Test-Time Zero-Shot Temporal Action Localization
Supplementary Material

In this Supplementary Material, we provide additional quantitative and qualitative results of the proposed $T3AL$. In Sec. A we provide details on the preliminary experiment reported in the main manuscript, in Sec. B we discuss per-class results of $T3AL$, and in Sec. C we show captions generated by the model. The supplementary material is also accompanied by qualitative results in video format that are easily accessible at https://github.com/benedettaliberatori/T3AL. These videos can better aid in understanding the results presented in the paper.

A. Cross-dataset generalization analysis

In the experiment reported in Sec. 3 of the main manuscript we consider two state-of-the-art Zero-Shot Temporal Action Localization (ZS-TAL) methods [1, 4] that, to the best of our knowledge, are the only works with publicly available code.

For STALE [4] we use the model pre-trained on the ActivityNet-v1.3 dataset for the ZS-TAL task. For EffPrompt [1], which does not provide models pre-trained on ZS-TAL datasets, we use a model pre-trained on HMDB51 [2] for the video action recognition task. EffPrompt is a two-stage method, *i.e.*, it first detects region proposals and then classifies the obtained regions. For this reason, we employ the same action localizer [3] utilized in its first stage to generate action proposals from untrimmed videos, and then use the model pre-trained on trimmed videos to classify the obtained regions. The proposal detector is trained on the original training set of THUMOS14. We use the model pre-trained on THUMOS14 as it is the only one available in the official repository. Consequently, we evaluate its performance for each split using videos from the original test set. The results obtained for this out-of-distribution experiment are reported in Tab. 1, alongside the in-distribution numbers, *i.e.*, models trained and tested on THUMOS14.

STALE trained on ActivityNet-v1.3 is suboptimal when tested on out-of-distribution data. We attribute this performance reduction to different datasets characteristics, as the model trained on ActivityNet-v1.3 learns to predict fewer and longer proposals, but for THUMOS14 regions are generally sparser and shorter. Also EffPrompt shows significantly lower results when evaluated on THUMOS14, despite the pre-training of the proposal detector on the out-of-distribution dataset. This experiment shows that the model is unable to generalize from HMDB51 to THUMOS14 classes.

Model	Training Data	Split	mAP (%) ↑					
			0.3	0.4	0.5	0.6	0.7	Avg.
EffPrompt	THUMOS14	75:25	39.7	31.6	23.0	14.9	7.5	23.3
EffPrompt	THUMOS14	50:50	37.2	29.6	21.6	14.0	7.2	21.9
STALE	THUMOS14	75:25	40.5	32.3	23.5	15.3	7.6	23.8
STALE	THUMOS14	50:50	38.3	30.7	21.2	13.8	7.0	22.2
EffPrompt	HMDB51	75:25	7.1	5.9	4.5	3.4	2.2	4.6
EffPrompt	HMDB51	50:50	5.4	4.4	3.5	2.7	1.9	3.6
STALE	ActivityNet-v1.3	75:25	0.5	0.3	0.2	0.2	0.2	0.3
STALE	ActivityNet-v1.3	50:50	1.3	0.7	0.6	0.6	0.4	0.7

Table 1. **Cross-dataset generalization.** We show the average mAP, computed at IoU thresholds of [0.3:0.1:0.7], for EffPrompt and STALE trained and tested on THUMOS14, and trained on a different dataset and tested on THUMOS14. We report results for the 75:25 (75% seen classes) and 50:50 (50% seen classes) evaluation settings.

B. Experiments

We report per-class results of $T3AL$ on THUMOS14 for both the evaluation settings, *i.e.*, 50%-50% split in Tab. 2 and 75%-25% split in Tab. 3. Note that the latter contains only 18 of the total 20 classes as the labels *Basketball dunk* and *Long jump* are not contained in any of the test splits for the 75%-25% setting. Following [1], the results are the averages of the individual results obtained across all class splits. Both tables show high variance in performance among the classes. In particular, classes that have less in common with the surrounding scene (*e.g.*, *Clean and jerk*, *Pole vault*, and *Long jump*) exhibit considerably higher results (*e.g.*, 23.8%, 24.7%, and 31.9% avg. mAP on 50:50) compared to classes that share more visual cues with the surrounding context, as observed for *Tennis swing* or *Billiards* (*i.e.*, 1.5%, 2.9% avg. mAP on 50:50). We attribute the fact that the model underperforms on videos of class *Tennis swing* to the atomicity of the action: the swing movement bears a subtle difference from a person with a tennis racket in hand who is not actively swinging but is poised and waiting for the ball. Billiards, instead, serves as an example of an action class that is not atomic but rather encompasses a broad range of

