Supplementary materials for Boosting Weakly-Supervised Temporal Action Localization with Text Information

A. The Proposed Method

A.1. Modal Modulation Strategy

Similar as CCM proposed in [1], in this paper we design a modal modulation strategy to fuse RGB and Flow feature for downstream task. Specifically, an auto-encoder is used to obtain intermediate modal features of RGB and FlOW features. Then, we calculate the similarity between the intermediate modal features and the global RGB features, and modulate it into Flow feature with sigmoid function as [1] to get augmented Flow features. Similiarly, we also use the same method to augmented RGB features. Finally the fused RGB and Flow features are concatenated in the channel dimension to generate the fused video features.

A.2. Video-Text Language Completion

Following [2], when using contrastive loss $\mathcal{L}_c$ in video-text language model, the reconstructor loss $\mathcal{L}_{rec}$ is first used to update the transformer reconstructor while freezing the attention mechanism. Then the contrastive loss $\mathcal{L}_c$ is used to update attention mechanism while freezing the transformer reconstructor.

Table 1. Comparisons with different λ in The proposed framework on THUMOS14 dataset.

λ	mAP@IoU(%)			Avg
	0.3	0.5	0.7	
0.25	55.0	37.8	13.7	35.8
0.5	55.6	38.2	14.5	36.5
0.75	55.1	38.1	14.3	36.1
1.0	55.6	38.8	14.9	36.7
1.25	54.5	39.1	15.2	36.6
1.5	**56.2**	**39.3**	**15.2**	**37.2**
1.75	55.0	39.2	14.9	36.8
2.0	54.5	38.1	14.6	36.3

B. Experiments

B.1. Hyper-parameter sensitivity analysis.

There are six Hyper-parameter in the proposed framework: α,β,λ is used to balance the final loss function; γ_1 and γ_2 is the thresholds in the $\mathcal{L}_c$ and the length of the learnable prompt in TSM. We tested the sensitivity of these hyper-parametric design ablation experiments respectively. Figure 1 illustrates the model performance in terms of average mAP with varying values of α from 0.5 to 1.5 on THUMOS14 dataset. We can clearly see that the model could obtain best performance when α = 1.0.

Table 2. Comparisons with different length of prompts in TSM on THUMOS14 dataset.

Length	mAP@IoU(%)			Avg
	0.3	0.5	0.7	
6	54.7	38.8	14.3	36.2
7	54.9	37.7	14.3	35.8
8	55.2	38.4	14.8	36.4
9	**56.2**	**39.3**	**15.2**	**37.2**
10	56.6	38.8	14.8	36.7
11	54.9	38.3	14.4	36.1
12	55.3	38.6	13.8	36.0

Figure 1 illustrates the model performance in terms of average mAP with varying values of β from 0.5 to 1.5 on THUMOS14 dataset. We can clearly see that the model could obtain best performance when β = 1.0.

Table 1 illustrates the model performance in terms of average mAP with varying values of λ from 0.0 to 2.0 on THUMOS14 dataset. We can clearly see that the model could obtain best performance when λ = 1.5.

Figure 2 illustrates the model performance in terms of average mAP with varying values of γ_1 from 0.0 to 0.25 on THUMOS14 dataset. We can clearly see that the model could obtain best performance when γ_1 = 0.1.

Figure 2 illustrates the model performance in terms of average mAP with varying values of γ_2 from 0.0 to 0.25 on THUMOS14 dataset. We can clearly see that the model could obtain best performance when γ_2 = 0.15. Table 2

Table 3. Comparisons with different prompts in TSM on THUMOS14 dataset.

Method	mAP@IoU(%)			Avg
	0.3	0.5	0.7	
a [CLS]	55.1	38.1	14.1	36.1
a [CLS] video	54.9	38.4	13.9	36.0
a video of [CLS]	55.4	38.3	14.6	36.4
a video of action [CLS]	55.1	38.4	15.0	36.5
learnable prompt	**56.2**	**39.3**	**15.2**	**37.2**

illustrates the model performance in terms of average mAP with varying values of the length of the learnable prompt in TSM from 3 to 15 on THUMOS14 dataset. We can clearly see that the model could obtain best performance when the length is set as 9.

Table 4. Ablation studies of addition losses on THUMOS14.

