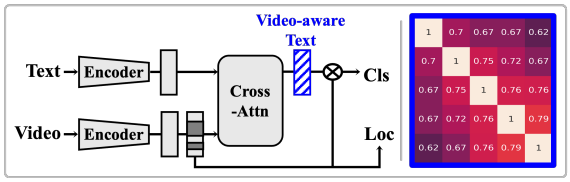
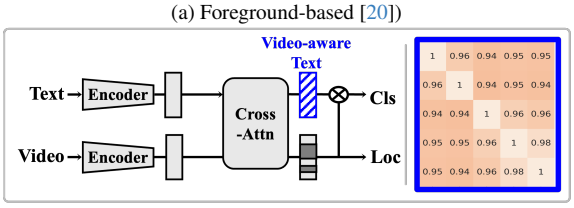
TF-CADE: Foreground-Concentrated Text-Video Alignment for Zero-Shot Temporal Action Detection

Supplementary Material



(a) Foreground-based [20])



(b) Foreground-free ([13])

Figure A. **Comparison of existing approaches**: (a) foreground-based and (b) foreground-free cross-modal text updating methods. The right column shows cosine similarity heatmaps of the updated video-aware text features across action classes on ActivityNet v1.3. [5].

A. Comparison of Structure

Fig. A illustrates a comparison between foreground-based and foreground-free architectures. As discussed in Sec. 1, foreground-based methods extract foreground regions solely from the video stream without conditioning on text. These text-unrelated foreground regions limit the model's ability to utilize textual cues in the detection process. In contrast, foreground-free methods eliminate proposal extraction and instead update the text through cross-attention over the entire video sequence. While the foreground-free methods avoid relying on the pre-extracted proposals, these methods have a different problem: because untrimmed videos contain a substantial amount of background, the updated video-aware text becomes dominated by background signals. As shown in Fig. A (b), the resulting video-aware text exhibits uniformly higher inter-class similarity than in Fig. A (a). This indicates that aligning text with video features containing background regions causes the model to fail to encode class-specific textual cues.

B. Additional Performance Comparison

Evaluation under Ti-FAD. Table A presents results obtained by integrating our method into Ti-FAD. In the in-distribution setting, performance remains similar when using external classification scores, as localization is relatively easy and thus largely dominate detection performance. In contrast, in the cross-dataset setting, our methods consistently improves the performance.

Table A. **Performance comparison under the original Ti-FAD.**

External info.	Model	In-distribution								Cross-dataset generalization			
		THUMOS14				ActivityNet v1.3				THUMOS14			
		0.3	0.5	0.7	Avg.	0.5	0.75	0.95	Avg.	0.3	0.5	0.7	Avg.
✓	Ti-FAD [12]	57.0	43.3	**21.2**	**41.2**	50.6	**32.2**	**5.2**	32.0	29.9	16.0	3.1	16.2
	Ti-FAD [12] + Ours	**57.8**	**43.6**	20.1	**41.2**	**51.2**	32.1	5.0	**32.1**	**41.0**	**26.1**	**8.2**	**25.3**
✗	Ti-FAD [12]	20.4	17.0	9.6	16.0	10.6	7.8	**1.9**	7.4	17.9	11.1	4.0	11.1
	Ti-FAD [12] + Ours	**24.9**	**20.1**	**10.8**	**18.9**	**14.7**	**9.2**	1.4	**9.2**	**36.6**	**23.9**	**7.5**	**22.9**

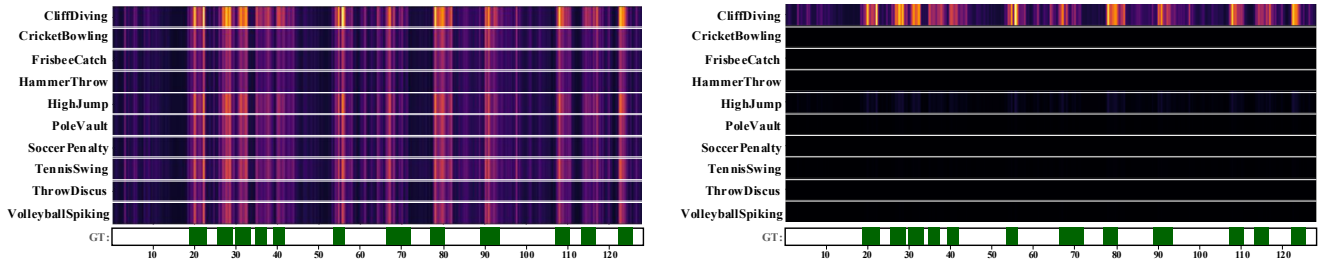
Table B. **Component analysis under cross-dataset generalization setting.** The models are trained on ActivityNet v1.3 and evaluated on THUMOS14.