Class Name	mAP (%) ↑					
	0.3	0.4	0.5	0.6	0.7	**Avg.**
BaseballPitch	13.2	9.4	4.4	2.5	1.6	6.2
BasketballDunk	23.5	13.7	7.7	4.1	1.4	10.1
Billiards	6.9	4.0	2.2	1.2	0.2	2.9
CleanAndJerk	43.4	31.8	22.8	14.1	6.8	23.8
CliffDiving	33.4	23.4	14.5	9.1	4.7	17.0
CricketBowling	7.6	2.6	0.9	0.3	0.1	2.3
CricketShot	7.1	3.3	1.2	0.5	0.2	2.5
Diving	23.9	17.8	11.7	6.2	2.8	12.5
FrisbeeCatch	8.2	4.0	1.7	0.7	0.4	3.0
GolfSwing	18.1	10.6	3.6	1.3	0.8	6.9
HammerThrow	34.9	30.8	23.3	15.4	10.7	23.0
HighJump	30.3	20.3	12.2	5.9	2.9	14.3
JavelinThrow	29.2	21.0	13.8	8.3	4.6	15.4
LongJump	51.2	42.6	32.5	21.6	11.4	31.9
PoleVault	42.0	33.3	23.9	16.6	7.9	24.7
Shotput	17.1	12.2	8.0	4.8	2.8	9.0
SoccerPenalty	26.7	14.0	6.7	3.0	0.8	10.3
TennisSwing	3.8	2.0	1.0	0.5	0.1	1.5
ThrowDiscus	4.8	3.5	2.1	1.7	0.7	2.6
VolleyballSpiking	19.6	14.3	7.7	4.1	2.0	9.5

Table 2. **Per-class results on THUMOS14 (50%-50%).** Numbers are computed at IoU thresholds of [0.3:0.1:0.7] and averaged across all class splits.

Class Name	mAP (%) ↑					
	0.3	0.4	0.5	0.6	0.7	**Avg.**
BaseballPitch	12.2	7.6	2.5	1.8	1.5	5.1
Billiards	2.0	1.4	0.3	0.2	0.0	0.8
CleanAndJerk	29.0	20.0	11.4	5.8	3.6	14.0
CliffDiving	37.3	25.8	16.5	10.6	5.1	19.0
CricketBowling	7.6	2.5	1.1	0.4	0.1	2.3
CricketShot	6.0	2.8	0.9	0.4	0.1	2.0
Diving	23.8	18.0	11.6	7.0	3.4	12.8
FrisbeeCatch	5.2	2.6	1.2	0.2	0.1	1.9
GolfSwing	15.8	9.8	2.5	1.3	0.9	6.1
HammerThrow	41.2	34.6	25.2	16.4	11.0	25.7
HighJump	32.7	20.5	12.5	5.7	2.2	14.7
JavelinThrow	25.0	17.2	11.3	7.6	3.4	12.9
PoleVault	50.2	37.4	25.5	18.1	8.2	27.9
Shotput	17.0	9.0	5.0	2.6	1.8	7.1
SoccerPenalty	26.9	15.4	7.1	2.3	1.0	10.5
TennisSwing	3.4	2.0	1.0	0.4	0.1	1.4
ThrowDiscus	3.3	1.8	1.2	0.7	0.2	1.5
VolleyballSpiking	19.2	11.5	5.7	2.7	1.4	8.1

Table 3. **Per-class results on THUMOS14 (75%-25%).** Numbers are computed at IoU thresholds of [0.3:0.1:0.7] and averaged across all class splits.

potential movements, *e.g.*, holding the billiard cue, striking the ball, or preparing the billiard table. The classes of the datasets contain a mixture of action verbs, nouns describing actions, and activities. The lack of a well-defined taxonomy poses a challenge for TAL methods, as explained in the main manuscript in Sec. 6.

C. Qualitative Results

In this section, we show some of the captions generated with CoCa [5] on THUMOS14. It can be seen that captions generated from frames within ground truth regions often contain the ground truth class. Moreover, there are instances where captions contain words related to the annotated class, even when the action is not depicted in the frame, *e.g.*, Fig. 2 containing the word *"diving"* when the individuals in the scene are stationary on the diving board and not engaged in the actual action of diving, or *"pole vaulting"* in Fig. 1 related to a static scene without the performed action. Certain captions may contain words associated with classes different from the ground truth, as illustrated by the example in Fig. 3 where the word *"frisbee"* is present. In this case, the caption shares more semantics with *Frisbee catch* than with the ground truth *Shot put*. There are also cases where words related to the captions (*e.g.*, *"pool"* for the action *Billiards*) are present in captions of images that may or may not depict the action happening, as shown in Fig. 4. In the case of *Soccer penalty*, the word *"penalty"* is not present in any caption, but the term *"soccer"* is consistently contained in most of them, as shown in Fig. 5.

References

[1] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *ECCV*, 2021. 1

[2] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *ICCV*, 2011. 1

[3] 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*, 2021. 1

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

[5] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv*, 2022. 2