Zero-Shot Temporal Action Detection by Learning Multimodal Prompts and Text-Enhanced Actionness

Asif Raza, Bang Yang, and Yuexian Zou, *Senior Member, IEEE*

***Abstract*— Zero-shot temporal action detection (ZS-TAD), aiming to recognize and detect new and unseen video actions, is an emerging and challenging task with limited solutions. Recent studies have adapted the vision-language pre-trained model CLIP for this task in a parameter-efficient fine-tuning fashion to achieve open-vocabulary detection. However, they suffer from insufficient vision-text alignment because of the dual-stream structure of CLIP and yield inferior TAD results due to the lack of accurate action prior. In this paper, we target the above limitations and propose to learn multimodal Prompts and Text-Enhanced Actionness (mProTEA) for ZS-TAD. Specifically, we insert learnable layer-wise prompts into the vision and text branches of the frozen CLIP and establish a strong coupling between them, resulting in multimodal prompts that can boost cross-modal alignment. To ease computation costs, we propose to conduct multimodal prompt learning on an image recognition dataset with rich concepts (e.g., ImageNet) first and then keep them frozen during TAD fine-tuning. For improving TAD, we introduce text-enhanced actionness modeling, where we leverage the concise semantics of text to assist the calculation of class-agnostic actionness scores, to offer accurate prior information for both action classification and localization. With the above designs, our mProTEA excels in extensive TAD experiments, surpassing the strong competitor STALE by 5.1% on ActivityNet under the zero-shot setting and achieving state-of-the-art performance in conventional supervised scenarios. Ablation studies confirm the effectiveness of our proposals and show superior domain generalization of multimodal prompts learned on ImageNet against the other 10 image recognition datasets.**

***Index Terms*— Zero-shot, temporal action detection, multimodal prompt learning, actionness modeling.**

I. Introduction

TEMPORAL action detection (TAD) is one of the fundamental tasks in video understanding that aims to predict the semantic label and the time range of every action instance in an untrimmed video. Traditional TAD methods [1], [2], [3], [4], [5], [6] follow a supervised learning paradigm and assume that the action categories within the training and testing sets remain identical. Nonetheless, this assumption confines the applicability of TAD to new and diverse scenarios, often necessitating model re-training to accommodate novel actions. To deploy TAD systems in the real world, the concept of *open-vocabulary* detection becomes imperative. In response, zero-shot TAD (ZS-TAD) is introduced [7], [8], [9], and it poses a distinctive challenge – no overlap exists between action categories in the training and testing sets. Consequently, addressing ZS-TAD requires an effective backbone network encompassing a broad spectrum of world knowledge beyond specific datasets, which can be obtained via, e.g., pre-training.

Recently, vision-language pre-trained models (PTMs) like CLIP [10] have demonstrated impressive open-vocabulary classification and zero-shot transfer abilities in a wide range of vision-language tasks [11], [12], [13], [14] due to the flexibility of representing curated categories in the text format and the semantic alignment between the vision and text modalities. Although these abilities of CLIP-like PTMs make them suitable for ZS-TAD, adapting PTMs to a downstream task risks the loss of knowledge with improper fine-tuning techniques. Prior ZS-TAD methods, e.g., the state-of-the-art STALE [8], tackle this challenge with the emerging parameter-efficient fine-tuning technique [15]. As illustrated in Figure 1(a), STALE freezes the PTM (i.e., CLIP) to preserve its internal knowledge, jointly trains text prompts and TAD modules, and specifically improves action classification by learning mask representations from videos. Despite its effectiveness, STALE suffers from two major demerits: 1) insufficient vision-text interactions because of the dual-stream structure of CLIP and 2) inaccurate TAD results due to the lack of accurate action prior.

In this paper, we target the above limitations and aim to develop a novel ZS-TAD network with multimodal Prompts and Text-Enhanced Actionness (mProTEA) by drawing inspiration from recent advances in prompting PTMs [9], [15], [16], [17], [18] and video understanding [19], [20], [21]. Figure 1(b) illustrates the diagram of mProTEA. In particular, we insert learnable layer-wise prompts into the vision and text branches of the frozen CLIP and establish a strong coupling between them, resulting in multimodal prompts that can boost cross-modal alignment. To ease computation costs,

Manuscript received 8 November 2023; revised 11 April 2024 and 9 May 2024; accepted 10 June 2024. Date of publication 13 June 2024; date of current version 27 November 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62176008 and in part by Shenzhen Science and Technology Research Program under Grant GXWD20201231165807007-20200814115301001. This article was recommended by Associate Editor D. Zeng. *(Asif Raza and Bang Yang contributed equally to this work.) (Corresponding author: Yuexian Zou.)*

Asif Raza and Yuexian Zou are with the School of Electronic and Computer Engineering, Peking University, Shenzhen 518055, China (e-mail: asifraza151@pku.edu.cn; zouyx@pku.edu.cn).

Bang Yang is with the School of Electronic and Computer Engineering, Peking University, Shenzhen 518055, China, and also with the Peng Cheng Laboratory, Shenzhen 518052, China (e-mail: yangbang@pku.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2024.3414275.

Digital Object Identifier 10.1109/TCSVT.2024.3414275

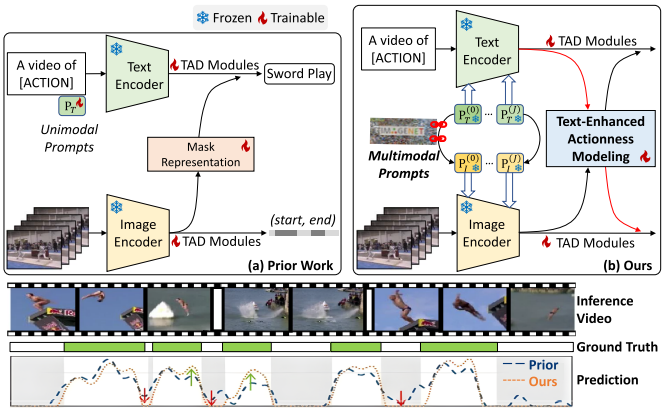


Fig. 1. Comparison with the prior work STALE [8] for ZS-TAD. (a): Prior work fails to establish sufficient vision-language interactions and specifically improves action classification by learning mask representations from videos. (b): Our work bridges text and vision branches by pre-training multimodal prompts on a recognition dataset with rich concepts (e.g., ImageNet) and introduces text-enhanced actionness modeling to boost both action classification and localization. Bottom: our approach yields superior TAD results against STALE.

we propose to pre-train multimodal prompts on an image recognition dataset with rich concepts (e.g., ImageNet [22]) first and then keep them frozen during TAD fine-tuning. For improving TAD, we introduce text-enhanced actionness modeling, where we leverage the concise semantics of text to assist the calculation of class-agnostic actionness scores, to offer accurate prior information for both action classification and localization. By incorporating these core designs, our mProTEA excels in both ZS-TAD and supervised TAD tasks on two widely adopted TAD benchmarks, i.e., ActivityNet [23] and THUMOS14 [24]. One example is given at the bottom of Figure 1, where compared with the state-of-the-art STALE, our mProTEA suppresses the over-prediction of background clips and boosts the responses of action clips. With a comprehensive ablation study, we gain insights into the inner workings of mProTEA and find that learning multimodal prompts on image recognition datasets with rich visual concepts demonstrates a superior domain generalization ability.

To summarize, Our main contributions are as follows.

1) **Multimodal Prompt Learning:** Our mProTEA involves a novel multimodal prompting strategy to bridge the gap between visual and language encoders and promote their synergy. Meanwhile, the proposed multimodal prompt learning can be agnostic to downstream TAD tasks.
2) **Text-Enhanced Actionness Modeling:** Our mProTEA diverges from conventional methods by assembling with text-enhanced actionness modeling, which can accommodate both class-independent and inter-class variations to produce action priors and thus facilitate precise TAD.
3) **Effective Solution for Open-Vocabulary TAD:** mProTEA's efficacy in addressing open vocabulary challenges in TAD is validated through extensive experiments on ActivityNet and THUMOS14 datasets. It surpasses state-of-the-art ZS-TAD methods and attains good generalization abilities from diverse image recognition datasets.

II. Related Work

A. Vision-Language (VL)

In recent years, significant research has focused on Vision-language (VL) models, merging computer vision and natural language processing strengths for tasks like image-text retrieval, visual captioning, and visual question-answering [25], [26], [27], [28]. Studies on image-text bonds include paired documents [29] and joint image-text embedding with category annotations [30], [31]. CLIP, a large-scale pretrained VL model developed by Radford et al. [10], has been a game-changer with 400 million visual-text pairs. CLIP and variants like ALIGN [32] and [33] learn potent visual representations from paired data. VL models show impressive adaptation in zero-shot classification [8], [9]. Yet, VL adaptation to video tasks, e.g., text-based action localization [34] and action recognition [35], often rely on error-prone classification-based methods involving cropping and propagating errors, semantic gaps, and generalization issues. To address this, we propose a unified classification and localization approach, optimizing pre-trained CLIP for video understanding techniques to preserve information and boost model performance.

B. Prompt Learning

Prompting guides pre-trained language models by using a few examples to demonstrate desired outputs. GPT-3 [36] showcases robust generalization through handcrafted prompt templates in few-shot or zero-shot learning. However, crafting these templates requires expertise and limits flexibility [15], [37]. While complete fine-tuning might harm previously acquired knowledge, the joint V-L representation and linear probing helps retain CLIP's zero-shot capability [38]. Recent prompt engineering trends can be categorized as discrete [39], [40], [41], [42] and continuous prompts [17], [38], [43]. Building on prompt learning in NLP [16], [17], [43], [44], researchers suggest enhancing V-L models through end-to-end training of prompt tokens. CoOp [17] refines CLIP for few-shot transfer by optimizing a continuous set of prompt vectors in the language branch. Addressing CoOp's [17] generalization on new classes, Co-CoOp [43] explicitly conditions prompts on image instances. Gao et al. propose optimizing multiple prompt sets by learning their distribution [39]. Nag et al. tailors CLIP for video understanding [8] using prompt learning, and Huang et al. perform visual prompt tuning on CLIP through the vision branch of CLIP [44]. Recent research has explored prompt learning for transferable representation and text-based action localization in videos [8], [9], [39], [45]. Existing methods typically focus on unimodal solutions and learn prompts in either CLIP's language or vision branch. Motivated by the success of continuous prompt learning [17], [38], [43], and noting limited success in video domains [8], [9], we investigate whether complete prompting (in both language and vision branches) is better for CLIP adaptation, given its multimodal nature. We address this question and explore the effectiveness of multimodal prompt learning for improving alignment between VL representations in Zero-Shot Temporal Action Detection (ZS-TAD). Notably, we propose a pretrained

CLIP tuning scheme on 11 different image recognition datasets to enhance generalization and bridge the gap between efficient knowledge transfer and domain adaptation. Our approach also facilitates efficient adaptation from image to video tasks in ZS-TAD.

