Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence
Canadian AI 2026
25–29 May 2026, Vancouver, BC

ConTrans: Learning Text-enhanced Local–global Temporal Representations for Zero-shot Temporal Action Localization

Kanchan Keisham[†*], Thenukan Pathmanathan[‡] Thangarajah Akilan[‡*]
[†] Vellore Institute of Technology, Tamil Nadu, India
[‡] Lakehead University, Thunder Bay, Canada

Abstract

Zero-shot Temporal Action Localization (ZS-TAL) aims to detect and locate previously unseen actions in untrimmed videos. However, existing approaches primarily focus on modeling long-range contextual information, often neglecting the critical relative-offset-based local correlations between video frames. Furthermore, their performance is hindered by limited feature representation capabilities due to the shallow nature of their network architectures. In this paper, we address these limitations by introducing a novel local-global multi-scale feature representation module. We propose a novel multi-scale encoder architecture, termed `ConTrans`, that integrates convolutional (Conv) inductive biases with transformer Self-attention to jointly capture fine-grained local dependencies and long-range global context, leading to more comprehensive feature representations than existing methods. Experimental evaluations on the ActivityNet-1.3 and THUMOS14 datasets demonstrate that `ConTrans` significantly outperforms existing methods, establishing a new benchmark for ZS-TAL.

Keywords: Cross-modal representation learning, hierarchical feature representation, temporal action localization, zero-shot learning.

1. Introduction

Temporal action localization (TAL) aims to detect and classify actions in long, untrimmed videos. Most existing TAL approaches [1, 2] operate under a closed-set assumption, where training and inference share the same action categories. However, real-world applications such as video retrieval and anomaly detection require recognizing unseen actions, motivating zero-shot temporal action localization (ZS-TAL). ZS-TAL enables models to localize actions from novel categories without labeled training examples. Large-scale vision–language (ViL) models, including CLIP [3], ALIGN [4], UniCL [5], and ZIM [6], have demonstrated strong zero-shot generalization by aligning visual and textual representations using large-scale web data. Beyond image understanding [7, 8], these models have significantly improved robustness and generalization in video tasks such as action recognition [9, 10], captioning [11, 12], and object tracking [13]. zero-shot inference is typically performed by measuring similarity between visual features and text embeddings representing semantic queries, making ViL representations promising for ZS-TAL.

Recently, zero-shot temporal action detection (ZS-TAD) has gained attention as a related task. STALE [14] is an early and influential method that aligns visual and textual embeddings within a one-stage detection framework, preserving foreground information via representation masking and improving classification through text prompt tuning (TPT). Subsequent work [15] further explores multi-modal prompt learning to adapt CLIP for temporal detection while reducing computational cost by pretraining prompts on image datasets and freezing them during TAD training.

Despite this progress, existing ZS-TAD methods struggle to capture fine-grained temporal cues and lack effective modeling of contextual interactions between visual and textual modalities, leading to imprecise action boundaries and limited robustness in complex videos. To address these limitations, we propose `ConTrans`, a novel framework that enhances zero-shot

[*]kanchan.keisham@vit.ac.in, takilan@lakeheadu.ca

This article is © 2026 by author(s) as listed above. The article is licensed under a Creative Commons Attribution (CC BY 4.0) International license (https://creativecommons.org/licenses/by/4.0/legalcode), except where otherwise indicated with respect to particular material included in the article. The article should be attributed to the author(s) identified above.

temporal action localization through Text-enhanced temporal representations. `ConTrans` integrates semantic textual information with visual features via multi-scale fusion and attention mechanisms, capturing both global and local context across modalities. This design improves sensitivity to subtle temporal variations and enables more accurate localization and recognition of unseen actions in open-set video scenarios.

In summary, the main contributions of this work are as follows:

- **A novel self-attention model for ZS-TAL:** We propose `ConTrans`, combining self-attention and Conv to capture long-range temporal dependencies and fine-grained local motion, addressing both contextual reasoning and precise action boundary localization.
- **Text-enhanced multi-scale visual–textual fusion:** A hierarchical mechanism aligns visual and textual features across temporal scales, leveraging semantic guidance to improve cross-modal interactions and zero-shot action localization.
- **Rich temporal representation for open-set detection:** Attention-based cross-modal reasoning produces rich temporal features sensitive to subtle boundary variations, enabling accurate detection of unseen actions in complex videos.
- **Improved performance in ZS-TAL**: `ConTrans` achieves state-of-the-art results on benchmark datasets ActivityNet-1.3 [16] and THUMOS14 [17], demonstrating robust performance in zero-shot temporal action localization.

2. Related Work

2.1. Temporal action localization

TAL aims to detect and classify action segments in long, untrimmed videos. Existing methods fall into two categories: (i) Two-stage approaches that first generate temporal proposals and then classify them [18, 19], and (ii) Single-stage approaches that perform classification and localization in a single forward pass [20]. Recent state-of-the-art methods focus on proposal-free models for better efficiency and accuracy. Transformer-based models like Actionformer [2] generate multi-scale feature representations while simultaneously classifying actions and detecting boundaries. However, they require full supervision and large datasets, which are often impractical. Inspired by proposal-free methods, we introduce a simple yet effective solution to generalize to unseen actions with minimal training data.

