ZEETAD: Adapting Pretrained Vision-Language Model for Zero-Shot End-to-End Temporal Action Detection

Thinh Phan*, Khoa Vo*, Duy Le†, Gianfranco Doretto‡, Donald Adjeroh‡, and Ngan Le*
*AICV Lab, University of Arkansas, Fayetteville, Arkansas, USA
†FPT Software AI Center, Vietnam
‡West Virginia University, Morgantown, West Virginia, USA
{thinhp, khoavoho, thile}@uark.edu, duylda1@fpt.com
{gianfranco.doretto, donald.adjeroh}@mail.wvu.edu

Abstract

Temporal action detection (TAD) involves the localization and classification of action instances within untrimmed videos. While standard TAD follows fully supervised learning with closed-set setting on large training data, recent zero-shot TAD methods showcase the promising open-set setting by leveraging large-scale contrastive visual-language (ViL) pretrained models. However, existing zero-shot TAD methods have limitations on how to properly construct the strong relationship between two interdependent tasks of localization and classification and adapt ViL model to video understanding. In this work, we present ZEETAD, featuring two modules: dual-localization and zero-shot proposal classification. The former is a Transformer-based module that detects action events while selectively collecting crucial semantic embeddings for later recognition. The latter one, CLIP-based module, generates semantic embeddings from text and frame inputs for each temporal unit. Additionally, we enhance discriminative capability on unseen classes by minimally updating the frozen CLIP encoder with lightweight adapters. Extensive experiments on THUMOS14 and ActivityNet-1.3 datasets demonstrate our approach's superior performance in zero-shot TAD and effective knowledge transfer from ViL models to unseen action categories. Code is available at https://github.com/UARK-AICV/ZEETAD.

1. Introduction

With the rapid growth of video content on the internet and social media, video understanding, which is about analyzing and interpreting action sequences, has gained a lot of interest. While video action recognition requires categorizing a standardized snippet with a single label, temporal action detection (TAD) aims to both localize and classify every action instances from long untrimmed videos. This task is challenging because existing supervised methods needs training with large amount of video data to attain decent performance. At the same time, obtaining multiple annotation pairs of temporal regions and corresponding action labels per video is laborious and expensive. These issues restrict current TAD works to closed-set learning setting, where the same set of categories apply to training and inference stage. Hence, there has been an increasing demand for expanding TAD methods to unseen classes with little additional annotation cost via few-shot or preferably zero-shot (ZS) learning strategies.

The recent achievements on vision-language (ViL) pretrained models with representatives such as CLIP [40], ALIGN [16], UniCL [52], Grounding DINO [30], have not only been beneficial to ZS image understanding tasks [5, 11, 12, 29, 36, 39, 56, 60] but promoted the generalizability of video analysis issues [19, 28, 41, 49–51]. Their success is ascribed to the rich semantics aligned with strong visual representation acquired from web-scale image-text pairs. The ZS transferability is accomplished by matching the similarity between query text embeddings and novel image features, or vice versa. Based on pretrained ViL models, there have been many literature that centered around ZS image recognition [12, 36, 56], open-vocabulary image segmentation [5, 11, 29, 60], interpretable AI in medical [36] and some related to ZS action recognition [19, 28, 49], video captioning [50, 51], objects tracking [41]. ZS TAD, which is the extension of ZS video action detection, has recently also received more attention and demonstrated promising performance in detecting actions within open-set settings.

As a foundational study, Efficient-Prompt [20] focuses on enhancing CLIP's text encoder to maximize similarity between visual proposals and textual embeddings. To achieve this, frame embeddings extracted from CLIP visual encoder are passed through a lightweight Transformer [43],