C. Temporal Action Detection

Recent advancements in Temporal Action Detection (TAD) [2], [6], [46], [47], [48] showcase notable progress. For instance, R-C3D [49] employs a region-convolutional 3D network for activity localization in video streams, utilizing anchor boxes from static images [50] to create and classify proposals. TSM [51] proposes a mapping operation to represent the video's temporal-spatial features as a 2D VideoMap for effective modeling. CSL [52] proposed the semantic boundary detection method by formulating it as a reinforcement learning problem. while CBMN [46] enhances proposal generation with confidence maps to form a capsule boundary network based on U-BlockConvCaps for dense temporal action proposal generation.

Gao et al. introduced RapNet for accurate temporal action proposal generation, capturing global contextual information for actions with varying durations [53]. SMEN [54] employs a slow-motion enhanced network comprised of a mining module to mask slow action and a localization module for detection. Zhang et al. proposed TD-3DCNN, utilizing temporal dropout during training to identify frame inconsistencies [55]. Chen et al. advanced skeleton-based action recognition with FDGCN, integrating attention mechanisms and transformer encoder layers for effective spatial-temporal context [56]. Hu et al. addressed open-set temporal action localization with GOTAL, utilizing a Transformer network and sharpness minimization algorithm for generalized action representations [57]. ASK [47] improves the proposal's ask-adaptive attention for image Captioning. Gait [48] learns action representations of subjects from multi-scale features using a cross-view spatiotemporal aggregation network. In contrast to the above sequential localization and classification pipeline methods, TAGS [6] introduces a parallel strategy of localization and classification. The above techniques are supervised and rely on extensive training datasets, making it problematic in low-data settings on TAD. Therefore, we extend the parallel strategy by integrating a dedicated modality refiner regulated by actionness score modeling for TAD.

D. ZS-TAD

Zero-shot learning (ZSL) aims to recognize new, unseen classes by transferring shared knowledge from known to unknown categories [58]. In contrast, traditional approaches often rely on learning from multiple noisy annotators to improve classifier performance, a strategy known for its robustness across various datasets [59]. However, ZSL presents a novel perspective, highlighting the significance of leveraging prior knowledge. One common approach involves creating a shared representation space for both seen and unseen categories, such as attribute space [58], [60], [61] or semantic space [58], [62], [63]. Alternatively, some methods synthesize features for unseen actions [64] or use objects to establish a shared space for unseen actions [6]. In attribute space, researchers rely on prior information like visual attributes (e.g., color, shape) [65]. Parikh et al. [66] focused on learning relative attributes, a promising but less scalable approach due to manual attribute definition. In semantic space, researchers use semantic embeddings of seen and unseen concepts as prior information. These embeddings are typically learned unsupervised using Word2Vec [67] or GloVe [68]. Zhang et al. [7] pioneered the application of Word2Vec to ZSL on TAD, marking a significant milestone. Likewise, recent work [8], [9] employed image-text pre-training from CLIP to address zero-shot action recognition, particularly in ZS-TAD. We propose an interactive stepwise training scheme that adapts CLIP on diverse image datasets, leveraging shared representation spaces and semantic embeddings for efficient generalization, conserving computational resources. Our study improves ZS-TAD performance, addressing its limitations and offering the potential solution for Open Vocabulary in TAD tasks.

III. Method

In this section, we introduce our mProTEA model for ZS-TAD. The overall framework is shown in Figure 2. As we can see, the key of mProTEA lies in three factors: 1) a vision-language pre-trained model that supports open-vocabulary classification (i.e., CLIP [10]), 2) multimodal prompts that bridge vision and text branches to enhance cross-modal alignment, and 3) text-enhanced actionness modeling that provides accurate action prior for TAD. In the following, we will first review how to use CLIP as the backbone to solve ZS-TAD in Section III-A, followed by the introduction of our proposed multimodal prompt learning in Section III-B. Next, we will elaborate on our mProTEA and detail text-enhanced actionness modeling in Section III-C. Finally, we will present the training procedure of our approach in Section III-D.

A. Preliminaries of TAD

1) Problem Definition: Given a dataset D with training set D_{train} and validation set D_{val}, each untrimmed training video V in D_{train} is associated with temporal segmentation $\Psi = (\psi_j, \xi_j, y_i)_{j=1}^{M}$ with M tuples, where ψ_j and ξ_j represent the start and end time of the j-th action and $y_j \in Y$ denote the action category. TAD considers both closed-set and open-set scenarios. The action categories for training and evaluation are identical in the closed-set setting (i.e., $Y_{D_{\text{train}}} = Y_{D_{\text{val}}}$) but are disjoint in the open-set case (i.e., $Y_{D_{\text{train}}} \cap Y_{D_{\text{val}}} = \emptyset$). In particular, TAD under the open-set scenario, a.k.a. ZS-TAD, is a more practical and challenging setup. The goal of TAD is to localize the action snippets of the video accurately and classify them into correct action categories.

2) CLIP as the Backbone: Recent studies [8], [9] have treated CLIP [10] as the backbone to facilitate ZS-TAD due to its impressive zero-shot transfer ability. In particular, CLIP is pre-trained on a dataset of 400 million image-text pairs, allowing images and texts to be mapped into a common latent space. For TAD, K action categories Y (i.e., $K = |Y|$) are represented in text format for recognition, akin to replacing the

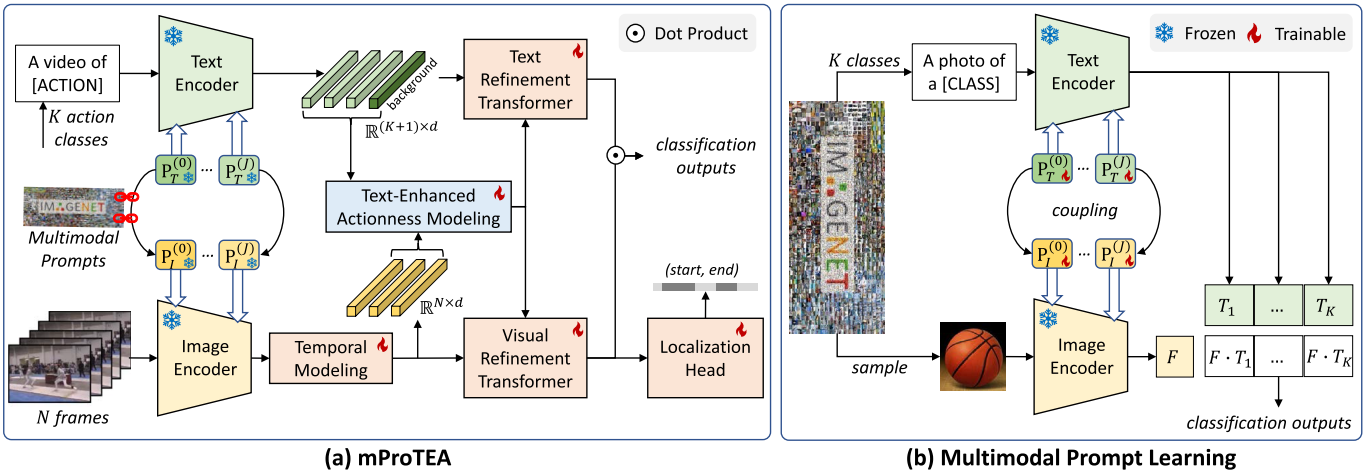


Fig. 2. Overview of our approach for zero-shot temporal action detection. **(a)**: mProTEA builds upon a frozen vision-language pre-trained model (i.e., CLIP) and consists of two key components: (1) multimodal prompts that bridge text and vision encoding and (2) text-enhanced actionness modeling that offers accurate action prior for action classification and localization. **(b)**: To ease computational costs, multimodal prompts can be pre-trained on a recognition dataset of rich visual concepts (e.g., ImageNet).

[ACTION] in the template "a video of [ACTION]" with each $y \in Y$, resulting in text representations $\boldsymbol{T} = \{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_K\} \in \mathbb{R}^{K \times d}$:

$$\boldsymbol{t}_k = \text{TextEncoder}(E_T(y_k)), \quad k \in [1, K], \tag{1}$$

where $E_T(y_k)$ produces the token embeddings associated with y_k and d denotes the feature dimension. Besides, recognizing background snippets is necessary for TAD. Therefore, we add an additional "background" class, randomly initialize its embedding, and prepend the embedding to $\boldsymbol{T}$ to obtain $\boldsymbol{T}_+ \in \mathbb{R}^{(K+1)\times d}$. For visual encoding, we uniformly sample N frames from the video V, i.e., $V = \{v_1, \ldots, v_N\}$, and use the image encoder of CLIP to extract visual features $\boldsymbol{F} = \{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_N\} \in \mathbb{R}^{N \times d}$:

$$\boldsymbol{f}_n = \text{ImageEncoder}(E_I(v_n)), \quad n \in [1, N], \tag{2}$$

where $E_I(v_n)$ produces the patch embeddings of v_n. After encoding, $\boldsymbol{T}$ and $\boldsymbol{F}$ are fed into the TAD network and interact with each other, as we will introduce later.

3) Prompting CLIP for Better Representations: To effectively adapt CLIP for a specific domain, the common practice [17], [43] is to insert learnable continuous prompts $\boldsymbol{P}_T \in \mathbb{R}^{C \times d_T}$ before the text encoder of CLIP, where C denotes the number of soft prompts and d_T the text model dimension. Thus, Eq. (1) is modified as follows:

$$\boldsymbol{t}_k = \text{TextEncoder}([\boldsymbol{P}_T; E(y_k)]), \quad k \in [1, K], \tag{3}$$

where $[;]$ denotes the concatenation operation along the sequence dimension.