2.2. Zero-shot temporal action localization

In zero-shot settings, the goal is to detect and localize unseen action instances not present in the training data. EffPrompt [21] introduced zero-shot temporal action localization (ZS-TAL) with a two-stage approach, generating action proposals and classifying them using CLIP [3]. ZEETAD [22] employs CLIP to encode RGB frames and action categories, using frame-level similarity to construct semantic representations. However, such a formulation primarily captures single-scale semantics and may struggle with temporally complex or motion-dependent actions. Our method addresses this limitation by integrating visual and textual features at multiple scales, allowing semantic alignment to adapt across varying temporal extents. However, it is computationally expensive due to redundant proposals. STALE [14] offers a single-stage approach with parallel classification and localization, using class-agnostic masking for adaptability but lacking explicit boundary regression. UnLoc [23] builds a feature pyramid to classify actions and detect boundaries at each frame. [15] improves cross-modal alignment with multi-modality prompting but relies on a complex two-step training process. GAP [24] enhances action proposals using static CLIP information but remains inefficient due to its two-stage nature. To address these issues, we propose a one-stage model that efficiently classifies actions and detects temporal boundaries.

2.3. Vision Language

Vision-language models (VLM) combine computer vision and natural language processing to address tasks like image-text retrieval [25] and visual question-answering [26]. Pre-trained on large-scale image-text pairs, VLMs can be applied to visual recognition tasks without additional fine-tuning. CLIP [3], a large-scale VLM trained with an image-text contrastive objective, has shown outstanding zero-shot performance [27]. Recently, CLIP has been adapted for video tasks such as text-based action localization [28], typically using a two-stage approach involving foreground cropping and alignment. However, this process suffers from error propagation, where mistakes in the first stage affect the second. To overcome this, we propose a single-stage model for simultaneous classification and localization without additional CLIP fine-tuning.

3. Proposed method

This section provides a comprehensive overview of our proposed `ConTrans` model, with the overall framework illustrated in Fig. 1. We begin by defining the problem in Subsection 3.1. Next, Subsection 3.2 discusses how the pre-trained CLIP model is utilized as an embedding module for ZS-TAL. Finally, Subsection 3.3 delves into the proposed local-global multi-scale feature representation module, detailing its role in action classification and localization.

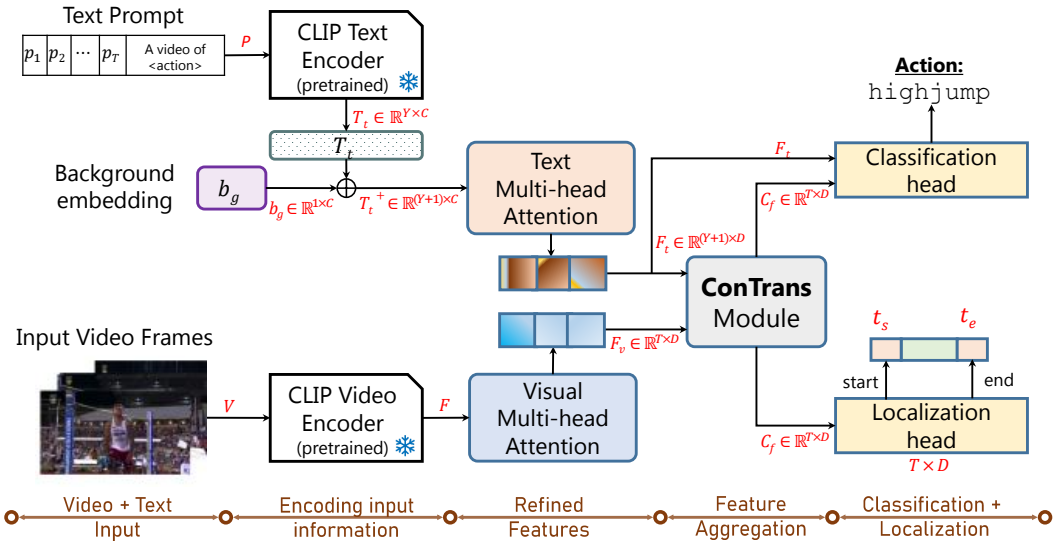

Figure 1. Overview of the proposed `ConTrans` architecture.

3.1. Problem definition

The main goal of the proposed method is to identify temporal action instances in unseen videos, defined by their start time, end time, and action label. For instance, in the dataset used in ZS-TAD, each untrimmed video V, consisting of S snippets, is associated with a set of action annotations $L = \{(t_s^i, t_e^i, y^i)\}_{i=1}^{S}$, where t_s^i and t_e^i denote the start and end times of the i-th action instance, and $y^i \in Y$ is its action label. Both closed-set and open-set scenarios are examined. In the closed-set scenario, the training and evaluation action labels are identical ($Y_{train} = Y_{val}$), whereas in the open-set scenario, the action labels for training and evaluation are mutually exclusive ($Y_{train} \cap Y_{val} = \emptyset$).

