. Thus, comparing its mAPs with ours solely evaluates the capability of action localization.

In terms of EffPrompt [5], the details about the action localization model is not enough to reproduce by ourselves. As it also adopted decoupled architecture as ours, we use our action localizer, but replace the action classifier with the prompt-tuned VLM. Most of the training details of prompt tuning are borrowed from EffPrompt [5], but we empirically find the training iterations for training it on THOMOS14 [4] and FineAction [9].

[1] In Issue 1 and Issue 2, the author mentioned that the method exploits UNet [16] as the external action classifier which is trained on all action classes in the dataset.

A.5. Training Details

AdamW [10] is chosen as the optimizer, coupled with a scheduler that linearly warms up the learning rate (lr) to its maximum value and decays to the minimum value ($1e^{-8}$) following a cosine function. Tab. 1 presents the hyperparameter values for each dataset. For the OD-ST experiments on THUMOS14, we use the subset of 100k videos. We empirically found the threshold values for obtaining pseudo labels. 0.2, 0.05, and 0.4 are used for ActivityNet, THUMOS14, and FineAction, respectively.

B. Additional Experimental Results

B.1. ActivityNet Results for OV-TAL Benchmark

The main paper presents the cross-category OV-TAL benchmark results of the THUMOS14 [4] and FineAction [9] datasets in Tab. 4 (Main). Here, we show the results of the ActivityNet v1.3 [2] in Tab. 2. As discussed in Sec. 4.6, ActivityNet v1.3 is not a proper dataset for evaluating the generalization capability in action localization, supported by the zero-shot performance (w/o ST) of action localization on par with that of full-shot in Tab. 6 (Main). In the proposed OV-TAL benchmark, we observe

Dataset	Stage	max lr	warm-up epoch	main epoch	batch size
ActivityNet	First	$1e^{-5}$	5	10	16
	Second	$1e^{-5}$	5	5	16
THUMOS14	First	$1e^{-4}$	5	30	2
	Second (ID-ST)	$1e^{-4}$	5	10	2
FineAction	First	$1e^{-5}$	5	10	4
	Second (ID-ST)	$1e^{-5}$	5	5	4
THUMOS14 FineAction	Second (OD-ST)	$1e^{-5}$	2	2	4

Table 1. **Training hyperparameters values.**

Methods	Backbone	ActivityNet				
		Generalized			Constrained	
		mAP^{50}_A	mAP^{50}_B	mAP^{50}_N	mAP^{50}_B	mAP^{50}_N
OpenTAL† [1]	I3D [3]	32.7	36.4	29.7	39.1	34.1
STOV-TAL (w/o ST)	ViFi-B [14]	42.9	47.5	39.1	51.3	44.1
STOV-TAL (ID-ST)	ViFi-B [14]	43.1	47.3	39.7	51.1	44.6
STOV-TAL (FS)	ViFi-B [14]	43.7	47.9	40.3	52.0	45.3

Table 2. **Evaluation of cross-category OV-TAL benchmark**. For reference of upper-bound, full-shot results are shown in gray. † is our reproduced result.

a similar trend. For instance, in the constrained setting, mAP^{50}_N of w/o ST and FS are 44.1 and 45.3, respectively. Based on these results, we decided not to include the ActivityNet dataset in the OV-TAL benchmark since there is only a small room for improvement in cross-category generalization ability.

B.2. ZS-TAL Benchmark Full Results

Due to space constraints, we present only partial results of the ZS-TAL benchmark in Tab. 6 (Main). In Tab. 3, we provide the complete results, which complement the tIoU values of 0.4 and 0.6 for TH14, and 0.95 for ANET. These results exhibit a similar trend to those presented in the main paper. In the case of full-shot results (100% Seen 0% Unseen), other methods achieve higher mAP, which is attributed to the use of fine-tuned classifiers. These methods fine-tune the action classifiers on the target action categories, resulting in identical action categories during training and testing. In contrast, we keep freeze and do not fine-tune the VLM for the target actions, and ours perform better for zero-shot settings. Therefore, the 100% Seen 0% Unseen results do not reflect the generalization ability of action localizers.

References

[1] Wentao Bao, Qi Yu, and Yu Kong. Opental: Towards open set temporal action localization. In *CVPR*, pages 2979–2989, 2022. 2, 3

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015. 1, 2, 4

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 2, 3

[4] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *CVIU*, 155:1–23, 2017. 1, 2, 4

[5] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *ECCV*, pages 105–124. Springer, 2022. 1, 2, 4

[6] Chen Ju, Zeqian Li, Peisen Zhao, Ya Zhang, Xiaopeng Zhang, Qi Tian, Yanfeng Wang, and Weidi Xie. Multi-modal prompting for low-shot temporal action localization. *arXiv preprint arXiv:2303.11732*, 2023. 4

[7] Chuming Lin, Chengming Xu, Donghao Luo, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. Learning salient boundary feature for anchor-free temporal action localization. In *CVPR*, pages 3320–3329, 2021. 1

[8] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *ICCV*, pages 3889–3898, 2019. 1

[9] Yi Liu, Limin Wang, Yali Wang, Xiao Ma, and Yu Qiao. Fineaction: A fine-grained video dataset for temporal action localization. *IEEE TIP*, 31:6937–6950, 2022. 1, 2

[10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 2

[11] Sauradip Nag, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Zero-shot temporal action detection via vision-language prompting. In *ECCV*, pages 681–697. Springer, 2022. 1, 4

[12] Thinh Phan, Khoa Vo, Duy Le, Gianfranco Doretto, Donald Adjeroh, and Ngan Le. Zeetad: Adapting pretrained vision-language model for zero-shot end-to-end temporal action detection. In *WACV*, pages 7046–7055, 2024. 1