B. Multimodal Prompt Learning

Unlike prior text-only prompting methods indicated by Eq. (3), we propose multimodal prompting to achieve the maximum use of CLIP's knowledge. Our proposal involves two core designs: 1) *branch- and layer-wise prompting* that adjusts the intermediate activations of image and text encoders of CLIP to learn new knowledge efficiently and 2) *vision-language prompt coupling* that enables early interactions between vision and text modalities by using prompts as a bridge.

1) Branch- and Layer-Wise Prompting: Inspired by the literature [17], [38], [43], [69], we highlight the importance of deep prompt learning. We propose to insert learnable soft prompts in the initial J layers of both branches of CLIP. These hierarchical prompts aim to amplify CLIP's embedded knowledge to learn contextual representations suitable for ZS-TAD. Let's denote the CLIP text encoder is composed of L_T blocks and one projection head. We modify Eq. (1) and decompose it into the following two equations:

$$[\widetilde{\boldsymbol{P}}_T^{(j)}; \widetilde{\boldsymbol{W}}^{(j)}] = \begin{cases} \text{Block}_T^{(j)}([\boldsymbol{P}_T^{(j)}; E(y_k)]), & j = 1 \\ \text{Block}_T^{(j)}([\boldsymbol{P}_T^{(j)}; \widetilde{\boldsymbol{W}}^{(j-1)}]), & j \in [2, J] \\ \text{Block}_T^{(j)}([\widetilde{\boldsymbol{P}}_T^{(j-1)}; \widetilde{\boldsymbol{W}}^{(j-1)}]), & j \in (J, L_T] \end{cases} \tag{4}$$

$$\boldsymbol{t}_k = \text{Head}_T(\widetilde{\boldsymbol{w}}_{[\text{EOS}]}^{(L_T)}), \tag{5}$$

where $\boldsymbol{P}_T^{(j)} \in \mathbb{R}^{C \times d_T}$ means the soft prompts inserted into the j-th text block and the final text feature $\boldsymbol{t}_k$ of the k-th class is obtained from the position of the special token [EOS]. Similarly, for the image branch, let's assume the CLIP image encoder comprises of L_I blocks and one projection head. We modify Eq. (2) and decompose it into the following two equations:

$$[\widetilde{\boldsymbol{P}}_I^{(j)}; \widetilde{\boldsymbol{V}}^{(j)}] = \begin{cases} \text{Block}_I^{(j)}([\boldsymbol{P}_I^{(j)}; E(v_n)]), & j = 1 \\ \text{Block}_I^{(j)}([\boldsymbol{P}_I^{(j)}; \widetilde{\boldsymbol{V}}^{(j-1)}]), & j \in [2, J] \\ \text{Block}_I^{(j)}([\widetilde{\boldsymbol{P}}_I^{(j-1)}; \widetilde{\boldsymbol{V}}^{(j-1)}]), & j \in (J, L_I] \end{cases} \tag{6}$$

$$\boldsymbol{f}_n = \text{Head}_I(\widetilde{\boldsymbol{v}}_{[\text{CLS}]}^{(L_I)}), \tag{7}$$

where $\boldsymbol{P}_I^{(j)} \in \mathbb{R}^{C \times d_I}$ means the soft prompts inserted into the j-th image block and the final visual feature $\boldsymbol{f}_n$ of the n-th frame is obtained from the position of the special token [CLS].

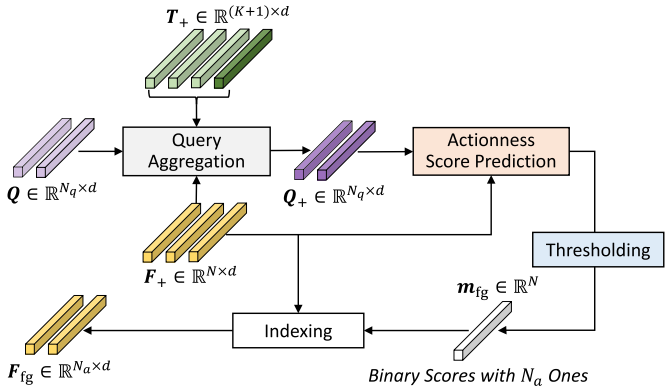


Fig. 3. Illustration of the proposed text-enhanced actionness modeling. It comprises query aggregation, actionness score prediction, thresholding and indexing to obtain the final foreground action features.

2) Vision-Language Prompt Coupling: Instead of using the disjoint image and text prompts, we propose to establish cross-modal connections between two branches by learning image prompts $\{\boldsymbol{P}_I^{(j)}\}_{j=1}^{J}$ from text prompts $\{\boldsymbol{P}_T^{(j)}\}_{j=1}^{J}$ with a fully-connected (FC) layer acting as V-L prompt coupling:

$$\boldsymbol{P}_I^{(j)} = \text{FC}(\boldsymbol{P}_T^{(j)}), \quad j \in [1, J]. \tag{8}$$

V-L prompt coupling is mapping of $\boldsymbol{d}_\text{T}$ dimensional text prompts to $\boldsymbol{d}_\text{I}$ of vision prompts. By doing so, we encourage mutual gradient propagation and interaction of two modalities, ensuring mutual synergy to optimize pre-trained CLIP. The reason to choose an explicit conditioning image prompt on text prompt is computational cost and trend [17], [43] as the processing of text is quite lighter than an image. Note: the dimension of vision prompts is always greater than the dimension of text prompts.

C. mProTEA

Besides the aforementioned multimodal prompt learning, mProTEA also contains several modules dedicated to TAD, as introduced next.

1) Temporal Modeling: Given the visual features $\boldsymbol{F} \in \mathbb{R}^{N\times d}$ produced by Eq. (7), we employ Transformer [70] to learn the temporal dependencies among N frames, resulting in temporal-aware visual features $\boldsymbol{F}_+ \in \mathbb{R}^{N\times d}$.

2) Text-Enhanced Actionness Modeling: Actionness refers to the likelihood of containing a human action for each video snippet. Different from the prior work [8] that uses N_q virtual queries with ambiguous meanings to aggregate visual features $\boldsymbol{F}_+ \in \mathbb{R}^{N\times d}$ for actionness modeling, we propose to create action-relevant queries from textual semantics, thereby suppressing background interference. The diagram of our proposal is illustrated in Figure 3. Let denote the original query features with no explicit meanings as $\boldsymbol{Q} \in \mathbb{R}^{N_q\times d}$, we refine $\boldsymbol{Q}$ with video- and text-related semantics as follows:

$$\boldsymbol{Q}_+ = \text{Aggregation}(\boldsymbol{Q}, \boldsymbol{F}_+, \boldsymbol{T}_+) \in \mathbb{R}^{N_q\times d}, \tag{9}$$

where for $\boldsymbol{q}_+ \in \boldsymbol{Q}_+$ and $\boldsymbol{q} \in \boldsymbol{Q}$, the aggregation process is formulated as:

$$\boldsymbol{q}_+ = \boldsymbol{q} + \text{FC}([\text{Mean}(\boldsymbol{F}_+); \text{Mean}(\boldsymbol{T}_+)]), \tag{10}$$

where Mean$(\cdot)$ denotes mean pooling. With aggregated action queries $\boldsymbol{Q}_+$, we then predict actionness scores $\boldsymbol{m}$ as follows:

$$\boldsymbol{m} = \text{FC}(\sigma(\boldsymbol{Q}_+ \cdot \boldsymbol{F}_+^\top)) \in \mathbb{R}^N, \tag{11}$$

where σ denotes the sigmoid activation and FC$(\cdot)$ is a fully connected layer. Next, we convert $\boldsymbol{m}$ into grayscale using a threshold θ_grey to obtain binary class-agnostic actionness scores, denoted as $\boldsymbol{m}_\text{fg}$. The obtained $\boldsymbol{m}_\text{fg}$ aids the model in emphasizing crucial video activity features. Therefore, we can obtain foreground action features $\boldsymbol{F}_\text{fg} \in \mathbb{R}^{N_a\times d}$ by indexing $\boldsymbol{F}_+$ based on $\boldsymbol{m}_\text{fg}$, where $N_a \leq N$. This yields class-agnostic foreground action features initially optimized for seen classes, yet designed for robust generalization to unseen classes.

3) Actionness-Guided Refinement: Instinctively, incorporating action priors will likely augment the text and visual representations, leading to a richer and more comprehensive understanding of the data. Motivated by this, we introduce Text Refinement Transformer (TRT) to perform cross-attention between text features $\boldsymbol{T}_+$ and action features $\boldsymbol{F}_\text{action}$ in the following (query, key, value) configuration:

$$\hat{\boldsymbol{T}}_+ = \text{TRT}(\boldsymbol{T}_+, \boldsymbol{F}_\text{fg}, \boldsymbol{F}_\text{fg}) \in \mathbb{R}^{(K+1)\times d}, \tag{12}$$

where $\boldsymbol{T}_+$ is designed to find related action clues across foreground snippets. Similarly, we refine visual features $\boldsymbol{F}_+$ via Visual Refinement Transformer (VRT) as follows:

$$\hat{\boldsymbol{F}}_+ = \text{VRT}(\boldsymbol{F}_+, \boldsymbol{F}_\text{fg}, \boldsymbol{F}_\text{fg}) \in \mathbb{R}^{N\times d}. \tag{13}$$

4) Action Classification and Localization: To achieve the classification, we first compute the dot product of refined text and visual features:

$$\boldsymbol{S} = \hat{\boldsymbol{T}}_+ \cdot \hat{\boldsymbol{F}}_+^\top \tag{14}$$

where $S \in \mathbb{R}^{(K+1)\times N}$. Then, we apply the SoftMax activation on the first dimension of $\boldsymbol{S}$ to obtain $\boldsymbol{P}$, where each element $p_{j,i} \in \boldsymbol{P}$ denotes the possibility of the i-th a temporal snippet belonging to the j-th class.