3.2. **Pre-trained CLIP as encoding module**

Recent studies [14] leverage CLIP as the backbone for ZS-TAD due to its strong zero-shot transfer capability. CLIP is pre-trained on 400 million image–text pairs and maps visual and textual inputs into a shared latent space using image and text encoders. For ZS-TAD, we use the pre-trained CLIP encoders to extract visual and textual features, keeping the model frozen during training. For visual encoding, we sample T consecutive frames from a video $V = \{l_1, l_2, \cdots, l_T\}$. The CLIP image encoder extracts visual features $F \in \mathbb{R}^{T \times D}$, where D is the feature dimension and T is the number of sampled frames. To capture global temporal context, we apply a multi-head attention (MHA) mechanism [29] with $(query, key, value) = (F, F, F)$, producing refined visual features F_v as shown in Eq. (3.1).

$$F_v = \text{MHA}(F, F, F) \in R^{T \times D}. \tag{3.1}$$

Textual Encoding– Each action category $y_i \in Y$ is represented using a template, "`a video of <action>`", where `<action>` is replaced by the class label y_i. This text prompt, P_T, is fed into the pre-trained CLIP text encoder to obtain textual embeddings. Let $T_t \in \mathbb{R}^{Y \times C}$ denote the resulting embeddings for all Y action classes, where C is the embedding dimension. Since temporal action localization also requires background classification, we introduce a learnable background embedding $b_g \in \mathbb{R}^{1 \times C}$ and append it to T_t, as in Eq. (3.2):

$$T_t^+ = [T_t; b_g] \in \mathbb{R}^{(Y+1) \times C}. \tag{3.2}$$

The refined textual features F_t are then obtained using a multi-head attention mechanism, as defined in Eq. (3.3). Positional encoding (PE) is excluded from both visual and textual branches, as our ablation study indicates that it degrades performance.

$$F_t = \text{MHA}(T_t^+, T_t^+, T_t^+) \in \mathbb{R}^{(Y+1) \times D}. \tag{3.3}$$

3.3. **ConTrans module**

Fig. 2 illustrates the high-level structure of the proposed `ConTrans` module, while Fig. 3 provides a detailed connectivity diagram of a `ConTrans` layer. The `ConTrans` module consists of multiple stacked `ConTrans` layers followed by strided depthwise 1D Conv layers. Each `ConTrans` layer is designed to capture both local and global cross-modal interactions

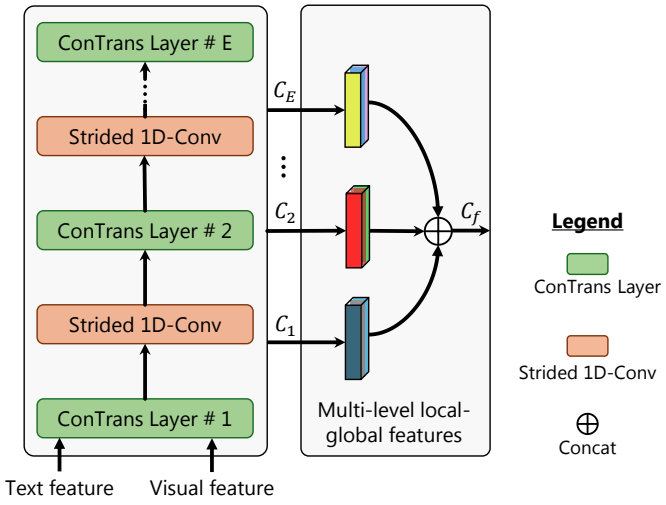

Figure 2. The proposed `ConTrans` module, which comprises "E" number of `ConTrans` layers, with each layer followed by a strided 1D-Conv for downsampling.

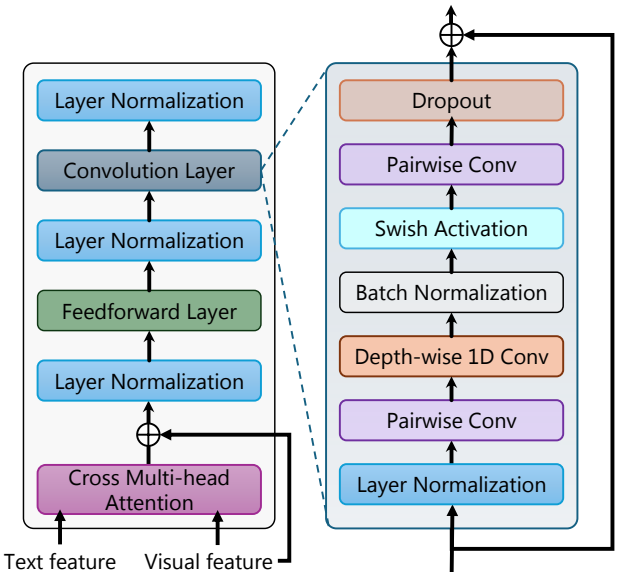

Figure 3. A detailed view of the `ConTrans` layer. It captures both local and global context via a combination of cross-attention and convolutional layers.

between textual and visual features at multiple hierarchical levels. This hierarchical design progressively refines cross-modal representations, enabling accurate detection of actions across varying temporal durations. As shown in Fig. 3, each `ConTrans` layer comprises a cross multi-head attention module, followed by a Conv layer and a feed-forward network, with layer normalization applied throughout to ensure training stability. Given visual features F_v and textual features F_t, the model uses contextual visual information to guide textual representations in identifying informative temporal regions. Specifically, text features act as queries, allowing the model to focus on video regions most relevant to each action category. This design is especially effective in zero-shot settings, where textual semantics provide critical guidance for localizing unseen actions. To this end, we employ cross-attention with $(query, key, value) = (F_t, F_v, F_v)$, as in Eq. (3.4).