Row	Method	m_{max}	m_{filter}	mAP@tIoU (%)			
				0.3	0.5	0.7	Avg.
1	Baseline			24.5	14.5	4.7	14.5
2	+$\mathcal{L}_{video}$	✓		23.0	14.9	5.3	14.5
3			✓	24.9	16.4	5.7	15.8
4		✓	✓	29.5	19.3	7.0	18.6
5	+CCR	✓		25.6	16.3	5.8	15.9
6			✓	25.6	16.3	5.8	15.9
7		✓	✓	25.6	16.3	5.8	15.9
8	+$\mathcal{L}_{video}$ & CCR	✓		35.7	21.7	6.7	21.3
9			✓	35.9	23.4	7.6	22.6
10		✓	✓	**42.8**	**28.6**	**10.1**	**27.4**

Additional Component Analysis. To further examine the contribution of each component in our framework, Table B provides an extended component analysis under the cross-dataset generalization setting. In the cross-dataset setting, using only the peak-based certainty m_{max} with $\mathcal{L}_{video}$ leads to limited improvements, as raw similarity peaks are sensitive to background biases under domain shift, as shown in Row 2. In contrast, applying the smoothing filter m_{filter} consistently improves performance, indicating that temporally coherent foreground estimation is critical for stable alignment. CCR alone (Rows 5–7) shows nearly identical performance across different certainty maps, indicating the need for guidance in handling noisy or fragmented foreground modeling. When $\mathcal{L}_{video}$ and CCR are jointly enabled, the effect of the smoothing filter becomes significantly amplified. In particular, combining m_{max} and m_{filter} achieves the best performance. These results indicate that the smoothing filter stabilizes foreground estimation and that $\mathcal{L}_{video}$ and CCR are complementary.

C. Further Analysis on CCR

Effect of CCR on Confidence Score Map. To better understand the effect of CCR, we visualize the video-text confi-



(a) w/o CCR (P_{cls}) (b) w/ CCR ($\tilde{P}_{\text{cls}}$)

Figure B. **Comparison of heatmaps without and with CCR.** (a) the original classification confidence P_{cls} and (b) the re-weighted confidence $\tilde{P}_{\text{cls}}$ with CCR. The heatmaps show classification confidence across time (horizontal axis) and action classes (vertical axis), with color intensity indicating confidence scores. The green bars below each plot indicate the ground-truth (*"CliffDiving"*) action segments.

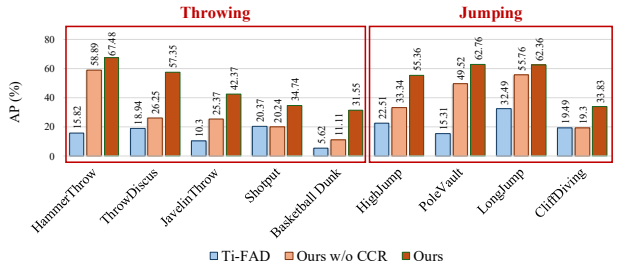


Figure C. **Per-class AP (%) at tIoU=0.5 under the cross-dataset generalization setting.** Compared to Ti-FAD [13] and our model without CCR, our full model (Ours) consistently achieves higher AP across semantically similar action categories. The models are trained on ActivityNet v1.3 and evaluated on THUMOS14.

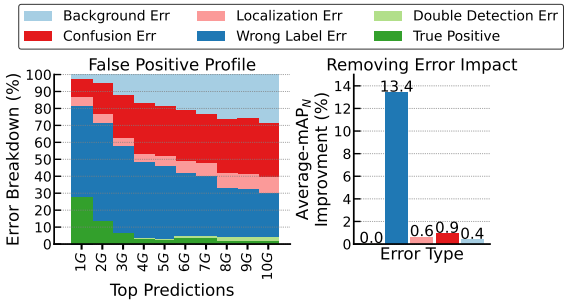


Figure D. **False positive analysis without applying CCR on THUMOS14 using DETAD [2].** Without applying CCR, the proportion of Wrong Label Error increase compared to our full in model Fig. 4.

dence score map with and without applying the re-weighted confidence during inference step. As shown in Fig. B, using the original classification confidence P_{cls} produces scattered and noisy activations across multiple classes, failing to clearly distinguish the target action. After applying CCR, the re-weighted confidence score map $\tilde{P}_{\text{cls}}$ becomes more compact and aligned with the ground-truth class, allowing the model to more reliably focus on the relevant action segments. This result indicates that our re-weighting strategy suppresses uncertain predictions while enhancing discriminative regions for classification.