To achieve action localization, we follow previous works [6], [8], so this branch predicts 1-D masks of action instances across the entire temporal span of the video. These masks are conditioned on the temporal location i, therefore leveraging the dynamic convolution filters [71] to learn the context of action (foreground) and background instances at each snippet location individually. To clarify, given the i-th snippet $(\hat{\boldsymbol{F}}_+)$, the dynamic decoupled filters head outputs a 1-D mask vector $\mathcal{A} = [k_1, \ldots, k_T] \in \mathbb{R}^{T\times 1}$, where each element $k_o \in [0, 1]$ indicates the foreground probability of the i-th snippet. This is achieved through a stack of three 1-D dynamic convolution layers C_m, to predict action instances across the entire temporal range of the video as follows:

$$\boldsymbol{A} = \text{sigmoid}(\text{LocalizationHead}(C_m(\hat{\boldsymbol{F}}_+)) \in \mathbb{R}^N, \tag{15}$$

where i-th column of A is the temporal action mask vector showcases the action probability of a specific i-th video snippet, and C_m is a stack of 3 convolution layers [71] implemented as a localization head.

D. Two-Stage Training

The training procedure of our mProTEA is carried out in two distinct stages to reduce GPU memory costs and save computation resources.

In the first stage, we freeze the pre-trained CLIP and conduct multimodal prompt learning (Section III-B) on an image recognition dataset with rich concepts (e.g., ImageNet). Given the text features of all K classes $\boldsymbol{T} \in \mathbb{R}^{K\times d}$ using the template "a photo of [CLASS]", the feature of an input image $\boldsymbol{f} \in \mathbb{R}^{d}$, and the corresponding label y, We train the model by minimizing the following Cross-Entropy:

$$L_1 = \text{CrossEntropy}(\text{SoftMax}(\boldsymbol{T} \cdot \boldsymbol{f}), y). \tag{16}$$

In the second stage, we keep the pre-trained CLIP fixed as usual and also freeze the multimodal prompts learned in the first stage. We train the model on the training set of a TAD dataset D_{train} in a standard supervised learning manner.

To train mProTEA, ground-truth data is organized into the designed format. Each snippet within a training video is labeled with the corresponding action class or as background, based on temporal intervals and class labels. Action snippets are assigned instance-specific binary masks, with all snippets of the same action instance sharing the same mask. Following [8], we use comprehensive training objectives that combine 1) the Cross-Entropy loss L_{ce} between the predicted probabilities $\boldsymbol{P} \in \mathbb{R}^{(K+1)\times N}$ and the ground-truth label for action recognition, 2) the binary dice loss L_m between the predicted mask $\hat{\boldsymbol{m}} \in \mathbb{R}^{N}$ and the ground-truth mask for action localization, 3) the action completeness loss L_{comp} to predict foreground masks that closely match the complete temporal extent of the action instances $\boldsymbol{m}_{fg}$, as the binary cross-entropy between the binarized predicted foreground mask ($\mathbf{m}_{fg}$) and the ground-truth one-hot foreground mask ($\mathbf{g}$):

$$\mathcal{L}_{comp} = -\left(\mathbf{g} \cdot \log(\mathbf{m}_{fg}) + (1-\mathbf{g}) \cdot \log(1 - t\mathbf{m}_{fg})\right) \tag{17}$$

and 4) the inter-branch consistency L_{const} to ensure foreground consistency between the classification and localization branches. This loss encourages the classification features ($\hat{\mathbf{F}}_{clf}$) to be similar to the features derived from the predicted foreground mask ($\mathbf{m}_{fg}$), computed as the cosine similarity:

$$\mathcal{L}_{const} = 1 - \text{cosine}(\mathbf{F}_{clf}, \mathbf{m}_{fg}) \tag{18}$$

where F_{clf} is the features obtained from top-scoring $topk$ foreground snippets obtained classification output same as [8]. Therefore, the overall loss function in the second phase is formulated as follows:

$$L_2 = L_{ce} + L_m + L_{\text{comp}} + L_{\text{const}}. \tag{19}$$

By disentangling the optimization of multimodal prompts and the TAD network, our mProTEA can achieve ZS-TAD efficiently.

IV. Experiments

A. Datasets

mProTEA introduces a novel two-stage training approach to enhance model generalization and training efficiency. In the first stage, we learn multimodal prompts on one of 11 diverse image recognition datasets: ImageNet [22], Caltech101 [72], Oxford-Pets [73], Stanford-Cars [74], Flowers102 [75], Food101 [76], FGVC-Aircraft [77], SUN397 [78], UCF-101 [79], DTD [80], and Euro-Sat [81]. This stage improves the model's ability to recognize various objects and scenes, fostering robustness. In the second stage, we conduct training of the rest model on two widely-used temporal action detection benchmarks: ActivityNet [23], and THUMOS14 [24]. ActivityNet comprises 19,994 videos across 200 action classes and is split into 2:1:1 training, validation, and testing sets. THUMOS14 comprises 200 validation videos and 2 testing videos across 20 categories, each labeled with temporal boundaries and action classes.

B. Implementations

Stage 1: We use a pre-trained ViT-B/16 CLIP model for prompt optimization with $\boldsymbol{d}_V = 768$, $\boldsymbol{d}_T = 512$ and $\boldsymbol{d}_{\text{VT}} = 512$. The prompt depth is set to 9, and the prompt length is fixed at 3. We train models for 5 epochs with a batch size of 4 and a learning rate of 0.0026. Training was performed on a single NVIDIA A100 GPU using the SGD optimizer. The first layer of text prompts uses pre-trained CLIP word embeddings for the template "a video of [CLASS]", whereas the rest prompts are randomly initialized from a normal distribution. Consistent hyperparameters are maintained for fair comparisons and reliable evaluations. *Stage 2*: Video frames are pre-processed to 224×224 spatial resolution before encoding. The feature sequence $\boldsymbol{F}$ of each video is rescaled to $T = 100/256$ snippets for ActivityNet and THUMOS14 using linear interpolation. During training, our model undergoes 15 epochs using the Adam optimizer, with a learning rate of 10^{-4} for ActivityNet and 10^{-5} for THUMOS14, respectively.

C. Comparison With State-of-the-Art Methods

1) Zero-Shot Setting: The action categories of training and testing sets are mutually exclusive in this setting. We follow the dataset splits and evaluation settings by [8] and [9]. Specifically, we conduct two evaluation settings on THUMOS14 and ActivityNet: 1) training with 75% action categories and testing on the remaining 25%; 2) training with 50% categories and testing on the other 50%. We carry out 10 runs for each evaluation setting to confirm statistical worth.

Competitors: As ZS-TAD is a relatively new problem, we have limited and competitive works as [8], [9] and [82], [83], [84], [85] for fair comparison based on similar settings (data-split, related approach). Particularly, the absence of open-source implementation necessitates replicating their reported [9] baselines for our evaluation, similar to [8]. Two baselines are introduced by extended existing TAD methods using CLIP as B-I and B-II.B-I represents a two-stage TAD baseline using BMN [86] with the proposal generator and CLIP. B-II: represent one-stage baseline using CLIP + TAD. The text encoders remain identical for both baselines using CLIP pre-training weights. Unfortunately, we cannot compare with the earlier zero-shot TAD method, ZS-TAD [7], due to the unavailability of code and a lack of common data-split between [7] and [9].

TABLE I

RESULTS OF ZERO-SHOT TEMPORAL ACTION DETECTION ON THUMOS14 AND ACTIVITYNET DATASETS. IN THIS SETTING, WE TRAIN AND TEST MODELS WITH A PORTION OF SEEN AND UNSEEN ACTION CATEGORIES. *CLIP+ PRETRAINED VISUAL MODEL, **CLIP+LLM

Train Split	Methods	THUMOS 14						ActivityNet v1.3			
		0.3	0.4	0.5	0.6	0.7	Avg	0.5	0.75	0.95	Avg
75% Seen 25% Unseen	B-II	30.6	22.4	18.2	11.4	7.2	17.9	34.8	19.8	4.3	20.2
	B-I	35.1	27.5	20.4	13.5	6.8	21.9	36.6	20.7	3.5	22.5
	EffPrompt	39.7	31.6	23	14.9	7.5	23.3	37.6	22.9	3.8	23.1
	STALE	40.5	32.3	23.5	15.3	7.6	23.8	38.2	25.2	6	24.9
	mProTEA (Ours)	**43.1**	**38.2**	**28.2**	**18.1**	**8.7**	**27.91**	**44.5**	**27.4**	**7.9**	**27.6**
	3 Encoder Models (CLIP + Extra Encoder)										
	MP-TAL*	46.3	39	29.5	18.3	8.7	28.4	42	25.8	3.2	25.9
	ZEETAD**	61.4	53.9	44.7	34.5	20.5	43.2	51	33.4	5.9	32.5
50% Seen 50%Unseen	B-II	21	16.4	11.2	6.3	3.2	11.6	25.3	13	3.7	12.9
	B-I	27.2	21.3	15.3	9.7	4.8	15.7	28	16.4	1.2	16
	EffPrompt	37.2	29.6	21.6	14	7.2	21.9	32	19.3	2.9	19.6
	STALE	38.3	30.7	31.2	13.8	7	22.2	32.1	20.7	5.9	20.5
	MUPPET	40.1	-	22.8	-	8.1	24.8	33.5	21.9	6.7	22
	UNLoc	36.9	-	-	-	-	-	-	-	-	-
	mProTEA (Ours)	41.2	36.3	26.3	16.8	8.4	26.1	**41.8**	**24.6**	**6.1**	**25.6**
	3 Encoder Models (CLIP + Extra Encoder)										
	MP-TAL*	42.3	34.7	25.8	16.8	7.5	25.3	34.3	20.8	3	21
	ZEETAD**	**45.2**	**38.8**	**30.8**	**22.5**	**13.7**	**30**	39.2	25.7	3.1	24.9

Performance: The performance of various methods in zero-shot temporal action detection (ZS-TAD) experiments is detailed in Table I. Our proposed mProTEA demonstrates strong performance across both THUMOS14 and ActivityNet v1.3 datasets. In the 75% seen and 25% unseen split, mProTEA achieves the highest average mAP of 27.6 on THUMOS14 and 27.91 on ActivityNet, surpassing all other methods. Even in the more challenging 50% seen and 50% unseen split. Interestingly, the one-stage baseline (B-II) exhibits a more significant performance gap than the two-stage baseline (B-I), with a substantial decrease in mAP as the amount of labeled data increases. This suggests the potential for the one-stage baseline to improve with more data. Moreover, it also reveals that localization uncertainty propagation can be problematic in low-data scenarios. STALE's performance also drops on THUMOS14 with stricter metrics, possibly due to the localization head struggling with foreground imbalance. while mProTEA may exhibit a partial performance decrease compared to the 3 encoder models, it remains competitive, particularly on the THUMOS14 dataset. This indicates that mProTEA competes effectively with these models in scenarios where the data skew is not as pronounced, such as in the 75%-25% split. It's noteworthy that the complexity of the ActivityNet v1.3 dataset may contribute to this partial performance decrease. Nevertheless, mProTEA still outperforms baseline methods B-II, B-I, EffPrompt, STALE, and MUPPET across both splits, showcasing strong generalizability and a class-agnostic property by learning multimodalby leveraging a pre-trained large multimodal model for ZS-TAD tasks.