$$\begin{aligned} q &= W_q\, \mathcal{LN}(F_t) \in \mathbb{R}^{(Y+1)\times d}, \\ k &= W_k\, \mathcal{LN}(F_v) \in \mathbb{R}^{T\times d}, \\ v &= W_v\, \mathcal{LN}(F_v) \in \mathbb{R}^{T\times d}, \end{aligned} \tag{3.4}$$

where $W_q, W_k, W_v \in \mathbb{R}^{D\times d}$ are learnable projection matrices. The output of the multi-head cross-attention mechanism, followed by a feed-forward layer, $FF(\cdot)$, and layer normalization, $\mathcal{LN}(\cdot)$, is defined in Eq. (3.5).

$$\begin{aligned} \mathrm{MHA}(q,k,v) = \mathrm{Concat}(head_1, head_2, \cdots, head_h); \quad head_i = \mathrm{Attention}(q,k,v) \\ C_a = \mathcal{LN}(FF(\mathrm{MHA})), \end{aligned} \tag{3.5}$$

where $C_a \in \mathbb{R}^{(Y+1)\times d}$ is the final cross-attention output. To model local cross-modal correlations, C_a is first projected into temporal dimension $T \times d$ and processed by a Conv block. This block complements the global context captured by attention by learning fine-grained local dependencies. The Conv output is combined with a residual connection and normalized to generate C_o, as expressed in Eq. (3.6).

$$C_o = \mathcal{LN}(C_a + \mathrm{Conv}(C_a)). \tag{3.6}$$

Inspired by [30], the Conv block consists of a pointwise Conv with a gated linear unit (GLU), followed by a 1D depthwise Conv and batch normalization (see Fig. 3). The resulting C_o captures both local and global cross-modal context at each `ConTrans` layer. To obtain rich multi-scale temporal representations, the output of each `ConTrans` layer is downsampled using 2× strided depthwise 1D Conv [2] before being passed to subsequent layers as:

$$\begin{aligned} C_o^{\wedge} = C_o(\downarrow); \quad C_e^{\wedge} = \mathcal{LN}(FF(\text{MHA}(C_o^{\wedge}))), \text{ and} \\ C_{e_i}^{\wedge} = \mathcal{LN}((C_e^{\wedge} + \text{Conv}(C_e^{\wedge})), i = \{1, \cdots, E\}. \end{aligned} \tag{3.7}$$

Thus, $\{C_e, C_{e_1}, C_{e_2}, \cdots C_E\}$ represents the output from E `ConTrans` layers. To enable the model to leverage both low-level and high-level representations, we then concatenate the multi-scale outputs as shown in Eq. (3.8).

$$C_f = \text{Concat}(C_e, C_{e_1}, C_{e_2}, \cdots, C_E). \tag{3.8}$$

For action classification, the class probability $Prob$ is computed by the dot product between C_f and the refined text features F_t as defined in Eq (3.9).

$$Prob = \text{Softmax}(F_t C_f^T) \tag{3.9}$$

where $Prob_{i,j}$ denotes the probability of the i-th class at the the j-th temporal snippet. For temporal localization, we follow prior work [14, 15] and predict 1D action masks over the full video duration. A stack of three 1D-Conv layers operates on C_f to produce foreground probabilities $A_L \in \mathbb{R}^{T\times 1}$, where each element represents the likelihood of an action at the corresponding temporal snippet as shown in Eq (3.10).

$$A_L = \text{Sigmoid}(\text{1D-CNN}(C_f)). \tag{3.10}$$

In addition, the model indicates the presence and strength of action at each snippet by generating a confidence score $Conf$ using a series of Conv layers, as shown in Eq (3.11). Higher confidence values facilitate more accurate boundary estimation,

$$Conf = \text{ReLU}(\text{1D-CNN}(C_f)). \tag{3.11}$$

4. **Training objective and Inference**

For label assignment, the ground-truth annotations are structured as follows. Given a training video with annotated temporal intervals and class labels, all snippets within the duration of an action instance are assigned the same action class, while snippets outside any action interval are labeled as background. For each action snippet in the class stream, we assign a binary instance mask in the action mask stream that spans the entire video length at the corresponding snippet index. All snippets belonging to the same action instance share the same instance-specific mask. For action classification, we employ the cross-entropy loss L_c between the predicted class probabilities $p \in Prob$ and the ground-truth labels $y \in \mathbb{R}^{(Y+1)\times C}$ for each action snippet, as defined in (4.1).

$$L_c = \text{CrossEntropy}(p, y). \tag{4.1}$$

For the action localization loss, binary dice loss [14] along with weighted cross-entropy loss is computed between the predicted action mask $A \in R^{(TX1)}$ of the $i-th$ action snippet and the ground-truth action mask $g \in R^{(TX1)}$ as defined in (4.2).