[13] 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 *ICML*, volume 139, pages 8748–8763. PMLR, 2021. 4

Evaluation Setting	Methods	Backbone	THUMOS14 [4]						ActivityNet v1.3 [2]			
			0.3	0.4	0.5	0.6	0.7	Avg.	0.5	0.75	0.95	Avg.
Full-shot 100% Seen 0% Unseen	ActionFormer† [20]	ViFi-CLIP-B [14]	72.8	67.4	57.3	45.2	29.7	54.5	49.5	31.2	4.3	30.9
	EffPrompt [5]	CLIP-B [13]	50.8	44.1	35.8	25.7	15.7	34.5	44.0	27.0	5.1	27.3
	STALE [11]	CLIP-B [13]	60.6	53.2	44.6	36.8	26.7	44.4	54.3	34.0	7.7	34.3
	UnLoc [19]	CLIP-B [13]	-	-	-	-	-	-	54.6	-	-	-
	UnLoc [19]	CLIP-L [13]	-	-	-	-	-	-	59.3	-	-	-
	STOV-TAL (FS)	CLIP-B [13]	47.5	41.7	33.0	24.7	15.4	32.5	37.5	23.1	1.8	22.7
	STOV-TAL (FS)	ViFi-CLIP-B [14]	65.3	60.6	50.2	39.0	26.0	48.2	43.7	26.8	2.0	26.4
Zero-shot 75% Seen 25% Unseen	EffPrompt [5]	CLIP-B [13]	39.7	31.6	23.0	14.9	7.5	23.3	37.6	22.9	3.8	23.1
	STALE [11]	CLIP-B [13]	40.5	32.3	23.5	15.3	7.6	23.8	38.2	25.2	**6.0**	24.9
	UnLoc [19]	CLIP-B [13]	-	-	-	-	-	-	40.2	-	-	-
	UnLoc [19]	CLIP-L [13]	-	-	-	-	-	-	48.8	-	-	-
	Ju *et al*. [6]	CLIP-B [13]	46.3	39.0	29.5	18.3	8.7	28.4	42.0	25.8	3.2	25.9
	STOV-TAL (w/o ST)	CLIP-B [13]	47.8	39.1	28.4	17.6	9.1	28.4	47.0	28.1	1.6	27.9
	STOV-TAL (w/o ST)	ViFi-CLIP-B [14]	56.7	47.2	34.3	22.8	11.3	34.5	51.7	**30.9**	1.8	**30.5**
	STOV-TAL (ID-ST)	ViFi-CLIP-B [14]	**59.5**	**50.2**	**37.5**	**24.6**	**12.5**	**36.9**	**52.0**	30.6	1.2	30.1
	STOV-TAL (OD-ST)	ViFi-CLIP-B [14]	58.2	48.2	35.1	23.0	11.8	35.2	-	-	-	-
	STOV-TAL (FS)	ViFi-CLIP-B [14]	67.5	60.8	47.7	34.8	21.8	46.5	52.6	31.5	2.3	31.3
Zero-shot 50% Seen 50% Unseen	EffPrompt [5]	CLIP-B [13]	37.2	29.6	21.6	14.0	7.2	21.9	32.0	19.3	2.9	19.6
	STALE [11]	CLIP-B [13]	38.3	30.7	21.2	13.8	7.0	22.2	32.1	20.7	**5.9**	20.5
	UnLoc [19]	CLIP-B [13]	-	-	-	-	-	-	36.9	-	-	-
	UnLoc [19]	CLIP-L [13]	-	-	-	-	-	-	43.7	-	-	-
	Ju *et al*. [6]	CLIP-B [13]	42.3	34.7	25.8	16.2	7.5	25.3	34.3	20.8	3.0	21.0
	STOV-TAL (w/o ST)	CLIP-B [13]	44.2	35.7	25.7	16.5	8.0	26.0	42.1	25.0	1.3	24.8
	STOV-TAL (w/o ST)	ViFi-CLIP-B [14]	53.4	43.1	31.3	19.7	9.8	31.5	48.1	28.4	1.3	**28.0**
	STOV-TAL (ID-ST)	ViFi-CLIP-B [14]	**56.3**	**46.1**	**34.4**	**21.9**	**11.3**	**34.0**	**48.4**	**28.7**	0.8	27.9
	STOV-TAL (OD-ST)	ViFi-CLIP-B [14]	54.3	43.7	32.5	21.4	10.6	32.5	-	-	-	-
	STOV-TAL (FS)	ViFi-CLIP-B [14]	68.2	62.5	50.7	38.2	24.6	48.8	49.4	29.9	2.2	29.6

Table 3. **Evaluation of ZS-TAL benchmark.** The results are based on RGB only without optical flow. In each setting, the best for each metric is bolded. Full-shot results are shown in gray for reference of upper-bound. † indicates our reproduced results. The values not provided are filled by "-".

[14] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *CVPR*, pages 6545–6554, 2023. 1, 2, 3, 4

[15] Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi. Open-vocabulary temporal action localization using vlms. *arXiv preprint arXiv:2408.17422*, 2024. 1

[16] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *CVPR*, pages 4325–4334, 2017. 2

[17] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 1

[18] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *CVPR*, pages 10156–10165, 2020. 1

[19] Shen Yan, Xuehan Xiong, Arsha Nagrani, Anurag Arnab, Zhonghao Wang, Weina Ge, David Ross, and Cordelia Schmid. Unloc: A unified framework for video localization tasks. In *ICCV*, pages 13623–13633, 2023. 4

[20] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *ECCV*, pages 492–510. Springer, 2022. 1, 4