Per-class Analysis under Cross-Dataset Setting. To further validate the effectiveness of CCR in enhancing inter-class discrimination, we conduct a semantic group-level analysis on semantically similar action classes that commonly tend to confusion in zero-shot inference step. This analysis is conducted under the cross-dataset setting, which presents a more challenging scenario for generalization and discrimination. We group the unseen classes in THUMOS14 [8] into two semantic clusters: *Throwing* and *Jumping*. These groups are chosen based on textual and motion-level similarity, where previous models often struggle due to class ambiguity (e.g., *"JavelinThrow"* vs *"ThrowDiscus"*). Fig. C shows per-class Average Precision (AP) at tIoU=0.5. Across both the *Throwing* and *Jumping* categories, our method outperforms Ti-FAD [13] and the variant without CCR. This result indicates that our CCR effectively re-weights the original classification confidence by applying video-level class prior and alleviates inter-class confusion, especially under domain shift.

Table C. **Additional analysis of weighting scheme in CCR.**

Type	mAP@tIoU (%)			
	0.3	0.5	0.7	Avg.
$\text{sigmoid}(P_{cls}) + \text{softmax}(S_{\text{fg}})$	22.1	17.2	8.0	16.0
$\text{sigmoid}(P_{cls}) + \text{sigmoid}(S_{\text{fg}})$	22.5	17.1	7.7	16.0
$\text{sigmoid}(P_{cls}) \odot \text{softmax}(S_{\text{fg}})$	27.3	21.7	10.3	20.2
$\text{sigmoid}(P_{cls}) \odot \text{sigmoid}(S_{\text{fg}})$	26.4	21.2	10.7	19.8
$\sqrt{\text{sigmoid}(P_{cls}) \odot \text{softmax}(S_{\text{fg}})}$	**28.6**	**22.6**	**10.7**	**21.1**
$\sqrt{\text{sigmoid}(P_{cls}) \odot \text{sigmoid}(S_{\text{fg}})}$	28.1	22.1	10.3	20.6

False Positive Analysis w/o CCR. To further show the impact of CCR, we present an additional false positive analysis when CCR is removed. Fig. D illustrates that the proportion of Wrong Label Errors increases compared to our full model in Fig. 4 (b). This result indicates that CCR reduces misclassification by re-weighting the original per-snippet classification confidence and suppressing irrelevant action classes during inference.

Hyperparameter Analysis. Table C reports the performance of different weighting schemes in CCR. The results indicate that square-root reweighting consistently achieves better performance than alternative weighting schemes.

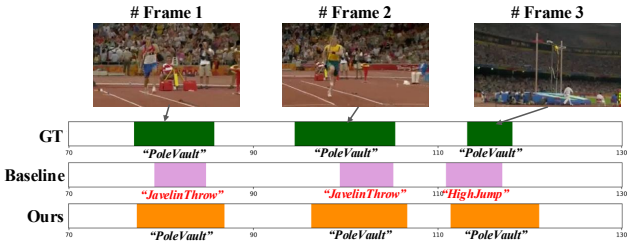


Figure E. **Qualitative results under cross-dataset generalization setting.** Each bar represents the predicted action segment, and the labels below indicate the predicted action categories for the baseline and our method. The models are trained on ActivityNet v1.3 and evaluated on THUMOS14.

D. Further Analysis on ACA

Qualitative Results. Fig. E illustrates the temporal action detection results of the baseline and our approach for ground-truth action *"PoleVault"* under the cross-dataset generalization setting. The prediction result of the baseline relying solely on snippet-level classification shows incorrect predictions due to visual ambiguity. For instance, although the ground-truth class is *"PoleVault"*, Frames 1 and 2 are incorrectly predicted as *"JavelinThrow"* because the athlete visually appears to hold a javelin-like object. Similarly, Frame 3 is misclassified as *"HighJump"* since the hurdle-clearing motion closely resembles a high-jump action. These examples demonstrate that relying solely on snippet-level predictions is insufficient and that video-level or segment-level context is critical for robust recognition. Our method addresses this limitation by explicitly learning temporally coherent foreground regions through ACA with $\mathcal{L}_{video}$.

E. Analysis on Computational Complexity

Table D presents the computational cost analysis including the number of parameters, inference time, and peak memory usage, comparing TF-CADE with state-of-the-art methods on THUMOS14 [8] and ActivityNet v1.3 [5]. We exclude inference time and peak memory usage for STALE [20] on THUMOS14, as the official implementation is not publicly available. For a fair comparison, inference time is measured per video across all models on a single NVIDIA A100 GPU. We utilize 10 random splits of the test classes and report the averaged results in the 50%-50% setting. TF-CADE has a comparable model size to Ti-FAD [13] and STALE [20], with a slightly higher inference time observed only on THUMOS14, but exhibiting faster inference on ActivityNet v1.3.