$$\begin{aligned} M = \beta_f \sum_{i=1}^{T} g(t) log(A(t)) + \beta_b \sum_{i=1}^{T} (1 - g(t))(1 - log(A(t))) \\ + \lambda \left(1 - \frac{A^T g}{\sum_{i=1}^{T} (A(t))^2 + (g(t))^2}\right), \end{aligned} \tag{4.2}$$

where β_f, β_b are the inverse of the foreground/background snippet's proportion and λ=0.4 is the loss trade-off coefficient. For the generated confidence score, we adopt $L2$ loss to calculate regression error and set the weight term λ=10. Thus, the overall loss is as follows:

$$Loss = L_c + M + \lambda \times L2. \tag{4.3}$$

During testing, action instance predictions for each test video are generated using the classification predictions $Prob$ and the mask predictions A. For $Prob$, we only consider snippets where the class probabilities surpass a threshold θ_c and select the highest-scoring snippets. For each of these high-scoring action snippets, the corresponding temporal mask is obtained by applying a threshold to the t_i-th column of A using the localization threshold Θ. To ensure a sufficient number of candidates, we use a set of thresholds $\Theta = \{\theta_i\}$. For these candidate snippets, the confidence score s is calculated by multiplying the classification score by the maximum mask score. Finally, SoftNMS [31] is applied to obtain the top-ranked results. The proposed method is evaluated using two standard ZS-TAL datasets: ActivityNet-1.3 [16], and THUMOS14 [17]. ActivityNet-1.3 consists of 200 action categories and a total of 19,994 videos. In accordance with the standard protocol, the dataset is split into training, validation, and test sets in a 2:1:1 ratio. THUMOS14 contains 20 action categories with 200 validation videos and 213 test videos. Both datasets include annotations for temporal boundaries and corresponding action labels.

4.1. **Implementation details**

For fair comparison with prior works [14, 15, 23], we adopt a pre-trained CLIP (ViT-B/16) visual and text encoders with feature dimension D=512, keeping them frozen during training. The encoder, i.e., ViT-B/16 + Transformer, is used to generate textual embeddings. Video frames are resized to 224×224, and 77 textual tokens are used for both datasets [14]. Extracted features are temporally rescaled to 100 and 256 for ActivityNet-1.3 and THUMOS14, respectively. The model employs 4 `ConTrans` blocks with 8 attention heads and is trained for 9 epochs using the Adam optimizer, with learning rates of 10^{-4} for ActivityNet-1.3 and 10^{-5} for THUMOS14. Performance is evaluated using the standard metric mean average precision (mAP) at multiple IoU thresholds.

5. **Comparison results with the state-of-the-art models**

5.1. **Experimental setup and analysis**

a. Zero-shot settings: The performance of the proposed model is assessed in open-set scenarios where $Y_{train} \cap Y_{val} = \emptyset$, meaning the training and evaluation labels do not overlap. We adhere to the evaluation settings and dataset splits outlined in [14]. Specifically, the two datasets, ActivityNet-1.3 and THUMOS14, are evaluated under two conditions: 1) training on 75% of the action categories and testing on the remaining 25%, and 2) training on 50% of the action categories and testing on the other 50%. For consistent and reliable evaluation, we average results across 10 random splits and compare only with methods that adhere to the same evaluation procedure.

Competitors– As ZS-TAD is a relatively new problem, only a few competitive methods [14, 15, 23], and GAP [24] are available for fair comparison. We additionally introduce two CLIP-based baselines: B-I, a two-stage approach combining BMN [32] with CLIP, and B-II, a one-stage TAD model integrated with CLIP. Both baselines use the same CLIP-initialized text encoder. Direct comparison with ZS-TAD [33] is not feasible due to unavailable code and inconsistent data splits with [21]. To ensure statistical reliability, we perform 10 random category samplings for each configuration, following [14].

Performance– Referring to Table 1, it is found that our approach achieves the highest

Data split Train:Test	Model	ActivityNet-1.3				THUMOS14					
		0.5	0.75	0.95	mean	0.3	0.4	0.5	0.6	0.7	mean
75:25	B-II	32.6	18.5	5.8	19.6	28.5	20.3	17.1	10.5	6.9	16.6
	B-I	35.6	20.4	2.1	20.2	33.0	25.5	18.3	11.6	5.7	18.8
	EffPrompt	37.6	22.9	3.8	23.1	39.7	31.6	23.0	14.9	7.5	23.3
	mProTEA	44.5	27.4	7.9	27.6	43.1	38.2	28.2	18.1	8.7	27.9
	STALE	38.2	25.2	6.0	24.9	40.5	32.3	23.5	15.3	7.6	23.8
	GAP	47.6	32.5	8.6	31.8	52.3	44.2	32.8	22.4	12.6	32.9
	GRIZAL	46.4	32.5	6.8	30.1	43.2	-	25.7	-	9.8	27.0
	Ours	**51.9**	**33.1**	**12.2**	**35.2**	**51.3**	**45.6**	**33.1**	**22.7**	**12.9**	**33.3**
50:50	B-II	32.1	20.7	3.7	12.9	21.0	16.4	11.2	6.3	3.2	11.6
	B-I	28.0	16.4	1.2	16.0	27.2	21.3	15.3	9.7	4.8	15.7
	EffPrompt	32.0	19.3	2.9	19.6	37.2	29.6	21.6	14.0	7.2	21.9
	mProTEA	41.8	24.6	6.1	25.6	41.2	36.3	26.3	16.8	8.4	26.1
	STALE	32.1	20.7	5.9	20.5	38.3	30.7	21.2	13.8	7.0	22.2
	GAP	41.6	26.2	6.1	26.4	44.2	36.0	27.1	15.1	8.0	26.1
	GRIZAL	39.9	25.7	6.6	25.7	40.0	-	25.0	-	9.1	25.2
	UnLoc	43.7	-	-	-	-	-	-	-	-	-
	Ours	**45.3**	**30.3**	**10.5**	**32.5**	**44.5**	**36.3**	**26.8**	**15.7**	**8.8**	**27.2**

Table 1. Comparison of the proposed model with other state-of-the-art models on the ActivityNet-v1.3 [16], and THUMOS14 [17] datasets for zero-shot Temporal Action Localization (ZS-TAL), evaluating the mean average precision (mAP) at different temporal Intersection over Union (tIoU) thresholds for open-set settings.