2) Conventional Supervised Setting: Training and testing sets consist of identical action categories in this setting. We adhere to the exact dataset splits in the existing literature to ensure a fair and consistent comparison.

Competitors: Besides STALE [8] and EffPrompt [9], we additionally consider seven representative TAD methods using the I3D [87] encoder backbone: TALNet, GTAN, MUSES, VSGN, Context-Loc, and BU-TAL. We also create three baselines: B-I (CLIP + sequential-TAD), B-II (CLIP + parallel-TAD), and B-III (CLIP single stage+ visual encoder replaced with Kinetics pretrained Video-Encoder I3D).

Performance: Analysis of the results in Table II unequivocally establishes the superior performance of our proposed approach across diverse settings. Notably, our method consistently outperforms existing TAD methods across different modes, including "RGB + Flow" and "RGB," as well as with varying encoder backbones, encompassing both I3D and CLIP. This consistency in superior performance underscores the efficacy and adaptability of our approach in the domain of temporal action detection. Particularly, our mProTEA consistently outperforms not only STALE but also its counterpart model B-III, utilizing the I3D backbone. On the THUMOS14 dataset, mProTEA achieves a substantial mAP improvement of 1.9 over B-III. Similarly, on the ActivityNet dataset, this improvement widens to 2.1. These findings underscore the efficacy of our multimodal approach and text-enabled action modeling to enhance temporal action detection accuracy. Furthermore, Table II bottom half depicts the performance comparison of models utilizing I3D and CLIP backbones, where mProTEA leads with the A2Net model by a huge margin. This highlights the importance of selecting an appropriate architecture to be more effective for temporal action detection tasks. In essence, our comparative analysis reaffirms the effectiveness of mProTEA, emphasizing the crucial role of multimodal prompts and the selection of an optimal backbone architecture in achieving state-of-the-art performance in temporal action detection. The results presented in this table provide valuable insights into the relative strengths of the investigated models, demonstrating the advantages of the mProTEA approach over alternative backbone architectures for this particular task. By carefully examining these performance metrics, researchers and practitioners can make informed decisions about which models and techniques to prioritize when tackling temporal action detection challenges.

3) Qualitative Analysis of mProTEA: Figure 4 presents six qualitative examples, with ZS-TAD results from ground truths,

TABLE II

RESULTS OF SUPERVISED TEMPORAL ACTION DETECTION ON THUMOS14 AND ACTIVITYNET DATASETS. IN THIS SETTING, THE ACTION CATEGORIES OF THE TRAINING SET ARE IDENTICAL TO THOSE OF THE TESTING SET

Methods	Mode	Encoder Backbone	THUMOS14						ActivityNet			
			0.3	0.4	0.5	0.6	0.7	Avg	0.5	0.75	0.95	Avg
TALNet	RGB+Flow	I3D	53.2	48.5	42.8	33.8	20.8	39.8	38.2	18.3	1.3	20.2
GTAN	RGB+Flow	P3D	57.8	47.2	38.8	-	-	-	52.6	34.1	8.9	34.3
MUSES	RGB+Flow	I3D	68.9	64.0	56.9	46.3	31.0	53.4	50.0	34.9	6.5	34.0
VSGN	RGB+Flow	I3D	66.7	60.4	52.4	41.0	30.4	50.1	52.3	36.0	8.3	35.0
Context-Loc	RGB+Flow	I3D	68.3	63.8	54.3	41.8	26.2	-	56.0	35.2	3.5	34.2
BU-TAL	RGB+Flow	I3D	53.9	50.7	45.4	38.0	28.5	43.3	43.5	33.9	9.2	30.1
STALE	RGB+Flow	I3D	68.9	64.1	57.1	46.7	31.2	52.9	56.5	36.7	9.5	36.4
B-III	RGB+Flow	I3D	70.1	66.2	58.4	47.4	32.3	53.5	57.2	37.1	9.7	37.2
mProTEA (Ours)	**RGB+Flow**	**I3D**	**72.2**	**67.8**	**59.3**	**49.2**	**33.7**	**55.4**	**58.9**	**38.4**	**10.1**	**39.1**
TALNet	RGB	I3D	42.6	–	31.9	–	14.2	–	–	–	–	–
A2net	RGB	I3D	45.0	40.5	31.3	19.9	10.0	29.3	39.6	25.7	2.8	24.8
B-I	RGB	CLIP	36.3	31.9	25.4	17.8	10.4	24.3	28.2	18.3	3.7	18.2
B-II	RGB	CLIP	57.1	49.1	40.4	31.2	23.1	40.2	51.5	33.3	6.6	32.7
EfficentPromot	RGB	CLIP	50.8	44.1	35.8	25.7	15.7	34.5	44.0	27.0	5.0	27.3
STALE	RGB	CLIP	60.6	53.2	44.6	36.8	26.7	44.4	54.3	34.0	7.7	34.3
mProTEA (Ours)	RGB	CLIP	**64.4**	**56.5**	**47.7**	**40.5**	**27.9**	**47.3**	**55.4**	**35.4**	**8.4**	**35.6**

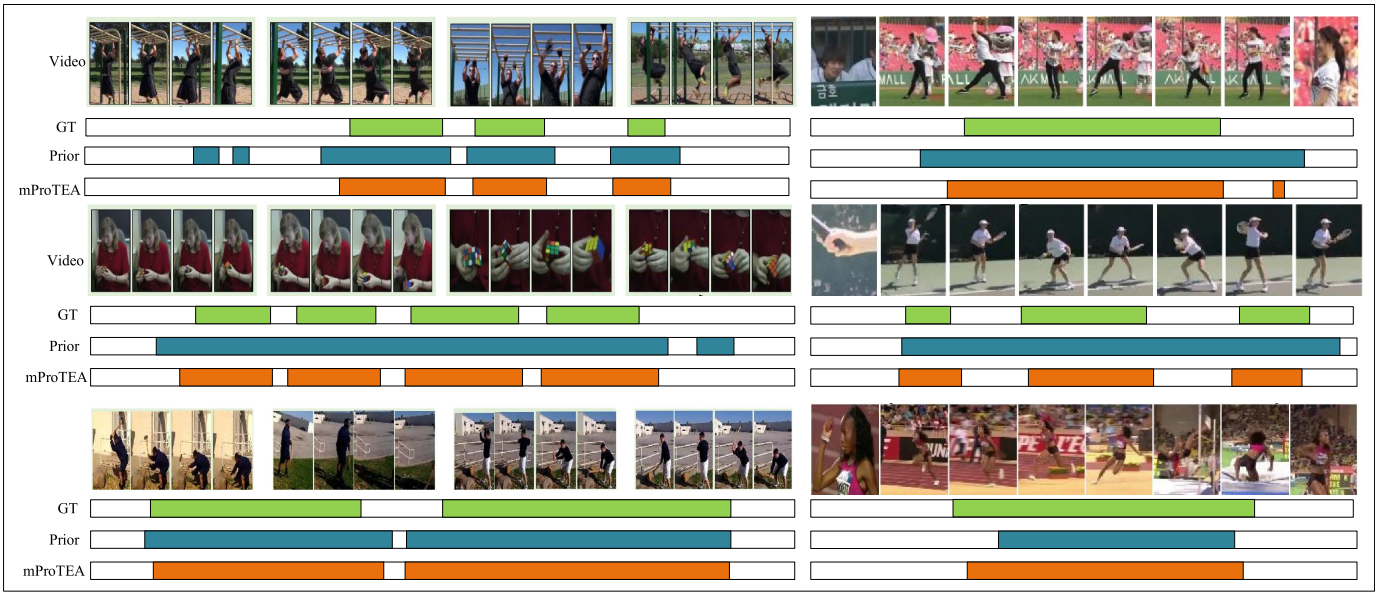


Fig. 4. Qualitative comparison of ZS-TAD results between mProTEA and prior work (i.e. STALE). "GT" is short for ground truth. The Left and right examples are from Activitynet and THUMOS, respectively. Compared with STALE, our mProTEA mitigates the over-prediction of background snippets and the miss-prediction of action activities.

the prior work – STALE, and our mProTEA. We can observe that mProTEA effectively resolves the issues of the prior work and contributes to a reduction in the over-prediction of background snippets and correction of miss-prediction instances. These visualized results emphasize the potential of mProTEA to enhance the precision and reliability of ZS-TAD.