F. Additional Qualitative Results

Fig. F compares the temporal action detection results of the baseline and our approach for two actions under the cross-dataset generalization setting. The baseline model often fails to accurately localize short action segments, showing limited ability to distinguish between foreground and background frames. In contrast, our method captures these short actions more precisely, demonstrating more accurate boundary localization and stronger discrimination against background motion. These visual results verify that our framework effectively enhances temporal precision and reduces false positives in challenging scenarios.

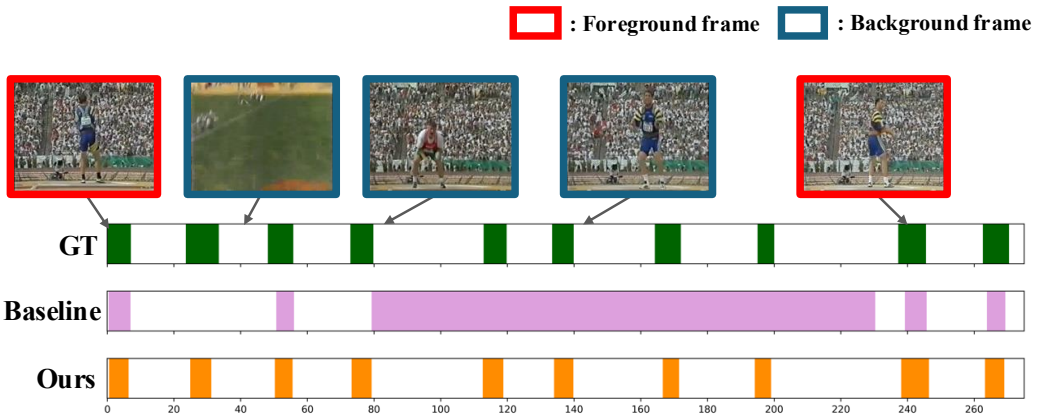
Table D. **Computational cost analysis in the 50%-50% setting under the in-distribution.**

Method	Parameters (M)	Inference Time (ms)	Peak Memory (GB)	Average mAP (%)
STALE [20]	170.72	–	–	–
Ti-FAD [13]	168.33	129.01 ± 3.77	3.91	16.0
TF-CADE	167.95	151.78 ± 1.79	3.23	21.1

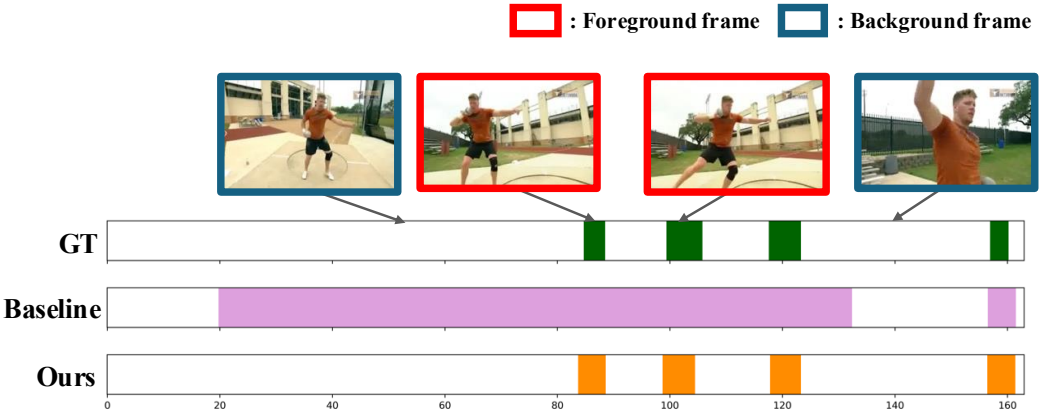
(a) Evaluation on THUMOS14 [8].

Method	Parameters (M)	Inference Time (ms)	Peak Memory (GB)	Average mAP (%)
STALE [20]	170.72	44.65 ± 1.18	2.53	4.4
Ti-FAD [13]	168.33	133.21 ± 1.42	3.49	7.4
TF-CADE	167.95	126.43 ± 2.78	2.81	10.5

(b) Evaluation on ActivityNet v1.3 [5].



(a) Ground-truth action: "HammerThrow"



(b) Ground-truth action: "Shotput"

Figure F. **Qualitative results under cross-dataset generalization setting.** Each bar represents the predicted action segments for the baseline and our method. The models are trained on ActivityNet v1.3 and evaluated on THUMOS14.