Model	Mode	ActivityNet-1.3				THUMOS14					
		0.5	0.75	0.95	mean	0.3	0.4	0.5	0.6	0.7	mean
A2Net+I3D	RGB	39.6	25.7	2.8	24.8	45.0	40.5	31.3	19.9	10.0	29.3
B-I+CLIP	RGB	28.2	18.3	3.7	18.2	36.3	31.9	25.4	17.8	10.4	24.3
B-II+CLIP	RGB	51.5	33.3	6.6	32.7	57.1	49.1	40.4	31.2	23.1	40.2
EffPrompt+CLIP	RGB	44.0	27.0	5.1	27.3	50.8	44.1	35.8	25.7	15.7	34.5
STALE+CLIP	RGB	54.3	34.0	7.7	34.3	60.6	53.2	44.6	36.8	26.7	44.4
`ConTrans`+CLIP	RGB	**54.5**	**41.8**	**15.5**	**36.6**	**62.1**	**53.5**	**45.1**	**36.9**	**27.2**	**45.1**
TALNet+I3D	RGB+Flow	38.2	18.3	1.3	20.2	53.3	48.5	42.8	33.8	20.8	39.8
MUSES+I3D	RGB+Flow	50.0	34.9	6.5	34.0	68.9	64.0	57.1	46.7	31.2	52.9
B-III+I3D	RGB+Flow	47.2	30.7	8.6	30.8	68.3	62.3	51.9	38.8	23.7	-
STALE+I3D	RGB+Flow	56.5	36.7	9.5	36.4	68.9	64.1	57.1	46.7	31.2	52.9
`ConTrans`+I3D	RGB+Flow	**60.2**	**42.4**	**22.3**	**38.5**	**69.1**	**64.6**	**57.4**	**47.1**	**31.5**	**53.2**

Table 2. Comparison of the proposed model with other state-of-the-art models using I3D or CLIP encoder backbones on the ActivityNet1.3 and THUMOS14 for zero-shot Temporal Action Localization (ZS-TAL), evaluating the mean average precision (mAP) at different temporal Intersection over Union (tIoU) thresholds on closed-set settings.

mAP, reaching 35.2 and 33.3 under the 75%–25% split, and continues to outperform state-of-the-art methods under the more challenging 50%–50% split with 32.5 and 27.2 mAP, respectively. The consistent gains across splits demonstrate strong robustness to limited training data and effective generalization to unseen classes. Performance improvements on ActivityNet-1.3 stem from modeling long temporal segments, while gains on THUMOS14 arise from capturing local–global context for densely occurring short actions.

b. Closed-set settings: In closed-set scenarios, the action labels used for training and evaluation are identical (i.e., $Y_{train} = Y_{val}$). To ensure a fair comparison, we adopt the same dataset splits as reported in previous studies.

Competitors– The proposed model's performance is evaluated against the latest state-of-the-art approaches, such as STALE [14]. Furthermore, we consider several temporal action localization(TAL) methods that leverage the CLIP backbone, as well as three baseline models for comparison: the two-stage CLIP-based baseline B-I, and two one-stage

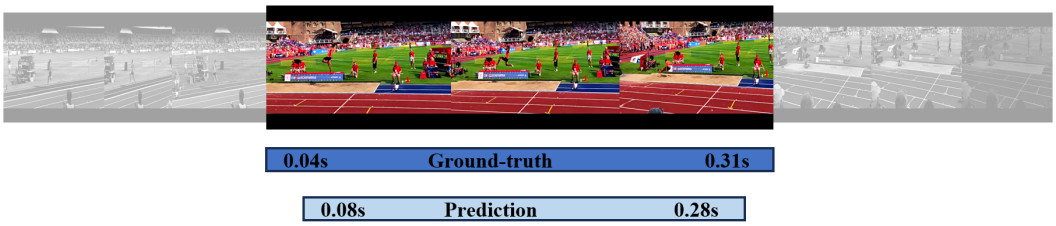

Figure 4. Qualitative analysis for action class "Long jump".

Model	Train time (hrs)	Infer time (s)
EffPrompt [21]	3.24	135
STALE [14]	3.96	151.8
mProTEA [15]	1.42	162.5
Ours	**1.3**	**31**

Table 3. Complexity analysis of `ConTrans`.

baselines—CLIP-based B-II and B-III, which utilize Kinetics pre-trained I3D.

Performance– As shown in Table 2, the proposed model improves consistently with more training samples on both ActivityNet-1.3 and THUMOS14. `ConTrans` outperforms existing methods across different input modalities (RGB and RGB+Flow) and encoder backbones (I3D and CLIP), highlighting its effectiveness across diverse feature representations. Its multi-level feature design captures both fine-grained and high-level temporal context, enabling robust detection of actions with varying durations and strong generalization across datasets and action categories.