4) Impact of Image-Text Interaction and Action Prior: In this section, we delve into the crucial discussion of image-text interaction in large pre trained model CLIP and action prior in the context of temporal action detection (TAD). Our analysis highlights their significance in enhancing model performance and providing a deeper understanding of the underlying dynamics in TAD systems. Our empirical analysis, as depicted in Table II, highlights the significance of these factors in determining the performance of TAD models. When comparing mProTEA with its counterpart, B-III, utilizing the same I3D backbone pre-trained on Kinetics, the disparity in performance can be attributed to the efficacy of action modeling within the TAD head. B-III's weaker action modeling leads to suboptimal detection in untrimmed videos, underscoring the importance of robust TAD modeling. Similarly, the absence or strength of image-text interaction becomes evident when comparing our method with others in Table I. Models like STALE and UNloc solely leverage text information, resulting in inferior visual-text interaction and performance. In contrast, later methods like MUFFET strive to enhance multimodal interaction through prompting and meta-learning, albeit with limited success. Additionally, our comparison with MP-TAL in Table I reveals the trade-off between computational feasibility and accuracy. While MP-TAL leverages triple encoders for improved accuracy, it comes at the cost of computational complexity. In contrast, mProTEA's dual encoder approach

TABLE III

ANALYSIS OF COMPUTATIONAL COMPLEXITY OF OUR MPROTEA WITH THE STATE-OF-THE-ART METHODS

Sr	METHOD		Paras(M)	Training (hours)	Inference	Epoch
1	Eff prompt		41.3	3.24 h	**2.25m**	15
2	STALE		50.2	3.96h	2.53m	15
3	mProTEA		**56.9**	**1.42h**	2.71m	**5**

TABLE IV

EFFECT OF LEARNING MULTIMODAL PROMPTS ON DIFFERENT IMAGE RECOGNITION DATASETS. WE CAN SEE THAT PROMPTS LEARNED ON IMAGENET EXHIBIT STRONG DOMAIN GENERALIZATION ABILITY

Dataset	mAp			
	@0.5	@0.75	@0.95	Average
Caltech101	41.2	28.2	5.6	25.1
Dtd	40.2	27.2	4.9	24.1
EuroSat	39.9	26.8	5.2	23.9
FGVC-Aircraft	39.8	25.7	4.8	23.1
Food101	40.7	26.8	5.5	24.1
Oxford_flowers	40.8	27.8	5.1	24.5
Oxford_Pets	40.2	26.8	4.9	24.0
Stanford_cars	40.7	27.7	5.1	24.6
Sun397	40.5	27.4	6.1	24.5
UCF101	40.7	28.2	5.4	24.9
ImageNet **(mProTEA)**	**41.8**	**29.1**	**6.1**	**25.6**

achieves competitive accuracy while maintaining computational efficiency.

5) Computational Complexity: The computational complexity analysis of our mProTEA model, as depicted in Table III, reveals several advantageous features. Notably, our model showcases remarkable efficiency in training, completing the process in a mere 1.42 hours. This efficiency is further underscored by its early convergence, requiring only 5 epochs compared to the 15 epochs needed by Efficent-prompt and STALE. Such swift convergence is facilitated by our innovative plug-and-play module, which integrates the TAD head with an optimized pre-trained CLIP model as a multimodal prompt with coupling. This specialized module not only streamlines the training process but also enhances the model's ability to extract meaningful representations from diverse data modalities. Despite possessing a slightly higher parameter count than STALE, our model maintains competitive inference times, further solidifying its appeal. This combination of efficient training, early convergence, and multimodal optimization makes mProTEA exceptionally well-suited for temporal action detection, while also highlighting its potential for facilitating advancements in other visual downstream tasks with its innovative approach.

D. Ablation Study

In the following, we delve deeper into various mProTEA's designs and conduct experiments on the "50% Seen and 50% Unseen" setting of ActivityNet.

1) Effective Datasets for Learning Multimodal Prompts: We here conduct experiments on eleven diverse image recognition datasets. From Table IV, observations reveal that not all datasets exhibit a generic nature. For instance, the EuroSat and FGVC-Aircraft datasets contain key content features that are unfamiliar to action-related activities, resulting in the poorest performance when transferring the learned multimodal prompts for ZS-TAD. By contrast, ImageNet, which holds a prominent position and is our preferred choice, outperforms other datasets in our framework. This superiority can be attributed to its rich visual concepts, which facilitate robust multimodal prompt learning.

2) Effective Prompting Method: Our mProTEA model bridges vision and text encoding via branch-aware multimodal prompting and prompt coupling. We consider four variants here: (a) no prompting, (b) deep vision prompting (DVP), (c) deep language prompting (DLP), and (d) independent vision and language prompting (I-VLP). Table V shows that (c) demonstrates improvement over (b), indicating better adaptation of CLIP where the semantic information learned in the language branch is richer than the visual branch. Although (d) further enhances the performance of ZS-TAD through integrating (b) and (c), it struggles to interchange information between vision and text branches. Based on (d), our approach (e) conditions vision prompts on the text ones and thus performs the best among all variants.

In addition, the computational complexity of different prompting methods plays a crucial role in the efficiency and practicality of implementing our model mProTEA and optimization of its backbone, the pre-trained CLIP. Table VI provides a comparative analysis of the computational complexities of different prompting methods employed in the mProTEA backbone. The metrics evaluated include GFLOPS (Giga Floating Point Operations Per Second), Parameters (in millions M), average Training time (in minutes m), and Frames Per Second (FPS). DLP has a relatively lower GFLOPS requirement (19) compared to other prompting methods. Conversely, DVP exhibits a notably higher GFLOPS requirement, suggesting intensive computational processing, leading to higher FPS. This indicates that the popular use of language prompts is common as a cheaper option. Notably, mProTEA with Independent Visual and Language Prompting (I-VLP) showcases a comparable computational demand to DVP while maintaining superior performance as well as training epoch (5, half as of DVP and DLP), as evidenced by Table VI. This signifies an optimized balance between computational complexity and performance. Moreover, mProTEA with prompt coupling demonstrates a slight increase in computational complexity compared to its counterpart with I-VLP. However, this marginal increment is justified by the improved performance and synergy achieved in visual and language encoders through coupling, as highlighted in Table V. Thus, it is imperative to strike a balance between computational demands alongside model efficacy to ensure the practical applicability and scalability of our proposed approach.

3) Performance Impact of Prompts in CLIP Pre-Training and mProTEA Training: In our article, we acknowledge differences in prompts used during pre-training of CLIP and training of mProTEA could indeed impact the performance of the model. Since CLIP is trained on web-based image-text pairs, while mProTEA is trained on the Imagnet dataset while fine-tuned on datasets (ActivityNet and THUMOS-14) with specific content and context relevant to temporal actions.

TABLE V
EFFECT OF DIFFERENT PROMPTING METHODS

Setting	Method	Map	
		@0.5	Avg
(a)	No Prompting	36.5	22.5
(b)	Deep Vision Prompting	39.9	23.9
(c)	Deep Language Prompting	40.6	24.7
(d) mProTEA	I- Visual and Language Prompting	41.5	25.4
(e) mProTEA	(d) + Coupling (Eq. (8))	**41.8**	**25.6**

TABLE VI
COMPUTATIONAL COMPLEXITY OF DIFFERENT PROMPTING METHODS ON MPROTEA BACKBONE

METHOD	GFLOPS	Parms(M)	Train(m)	FPS
DLP	19	0.0238	10.08m	13.8
DVP	167.8	0.3340	38.68m	64.5
mProTea (I-VLP)	168	0.3556	12.59m	62.5
mProTEA (w/ coupling)	168.1	0.3972	13.75m	60.2

TABLE VII
EFFECT OF DIFFERENT PROMPT INITIALIZATION METHODS

Setting	Method	Map	
		0.5	Avg
(a)	All Layers: "a video of"	41.2	25.2
(b)	All Layers: Random	41.6	25.4
(c) mProTEA	1st Layer: "a video of"; Rest: Random	**41.8**	**25.6**

TABLE VIII
EFFECT OF DIFFERENT TEMPORAL MODELING NETWORKS

Setting	Network	mAp	
		0.5	Avg
(a)	CNN	32.0	19.6
(b)	MS-TCN	35.4	22.1
(c) mProTEA	Transformer	**41.8**	**25.6**

TABLE IX
EFFECT OF DIFFERENT ACTIONNESS MODELING METHODS TO PROVIDE ACTION PRIORS

Setting	Method	mAp	
		0.5	Avg
(a)	No actionness modeling	22.5	11.3
(b)	1-D CNN	36.9	22.8
(c)	$\boldsymbol{Q}$ as the Queries (w/o Eq. (9))	38.3	24.3
(d) mProTEA	$\boldsymbol{Q}_+$ as the Queries (w/ Eq. (9))	**41.8**	**25.6**

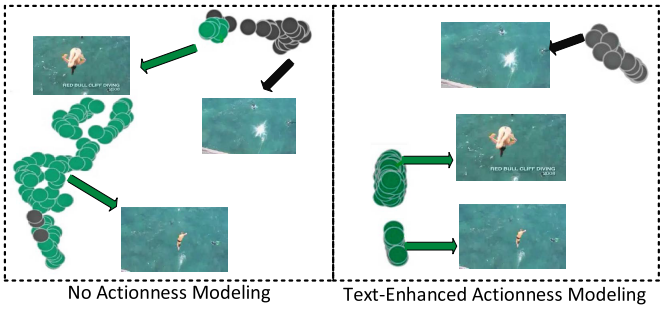


Fig. 5. UMAP visualization of learned features in a "Cliff Diving" video. Green dots represent action segments, while gray dots signify background segments. Our approach learns well-aligned features with text-enhanced actionness modeling.

Therefore, the prompts learned during mProTEA training may be more specialized and optimized for recognizing actions within videos.

Experiments in Table IV and V also validate successful domain adoption and the boost in performance of mProTEA compared to original CLIP in temporal action detection tasks. Especially IV highlights the significance of dataset characteristics in prompt learning. This suggests that the specialized prompts learned during mProTEA, aligned with the temporal action detection task at hand, lead to enhanced performance compared to generic prompts from CLIP pre-training. Additionally, mProTEA addresses the resource-intensive challenge (data collection) of video tasks as video-text pairs are harder to collect than image-text pairs and may suffer from misalignment. Computational Demands: Video tasks require more computational power than image tasks. Leveraging pre-trained models like mProTEA enhances computational efficiency. Temporal Dependencies: Videos have temporal dependencies effectively captured by image-based models like mProTEA.

4) Effective Prompt Initialization Method: In addition to prompt design, the effectiveness of prompt initialization significantly impacts the performance of our mProTEA model, as demonstrated in Table VII. Three different initialization strategies are compared:

1) All layers initialized with "a video of"
2) Random initialization for all layers
3) Only the first layer initialized with "a video of", with the rest initialized randomly

The results indicate that the approach where only the first layer is initialized with "a video of [action]" while the rest are initialized randomly yields the best performance, achieving a mean average precision (mAP) of 41.8 (Table VII).