$\mathcal{L}_{mil}$	$\mathcal{L}_{gui}$	$\mathcal{L}_{nor}$	$\mathcal{L}_{coa}$	TSM & VLC	Avg 0.3:0.7	Avg 0.1:0.7
✓	×	×	×	×	31.7	40.2
✓	✓	×	×	×	32.7	41.3
✓	✓	✓	×	×	33.8	42.6
✓	✓	✓	✓	×	34.9	43.8
✓	×	×	×	✓	32.6	40.9
✓	✓	×	×	✓	33.5	42.0
✓	✓	✓	×	✓	35.3	44.1
✓	✓	✓	✓	✓	37.2	46.0

The guide loss $\mathcal{L}_{gui}$, co-activity loss $\mathcal{L}_{coa}$ and normalization loss $\mathcal{L}_{nor}$ have been widely used in existing WTAL methods to improve model performance, and they are not the main contribution of this paper. We show the improvement of each additional loss in Table 4. As shown in Table 4, each additional loss can improve the performance of the baseline. Besides, the proposed TSM and VLC can improve the baseline performance in any case.

B.2. Comparisons with different prompts in text-segment mining model.

We further compare different handcraft prompt templates with learnable prompt template for generating text queries in TSM. As show in Table 3, compared with handcraft prompt templates, learnable prompt template obtain a better performance on TUHMOS14, verifying the effectiveness of adopting learnable prompt.

References

[1] Fa-Ting Hong, Jia-Chang Feng, Dan Xu, Ying Shan, and Wei-Shi Zheng. Cross-modal consensus network for weakly supervised temporal action localization. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1591–1599, 2021. 1

[2] Minghang Zheng, Yanjie Huang, Qingchao Chen, and Yang Liu. Weakly supervised video moment localization with contrastive negative sample mining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 1, page 3, 2022. 1

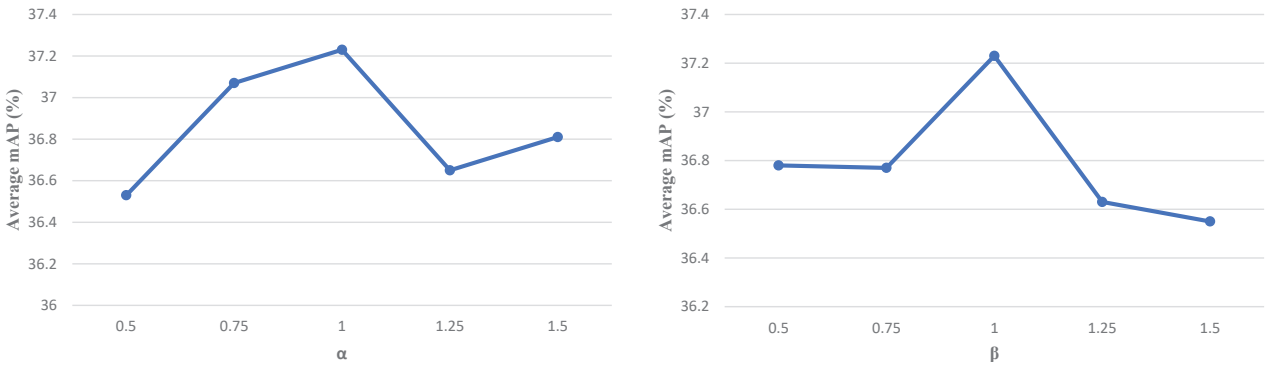

Figure 1. Average mAP with varying values of α and β on THUMOS14.

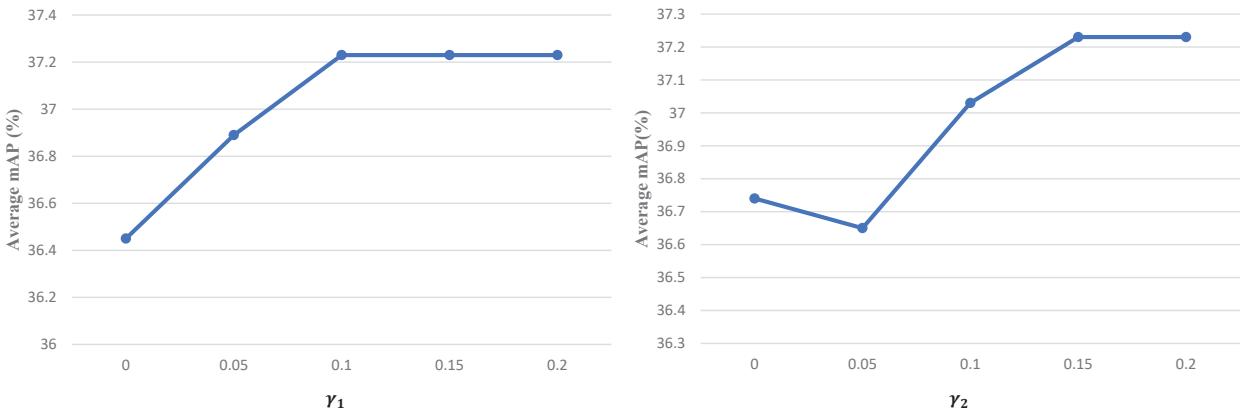

Figure 2. Average mAP with varying values of γ_1 and γ_2 on THUMOS14.