Computational complexity– As demonstrated in Table 3, our model exhibits faster performance in both training and inference times compared to prior SOTA methods. Its efficiency is further highlighted by its early convergence, achieving optimal results in just 9 epochs.

Qualitative analysis– The qualitative analysis of the proposed model is presented in Fig. 4. For the action class *Long jump*, the model accurately predicts the temporal boundaries in close alignment with the ground truth. As shown in the figure, the video spans a total duration of 31 seconds [0.04-0.31]s, and the proposed model predicts comparable boundaries at [0.08-0.28]s.

6. Ablation study

To assess the efficacy of the proposed approach, we perform extensive ablation studies on the ActivityNet-1.3 dataset using the 75%-25% split configuration.

a. Impact of local-global context dependencies: It can be observed from Table 4 that removing either the local or global context information results in a performance drop of 2.3%, highlighting the importance of incorporating both local and global temporal dependencies. Interestingly, the model's performance with either local attention or global attention alone is similar, with only a slight increase of 0.2% when local attention is used, demonstrating that both forms of attention contribute equally in capturing essential temporal relationships for ZS-TAL.

b. Impact of `ConTrans` layers: We also evaluate the model's performance based on the number of `ConTrans` layers used to generate multi-scale features. It can be seen from Table 5 that the model performs optimally with four `ConTrans` blocks. However, increasing the number of `ConTrans` blocks results in a rise in model parameters, leading to higher computational costs and the potential for overfitting.

Model	0.5	0.75	0.95	mean
w/o local	49.8	29.3	9.8	32.7
w/o global	50.1	30.3	9.8	32.9
full	**51.9**	**33.1**	**12.2**	**35.2**

Table 4. Impact of local-global dependencies, wrt. $\text{mAP}_{0.5-0.95}$ (%).

Number of layers	0.5	0.75	0.95	mean
1	50.2	30.4	10.6	33.4
3	51.3	32.1	10.0	34.1
4	**51.9**	**33.1**	**12.2**	**35.2**

Table 5. The impact of the number of `ConTrans` layers, wrt. $\text{mAP}_{0.5-0.95}$ (%).

Model	0.5	0.75	0.95	mean
with PE	50.9	31.0	9.5	33.3
w/o PE	**51.9**	**33.1**	**12.2**	**35.2**

Table 6. The impact of PEs on both the visual and text embeddings, wrt. $\text{mAP}_{0.5-0.95}$ (%).

Text encoder	Video encoder	mAP (%) 0.5	mean
✗	✓	47.1	32.6
✓	✗	49.8	33.1
✓	✓	**51.9**	**35.2**

Table 7. Effect of encoders in the proposed model.

mAP(%)	Max pooling	Avg pooling	Strided 1D-Conv
0.5	44.4	45.4	**51.9**
mean	32.6	33.1	**35.2**

Table 8. Effect of different down-samplers.

c. Effect of positional encodings on visual and text encoders: We examine the effect of positional encodings on visual and textual encoders. As shown in Table 6, adding them reduces performance, likely because dense frame sampling and cross-modal alignment in zero-shot TAD make explicit positional biases detrimental. Consequently, positional encodings are excluded in our model.

d. Effect of frame and text encoders: In this work, text is encoded using the pre-trained CLIP text encoder, and video frames are processed with Transformer multi-head attention. Table 7 shows performance drops when either encoder is removed, highlighting their crucial role in the model's effectiveness.

e. Effects of different downsampling methods: The ablation study at Table 8 on different downsampling methods for the feature pyramid shows that the strided 1D-CNN achieves the best performance, while max pooling performs the worst, due to the loss of contextual information from aggressively selecting only maximum values.

7. Conclusion

We propose `ConTrans`, a framework combining self-attention and convolution to capture both fine-grained local patterns and long-range temporal dependencies. Its multi-scale visual–textual fusion aligns semantic and visual features across temporal resolutions, enabling effective zero-shot action localization. Experiments on ActivityNet-1.3 and THUMOS14 show `ConTrans` consistently outperforms state-of-the-art methods, demonstrating strong generalization and effectiveness. Challenges remain for ambiguous visuals or overlapping actions with similar semantics, where precise boundaries are difficult. Future work could leverage finer motion cues or stronger temporal constraints to address these limitations.

References

[1] J. Shao, X. Wang, R. Quan, J. Zheng, J. Yang, and Y. Yang. "Action sensitivity learning for temporal action localization". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 13457–13469.

[2] C.-L. Zhang, J. Wu, and Y. Li. "Actionformer: Localizing moments of actions with transformers". In: *European Conference on Computer Vision*. Springer. 2022, pp. 492–510.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

[4] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. "Scaling up visual and vision-language representation learning with noisy text supervision". In: *International conference on machine learning*. PMLR. 2021, pp. 4904–4916.

[5] J. Yang, C. Li, P. Zhang, B. Xiao, C. Liu, L. Yuan, and J. Gao. "Unified contrastive learning in image-text-label space". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 19163–19173.

[6] B. Kim, C. Shin, J. Jeong, H. Jung, S.-Y. Lee, S. Chun, D.-H. Hwang, and J. Yu. "Zim: Zero-shot image matting for anything". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2025, pp. 23828–23838.

[7] J. Luo, Z. Zhao, and Y. Liu. "Zero shot domain adaptive semantic segmentation by synthetic data generation and progressive adaptation". In: *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2025, pp. 11531–11538.