Interpreting these results, we find that initializing all layers with the same generic template (a) leads to inferior performance, suggesting redundancy as prompts in higher layers may redundantly learn concepts already captured in the first layer. Conversely, complete random initialization surprisingly provides competitive performance, indicating the model's ability to learn effective prompts independently. The combination of a specific template in the first layer (providing initial guidance) and random initialization in subsequent layers (allowing for adaptation) seems to strike an optimal balance, resulting in the best performance for mProTEA. This outcome underscores the hierarchical nature of prompt learning [17], [43], as well in implementation, keeping the number of learnable prompts is fewer than the total words in the initial template, then only the first-word embeddings are considered learnable, with the remaining word embeddings treated as fixed inputs to the text encoder. Therefore, a proper prompt initialization method is crucial to enhance our model's capability.

5) Effective Temporal Modelling Network: We assess the selection of Transformer along with two alternatives for temporal modeling: (a) a one-dimensional CNN comprising two layers, each with dilation rates of 1, 3, and 5, and (b) a multi-stage Temporal CNN known as MS-TCN [88]. All of these options can be tailored to fit our approach easily. Table VIII

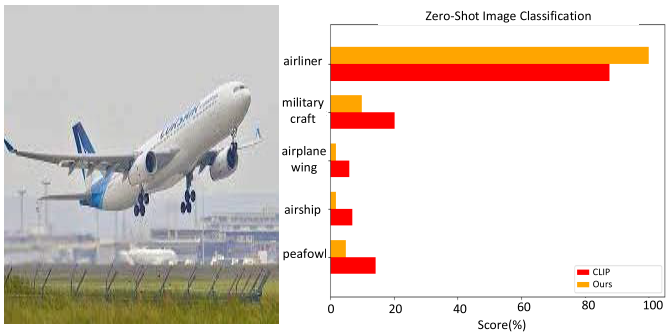


Fig. 6. Comparison of CLIP and CLIP with the proposed multimodal prompt learning on zero-shot image classification. Our approach enhances image-text alignment effectively, boosting the prediction of "airliner" while reducing irrelevant predictions like "airship" and "peafowl".

shows that Transformer demonstrates a better ability to learn long-range dependencies than CNN variants. Note, that we follow the transformer without positional encoding as the previous work [8] does.

6) Effective Actionness Modeling Method: To investigate the role of actionness modeling (AM), we here consider four variants: (a) the absence of AM, (b) AM achieved by a standard 1D-CNN approach, (c) AM achieved by using meaningless queries and temporal features, and (d) AM enhanced by text clues (i.e., using $\boldsymbol{Q}_+$ as the queries, indicated by Eq. (9)). As we can observe from Table IX, the setting (a) obtains extremely poor performance, showing the importance of action prior. Moreover, our approach (d) surpasses settings (b) and (c), indicating that leveraging textual semantics can offer more accurate action prior. In addition to quantitative analysis, we employ UMAP visualization to underscore the impact of the AM on action detection in Figure 5. As we can see, without our text-enhanced AM, action segments are often mistakenly included alongside background segments. In contrast, the influence of our approach is evident in correctly classifying action segment features.

7) Discussion on Efficiency: The efficiency of our model, mProTEA, is a crucial aspect to consider in its performance evaluation. Based on the data presented in presented in Table III and Table VI provides valuable insights into the efficiency of our mProTEA model compared to state-of-the-art methods. Firstly, in terms of model parameters and training time, mProTEA stands out as a competitive option and exhibits a favorable balance between model size, training time, and inference speed. Despite having a slightly larger parameter count compared to other methods, mProTEA significantly outperforms in training time, requiring only 1.42 hours, which is nearly three times faster than Efficent-Prompt and STALE, which require 3.24 hours and 3.96 hours respectively. mProTEA achieves a commendable inference time of 2.71 minutes, which is competitive with other methods while training efficiency is further underscored by its reduced epochs required for convergence. Moreover, when considering backbone optimization time separately, mProTEA still maintains its efficiency advantage. With only 13.7 minutes required for backbone optimization, mProTEA's overall training time remains substantially lower than competitors, showcasing its rapid convergence and effectiveness. The empirical evidence highlights mProTEA's superiority as an efficient approach for training Zero-Shot Temporal Action Detection (ZS-TAD) networks and positioning it as a promising solution for visual downstream tasks.

8) Discussion on Broader Implications: Although ZS-TAD is the target task in this paper, our approach can be extended to various tasks. Taking the zero-shot image classification task as an example, we compare CLIP with or without our proposed multimodal prompt learning in Fig 6. We can see that our model can align images and text well and produce more reasonable predictions than CLIP, e.g., higher response for the accurate class "airliner" and lower response for the irrelevant classes "airship" and "peafowl". This suggests that our approach can serve as a unified framework benefiting different tasks.

9) Limitation: Implementing the mProTEA model involves several considerations and trade-offs. Optimizing the mProTEA backbone on image datasets highlights the influence of dataset nature on domain adaptation and generalization. For temporal action detection got varying responses, among 11 datasets, ImageNet demonstrated the best performance. So this choice needs a trial method to choose a dataset. Balancing model complexity with computational efficiency is crucial. Exploring techniques to enhance CLIP backbone optimizing other than prompt and simplify the detection head is a priority. Addressing these challenges will refine the model and optimize resources for more efficient temporal action detection.

V. Conclusion

This paper has introduced mProTEA, an effective approach for zero-shot temporal action detection (ZS-TAD). mProTEA mitigates critical limitations in ZS-TAD by learning multimodal prompts for improved vision-text alignment and modeling text-enhanced actionness for accurate detection. mProTEA achieves state-of-the-art performance on benchmark datasets and meanwhile maintains efficiency due to the proposed step-wise training procedure. Besides, ablation studies reveal the importance of the source for multimodal prompt learning, appropriate prompt design, actionness modeling method, etc., all of which contribute to the superiority of mProTEA. Although we evaluate our approach solely on the CLIP model and the TAD task, we posit that our idea holds the potential to be adaptable to various vision-language pre-trained models and different tasks. In our future study, we will delve deeper into the role of prompts in cross-dataset transfer, with a specific focus on image-to-video transfer.

References

[1] M. Xu, J.-M. Perez-Rua, X. Zhu, B. Ghanem, and B. Martinez, "Low-fidelity end-to-end video encoder pre-training for temporal action localization," 2021, *arXiv:2103.15233*.

[2] M. Xu, C. Zhao, D. S. Rojas, A. Thabet, and B. Ghanem, "G-TAD: Sub-graph localization for temporal action detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10156–10165.

[3] S. Buch, V. Escorcia, C. Shen, B. Ghanem, and J. C. Niebles, "SST: Single-stream temporal action proposals," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6373–6382.

[4] L. Wang, Y. Xiong, D. Lin, and L. Van Gool, "UntrimmedNets for weakly supervised action recognition and detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4325–4334.

[5] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin, "Temporal action detection with structured segment networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2914–2923.

[6] S. Nag, X. Zhu, Y.-Z. Song, and T. Xiang, "Proposal-free temporal action detection via global segmentation mask learning," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 645–662.

[7] L. Zhang et al., "ZSTAD: Zero-shot temporal activity detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 876–885.

[8] S. Nag, X. Zhu, Y.-Z. Song, and T. Xiang, "Zero-shot temporal action detection via vision-language prompting," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 681–697.

[9] C. Ju, T. Han, K. Zheng, Y. Zhang, and W. Xie, "Prompting visual-language models for efficient video understanding," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 105–124.

[10] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[11] A. Sanghi et al., "CLIP-Forge: Towards zero-shot text-to-shape generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18582–18592.

[12] S. Subramanian, W. Merrill, T. Darrell, M. Gardner, S. Singh, and A. Rohrbach, "ReCLIP: A strong zero-shot baseline for referring expression comprehension," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 5198–5215.

[13] A. Zeng, "Socratic models: Composing zero-shot multimodal reasoning with language," in *Proc. Int. Conf. Learn. Represent.*, 2023, pp. 1–11.

[14] B. Yang, F. Liu, X. Wu, Y. Wang, X. Sun, and Y. Zou, "MultiCapCLIP: Auto-encoding prompts for zero-shot multilingual visual captioning," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics*, 2023, pp. 11908–11922.

[15] W. Jin, Y. Cheng, Y. Shen, W. Chen, and X. Ren, "A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models," 2021, *arXiv:2110.08484*.

[16] Y. Lu, J. Liu, Y. Zhang, Y. Liu, and X. Tian, "Prompt distribution learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 5206–5215.

[17] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, Sep. 2022.

[18] M. Shu et al., "Test-time prompt tuning for zero-shot generalization in vision-language models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 14274–14289.

[19] C. Zhang, M. Cao, D. Yang, J. Chen, and Y. Zou, "CoLA: Weakly-supervised temporal action localization with snippet contrastive learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16010–16019.

[20] M. Cao, C. Zhang, L. Chen, M. Z. Shou, and Y. Zou, "Deep motion prior for weakly-supervised temporal action localization," *IEEE Trans. Image Process.*, vol. 31, pp. 5203–5213, 2022.

[21] M. N. Rizve et al., "PivoTAL: Prior-driven supervision for weakly-supervised temporal action localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 22992–23002.

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Nov. 2009, pp. 248–255.

[23] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 961–970.

[24] H. Idrees et al., "The THUMOS challenge on action recognition for videos 'in the wild,"' *Comput. Vis. Image Understand.*, vol. 155, pp. 1–23, Feb. 2017.

[25] Z. Wang et al., "CAMP: Cross-modal adaptive message passing for text-image retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 5764–5773.

[26] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, vol. 37, 2015, pp. 2048–2057. [Online]. Available: https://dl.acm.org/doi/proceedings/10.5555/3045118

[27] B. Yang, M. Cao, and Y. Zou, "Concept-aware video captioning: Describing videos with effective prior information," *IEEE Trans. Image Process.*, vol. 32, pp. 5366–5378, 2023.