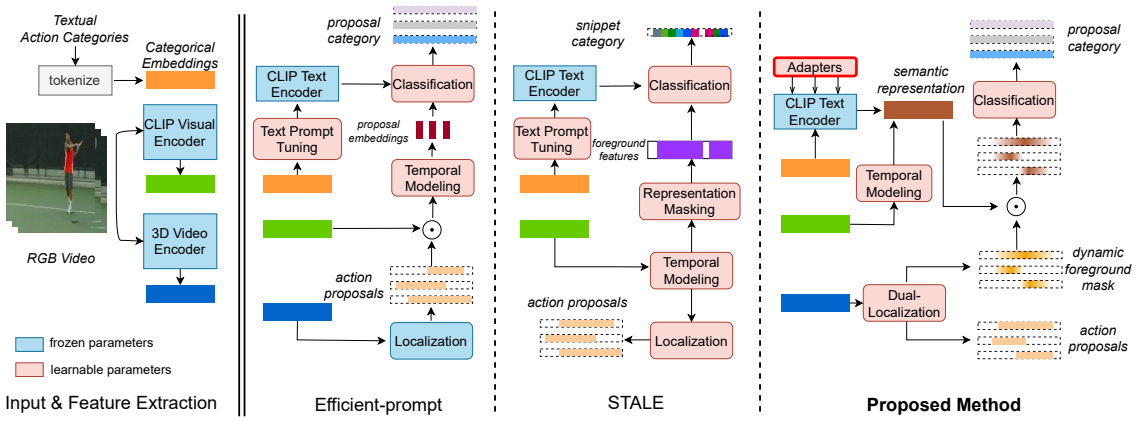

Figure 1. Network comparison between our ZEETAD and Efficient-Prompt [20], STALE [35]. Efficient-Prompt acquired the class-agnostic action proposals from pretrained action localization model and applied ZS video action recognition model on them. STALE used CLIP visual encoder for localization instead of 3D video backbone and predicted the action category and event duration on every snippet features. In our approach, we fuse CLIP text and visual features into temporal semantic representation, which we then correlate with a dynamic foreground mask to facilitate subsequent classification. Moreover, to enhance transferability, our technique employs Adapters in lieu of Text Prompt Tuning as in existing works.

capturing temporal information and encapsulating them into proposal embeddings. Nonetheless, due to the utilization of independently obtained proposal boundaries from a pretrained action localization model, the framework lacks integration between localization and classification [35] (limitation 1). Recognizing this limitation, STALE [35] designed a one-stage TAD model that reserved the foreground features through representation masking and jointly accommodated and classified every snippet embeddings. Additionally, the authors adopted text prompt tuning (TPT), a finetuning technique on CLIP, to determine optimal textual action descriptions. They also introduced a cross-modal adaptation to guide text features using contextual-level information. However, there is a dilemma of the localization input (limitation 2). According to CLIP's formation, novel classes are distinguished by comparing the similarity among CLIP-encoded visual and text embeddings. In STALE, the action localizer receives CLIP-encoded frame features as input. This impedes localization, as these features lack temporal relations and motion cues. Despite the subsequent temporal modeling module, the improvement remains marginal. Replacing this with conventional 3D video encoders for better localization results in an impaired action classifier, given the incongruity between visual features and text embeddings. Another issue is that STALE classified every video frames, instead of sequence of frames, and tied them to action proposals. Regarding the activity with several sub-actions, local temporal feature cannot be representative of an entire action (limitation 3).

Both STALE and Efficient-Prompt adopt ViL model leveraging learnable textual tokens from CLIP's text encoder to enhance ViL model adaptation. However, as noted in [53], TPT struggles to generalize to new classes and grapples with high intra-class variance in visual features (limitation 4). *To address the aforementioned issues, in this work, we design an effective end-to-end model architecture with the aid of CLIP pretrained model for ZS TAD called ZEETAD*. The network differences between the existing methods and ours are highlighted in Fig.1.

Our method focuses on solving two primary concerns: i) enhancing the ViL model for novel action detection; (ii) integrating the action localizer and classifier within the framework of TAD for an open-set scenario. Our network is a one-stage TAD model with learnable dual-localization module and ZS proposal classification module to address limitation 1. To rectify the limitation 2 identified in STALE, snippet features extracted from pretrained 3D convolutional neural network (CNN) are used for localization module while frame embeddings from CLIP's image encoder are allocated to the classification module, aligning with their respective objectives. Inspired by semantic image segmentation, the localization module employs video semantic embeddings generated by the classifier module to ascertain regions relevant to the action. Consequently, based on this segmented data, action proposals from seen or unseen classes are determined. Specifically, the frame embeddings, after being modeled with temporal relationship, are combined with their counterpart ones to generate a semantic representation that describe the action probabilities for each frames. Recognizing that not all frames contribute equally to action discrimination, we introduce the *dual-localization* module, facilitating the gathering of exclusively action-relevant embeddings. The classification of a class is attained by combining selected embeddings and identifying

the category with the highest matching score. Moreover, to achieve better transferability on video domain, we opt for an efficient finetuning technique known as AdaptFormer. As shown in [9], AdaptFormer merely updates the lightweight adapters injected inside the frozen CLIP Transformer sublayers, however, it surpasses the full-tuning solution by approximately 10% in video recognition task. The selection of the aforementioned components offers a simple pipeline as well as low computation cost but yields remarkable performance.