[8] J. Li, Q. Xie, R. Gu, J. Xu, Y. Liu, and X. Yu. "LGD: Leveraging Generative Descriptions for Zero-Shot Referring Image Segmentation". In: *arXiv preprint arXiv:2504.14467* (2025).

[9] M. Bosetti, S. Zhang, B. Liberatori, G. Zara, E. Ricci, and P. Rota. "Text-Enhanced Zero-Shot Action Recognition: A Training-Free Approach". In: *International Conference on Pattern Recognition*. Springer. 2024, pp. 327–342.

[10] G. Ye, L. Li, K. Li, J. Xiao, and L. Chen. "Zero-shot compositional action recognition with neural logic constraints". In: *Proceedings of the 33rd ACM International Conference on Multimedia*. 2025, pp. 3625–3634.

[11] Y. Tewel, Y. Shalev, R. Nadler, I. Schwartz, and L. Wolf. "Zero-shot video captioning with evolving pseudo-tokens". In: *arXiv preprint arXiv:2207.11100* (2022).

[12] P. Li, T. Wang, and Z. Pan. "Temporal prompt guided visual-text-object alignment for zero-shot video captioning". In: *Computer Vision and Image Understanding* (2025), p. 104601.

[13] K. Tran, A. D. Le Dinh, T.-P. Nguyen, T. Phan, P. Nguyen, K. Luu, D. Adjeroh, G. Doretto, and N. Le. "Z-gmot: Zero-shot generic multiple object tracking". In: *Findings of the Association for Computational Linguistics: NAACL 2024*. 2024, pp. 3468–3479.

[14] S. Nag, X. Zhu, Y.-Z. Song, and T. Xiang. "Zero-shot temporal action detection via vision-language prompting". In: *European Conference on Computer Vision*. Springer. 2022, pp. 681–697.

[15] A. Raza, B. Yang, and Y. Zou. "Zero-Shot Temporal Action Detection by Learning Multimodal Prompts and Text-Enhanced Actionness". In: *IEEE Transactions on Circuits and Systems for Video Technology* (2024).

[16] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. "Activitynet: A large-scale video benchmark for human activity understanding". In: *Proceedings of the ieee conference on computer vision and pattern recognition*. 2015, pp. 961–970.

[17] H. Idrees, A. R. Zamir, Y.-G. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah. "The thumos challenge on action recognition for videos "in the wild"". In: *Computer Vision and Image Understanding* 155 (2017), pp. 1–23.

[18] C. Zhao, A. K. Thabet, and B. Ghanem. "Video self-stitching graph network for temporal action localization". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 13658–13667.

[19] H. Su, W. Gan, W. Wu, Y. Qiao, and J. Yan. "Bsn++: Complementary boundary regressor with scale-balanced relation modeling for temporal action proposal generation". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. 3. 2021, pp. 2602–2610.

[20] C. Lin, C. Xu, D. Luo, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, and Y. Fu. "Learning salient boundary feature for anchor-free temporal action localization". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 3320–3329.

[21] C. Ju, T. Han, K. Zheng, Y. Zhang, and W. Xie. "Prompting visual-language models for efficient video understanding". In: *European Conference on Computer Vision*. Springer. 2022, pp. 105–124.

[22] T. Phan, K. Vo, D. Le, G. Doretto, D. Adjeroh, and N. Le. "Zeetad: Adapting pretrained vision-language model for zero-shot end-to-end temporal action detection". In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2024, pp. 7046–7055.

[23] S. Yan, X. Xiong, A. Nagrani, A. Arnab, Z. Wang, W. Ge, D. Ross, and C. Schmid. "Unloc: A unified framework for video localization tasks". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 13623–13633.

[24] J.-R. Du, K.-Y. Lin, J. Meng, and W.-S. Zheng. "Towards Completeness: A Generalizable Action Proposal Generator for Zero-Shot Temporal Action Localization". In: *International Conference on Pattern Recognition*. Springer. 2024, pp. 252–267.

[25] Z. Wang, X. Liu, H. Li, L. Sheng, J. Yan, X. Wang, and J. Shao. "Camp: Cross-modal adaptive message passing for text-image retrieval". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 5764–5773.

[26] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. "Vqa: Visual question answering". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 2425–2433.

[27] B. Yang, F. Liu, X. Wu, Y. Wang, X. Sun, and Y. Zou. "Multicapclip: Auto-encoding prompts for zero-shot multilingual visual captioning". In: *arXiv preprint arXiv:2308.13218* (2023).

[28] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman. "End-to-end learning of visual representations from uncurated instructional videos". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 9879–9889.

[29] A Vaswani. "Attention is all you need". In: *Advances in Neural Information Processing Systems* (2017).

[30] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, et al. "Conformer: Convolution-augmented transformer for speech recognition". In: *arXiv preprint arXiv:2005.08100* (2020).

[31] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis. "Soft-NMS–improving object detection with one line of code". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 5561–5569.

[32] T. Lin, X. Liu, X. Li, E. Ding, and S. Wen. "Bmn: Boundary-matching network for temporal action proposal generation". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 3889–3898.

[33] L. Zhang, X. Chang, J. Liu, M. Luo, S. Wang, Z. Ge, and A. Hauptmann. "Zstad: Zero-shot temporal activity detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 879–888.