[28] S. Antol et al., "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2425–2433.

[29] Y. Mori, H. Takahashi, and R. Oka, "Image-to-word transformation based on dividing and vector quantizing images with words," in *Proc. 1st Int. Workshop Multimedia Intell. Storage Retr. Manage.*, 1999, pp. 1–9.

[30] A. Frome et al., "Devise: A deep visual-semantic embedding model," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 1–6.

[31] C. Jia et al., "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, M. Meila and T. Zhang, Eds. Jul. 2021, pp. 4904–4916.

[32] L. Yao et al., "FILIP: Fine-grained interactive language-image pre-training," 2021, *arXiv:2111.07783*.

[33] M. Wang, J. Xing, and Y. Liu, "ActionCLIP: A new paradigm for video action recognition," 2021, *arXiv:2109.08472*.

[34] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9879–9889.

[35] S. Paul, N. C. Mithun, and A. K. Roy-Chowdhury, "Text-based localization of moments in a video corpus," *IEEE Trans. Image Process.*, vol. 30, pp. 8886–8899, 2021.

[36] T. B. Brown et al., "Language models are few-shot learners," in *Proc. NIPS*, 2020, pp. 1877–1901.

[37] Y. Yao, A. Zhang, Z. Zhang, Z. Liu, T.-S. Chua, and M. Sun, "CPT: Colorful prompt tuning for pre-trained vision-language models," 2021, *arXiv:2109.11797*.

[38] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "MaPLe: Multi-modal prompt learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19113–19122.

[39] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," 2020, *arXiv:2012.15723*.

[40] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig, "How can we know what language models know?" *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 423–438, Dec. 2020.

[41] T. Schick and H. Schütze, "Exploiting cloze questions for few shot text classification and natural language inference," 2020, *arXiv:2001.07676*.

[42] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh, "AutoPrompt: Eliciting knowledge from language models with automatically generated prompts," 2020, *arXiv:2010.15980*.

[43] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16816–16825.

[44] T. Huang, J. Chu, and F. Wei, "Unsupervised prompt learning for vision-language models," 2022, *arXiv:2204.03649*.

[45] H. Bahng, A. Jahanian, S. Sankaranarayanan, and P. Isola, "Exploring visual prompts for adapting large-scale models," 2022, *arXiv:2203.17274*.

[46] Y. Chen, B. Guo, Y. Shen, W. Wang, W. Lu, and X. Suo, "Capsule boundary network with 3D convolutional dynamic routing for temporal action detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 2962–2975, May 2022.

[47] C. Yan et al., "Task-adaptive attention for image captioning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 43–51, Jan. 2022.

[48] J. Chen, Z. Wang, C. Zheng, K. Zeng, Q. Zou, and Z. Xiong, "Understanding dynamic associations: Gait recognition via cross-view spatiotemporal aggregation network," *IEEE Trans. Circuits Syst. Video Technol.*, 2022.

[49] L. Xu, X. Wang, W. Liu, and B. Feng, "Cascaded boundary network for high-quality temporal action proposal generation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3702–3713, Oct. 2020.

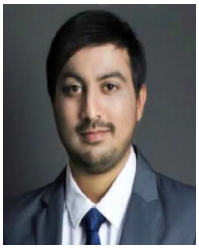
[50] J. Wang, W. Wang, and W. Gao, "Fast and accurate action detection in videos with motion-centric attention model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 117–130, Jan. 2020.

[51] X. Song, C. Lan, W. Zeng, J. Xing, X. Sun, and J. Yang, "Temporal–spatial mapping for action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 3, pp. 748–759, Mar. 2020.

[52] C. Wei, J. Zhao, W. Zhou, and H. Li, "Semantic boundary detection with reinforcement learning for continuous sign language recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 3, pp. 1138–1149, Mar. 2021.

[53] J. Gao et al., "Accurate temporal action proposal generation with relation-aware pyramid network," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2020, pp. 10810–10817.

[54] W. Sun, R. Su, Q. Yu, and D. Xu, "Slow motion matters: A slow motion enhanced network for weakly supervised temporal action localization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 1, pp. 354–366, Jan. 2023.
[55] D. Zhang, C. Li, F. Lin, D. Zeng, and S. Ge, "Detecting deepfake videos with temporal dropout 3DCNN," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 1288–1294.
[56] J. Gao, T. He, X. Zhou, and S. Ge, "Skeleton-based action recognition with focusing-diffusion graph convolutional networks," *IEEE Signal Process. Lett.*, vol. 28, pp. 2058–2062, 2021.
[57] J. Hu, L. Zhuang, W. Dong, S. Ge, and S. Wang, "Learning generalized representations for open-set temporal action localization," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 1987–1996.
[58] C. Gan, T. Yang, and B. Gong, "Learning attributes equals multi-source domain generalization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 87–97.
[59] S. Li, T. Liu, J. Tan, D. Zeng, and S. Ge, "Trustable co-label learning from multiple noisy annotators," *IEEE Trans. Multimedia*, vol. 25, pp. 1045–1057, 2023.
[60] M. Jain, J. C. van Gemert, and C. G. M. Snoek, "What do 15,000 object categories tell us about classifying and localizing actions?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 46–55.
[61] J. Liu, B. Kuipers, and S. Savarese, "Recognizing human actions by attributes," in *Proc. CVPR*, Jun. 2011, pp. 3337–3344.
[62] M. Jain, J. C. V. Gemert, T. Mensink, and C. G. M. Snoek, "Objects2action: Classifying and localizing actions without any video example," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4588–4596.
[63] Y. Li, S.-H. Hu, and B. Li, "Recognizing unseen actions in a domain-adapted embedding space," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 4195–4199.
[64] P. Mettes, W. Thong, and C. G. M. Snoek, "Object priors for classifying and localizing unseen actions," *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1954–1971, Jun. 2021.
[65] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 453–465, Mar. 2014.
[66] D. Parikh and K. Grauman, "Relative attributes," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 503–510.
[67] Y. Goldberg and O. Levy, "Word2vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method," 2014, *arXiv:1402.3722*.
[68] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
[69] M. Jia et al., "Visual prompt tuning," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2022, pp. 709–727.
[70] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–19.
[71] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11030–11039.
[72] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few examples: An unsupervised approach," in *Proc. Conf. Comput. Vision Pattern Recognit. Workshop*, 2004, pp. 1789–1798.
[73] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "A deep learning framework for object categorization," in *Proc. 3rd Int. Conf. Image, Vision Comput.*, 2012, pp. 2270–2277.
[74] J. Krause, M. Stark, J. J. Lim, J. P. H. Suh, and D. Koller, "3D object recognition without pose initialization: Real time scale performance," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2013, pp. 3986–3992.
[75] M.-E. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Oct. 2008, pp. 1–14.
[76] L. Bossard, M. Guillaumin, and L. V. Gool, "Learning multi-label classification models for large-scale image retrieval," in *Proc. Int. Conf. Multimedia Retr.*, 2014, pp. 35–48.
[77] S. Maji, A. C. Berg, and J. Malik, "A classification framework for aircraft images based on spatial and appearance features," *Remote Sens.*, vol. 10, no. 2, pp. 1549–1556, 2013.
[78] J. Xiao, J. Hays, and J. K. Aggarwal, "Learning high-level features of human actions with limited labeled data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2010, pp. 2057–2064.
[79] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*.
[80] M. Cimpoi, S. Maji, I. D. Reid, and D. G. Lowe, "Deep learning for pedestrian detection: A comparative review," *Multimedia Tools Appl.*, vol. 81, no. 10, pp. 3041–3048, 2014.
[81] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Introducing eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 1033–1041.
[82] S. Nag et al., "Multi-modal few-shot temporal action detection via vision-language meta-adaptation," 2022.
[83] 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 *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 7046–7055.
[84] S. Yan et al., "UnLoc: A unified framework for video localization tasks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 13623–13633.
[85] C. Ju et al., "Multi-modal prompting for low-shot temporal action localization," 2023, *arXiv:2303.11732*.
[86] T. Lin, X. Liu, X. Li, E. Ding, and S. Wen, "BMN: Boundary-matching network for temporal action proposal generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3889–3898.
[87] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.
[88] Y. A. Farha and J. Gall, "MS-TCN: Multi-stage temporal convolutional network for action segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3575–3584.



Asif Raza received the Ph.D. degree in controls science and control engineering from Shanghai Jiaotong University in 2022. He is currently a Post-Doctoral Researcher with the Advanced Data and Signal Processing Laboratory, Peking University. His contributions include several publications in famous journals and conferences and involvement in national projects and industry endeavors. His research interests include AI, machine learning, deep learning, pattern recognition, and multimedia content analysis. He actively contributes to the field by reviewing peer submissions and holding memberships in prestigious organizations, including IEEE Young Professionals, IEEE GRSS, AIAA, and SAAI.

Bang Yang received the B.E. degree from Sun Yat-sen University in 2018 and the M.S. degree from Peking University in 2021, where he is currently pursuing the Ph.D. degree. He is engaged in a joint training program with the Peng Cheng Laboratory. His research interests include multimodal learning and AI in healthcare.

Yuexian Zou (Senior Member, IEEE) received the B.Sc. degree from the University of Electronic Science and Technology in 1985 and the Ph.D. degree from The University of Hong Kong in 2001. She is currently a Full Professor with Peking University, where she is also the Director with the Advanced Data and Signal Processing Laboratory. She is also an Adjunct Professor with the Peng Cheng National Laboratory. Additionally, she is the Deputy Director of Shenzhen Association of Artificial Intelligence (SAAI). Since 2010, she has been dedicated to teaching and researching machine learning and its applications in video and audio analysis. She has led over 20 research projects, including those funded by NSFC and 863. With a prolific academic record, she has authored more than 260 papers in esteemed journals and flagship conferences and secured nine invention patents–two of which have been successfully transferred to a company. Her instructional roles encompass guiding graduate students through courses spanning machine learning, pattern recognition, digital signal processing, and array signal processing. Her primary research focus is on machine learning and scene understanding. In 2009, she received the Leading Figure for Science and Technology Award from Shenzhen Municipal Government.