The main contributions are summarized as follows:

- We introduce a dual-localization mechanism designed not only to determine proposal boundaries but also to segment the semantic embeddings synthesized from CLIP.
- We integrate an efficient finetuning method, known as Adapters, to adapt a large-scale ViL model to the video domain.
- As a result, we employ the dual-localization mechanism and Adapters to propose a highly effective end-to-end model architecture for Zero-shot Temporal Action Detection (ZEETAD), encompassing two modules: temporal action dual-localization and Zero-Shot (ZS) proposal classification.
- We conduct experiments on the THUMOS14 and ActivityNet-1.3 datasets. Our ZEETAD model, featuring dual-localization and conceptual-based classification, significantly outperforms other state-of-the-art (SOTA) methods. We also present comprehensive ablation studies to demonstrate the effectiveness of each individual component.

2. Related Works

2.1. Pretrained Vision-Language Models

Vision-Language models have rapidly evolved in recent years with the intention of improving vision models' generalizability upon unseen object classes. The key idea is to capitalize on large scale of pairs of images and natural language descriptions and then train a network to align the image representation with text embeddings through noise contrastive learning. While early approaches explored the semantic representation through word embeddings of class name [1] or attributes [22], recent works with representatives as CLIP [40] and ALIGN [16] augmented the training procedure with millions of image-text pairs as well as the backbone with modern Transformer [43]. Their rich vision-language correspondence knowledge serves as effective pretrained model for many few-shot to ZS tasks such as image captioning [33, 50, 51], image retrieval [32], semantic segmentation [11, 42], medical imaging [37, 39], object tracking [24, 41] . Along with that, adaptation methods [17, 18, 23] for large-scale ViL models have been becoming favored research which is about minimally finetuning these computation-heavy models but still boost the generalization capability on new tasks. In this work, we utilize CLIP for ZS action classification branch in TAD and further improve its performance on unseen video categories by incorporating adapters [9] to the text encoder backbone.

2.2. Temporal Action Detection

Temporal action detection is one of the key task in video understanding topic. Current methods could be roughly categorized into two-stage methods and one-stage methods. The former ones initially generate action proposals or foreground instances and then assign them with action categories. Commonly, more effort was put on the first stage and action labels were obtained from external classification scores. DCAN [8] followed the anchor-based localization, adjusting the pre-defined anchors based on boundary level and proposal level scores. BSN [27] and BMN [26] fall into action-guided localization direction, evaluating candidate proposals with probability of being a potential action. ABN [47] and [48] formulates TAD as a interaction between environment and agent. AEI [44], [46], AOE-Net [45] are also two stages TAD but they are are designed with explainable capability. Single-stage method [25, 54], also known as anchor-free localization method, simultaneously classifies every temporal units and regresses their boundaries. Our proposed model follows the single-stage method but it concurrently generates the action proposals and categorizes them. A Transformer encoder-decoder as the backbone is responsible for action boundary detection and proposal ranking through actionness score.

2.3. Zero-shot Temporal Action Detection

Zero-shot learning considers the model awareness of novel classes absent during training. Zero-shot learning in TAD is challenging because it deals with the joint localization and classification of multiple unseen instances. ZSTAD [55] adopted the R-C3D framework and optimized activity label mapping by considering common semantics between seen and unseen activities. However, ZSTAD allowed the unseen semantic embeddings to support the training stage via super-class classification loss, which is impractical in real-world scenarios. To address this issue, TranZAD [34] used semantic information of only seen classes at training phase and also proposed a network that learns to group action visual features and their corresponding class-specific semantic embedding. The semantic embeddings in these two methods are acquired from Word2Vec [13] or GLoVe [38]. More recent approaches, such as Efficient-Prompt [20] and STALE [35], harness large-scale pretrained ViL models, which inherently possess visual-textual alignment capabilities. he experiments clearly illustrate the effectiveness of incorporating ViL capabilities to address agnostic-

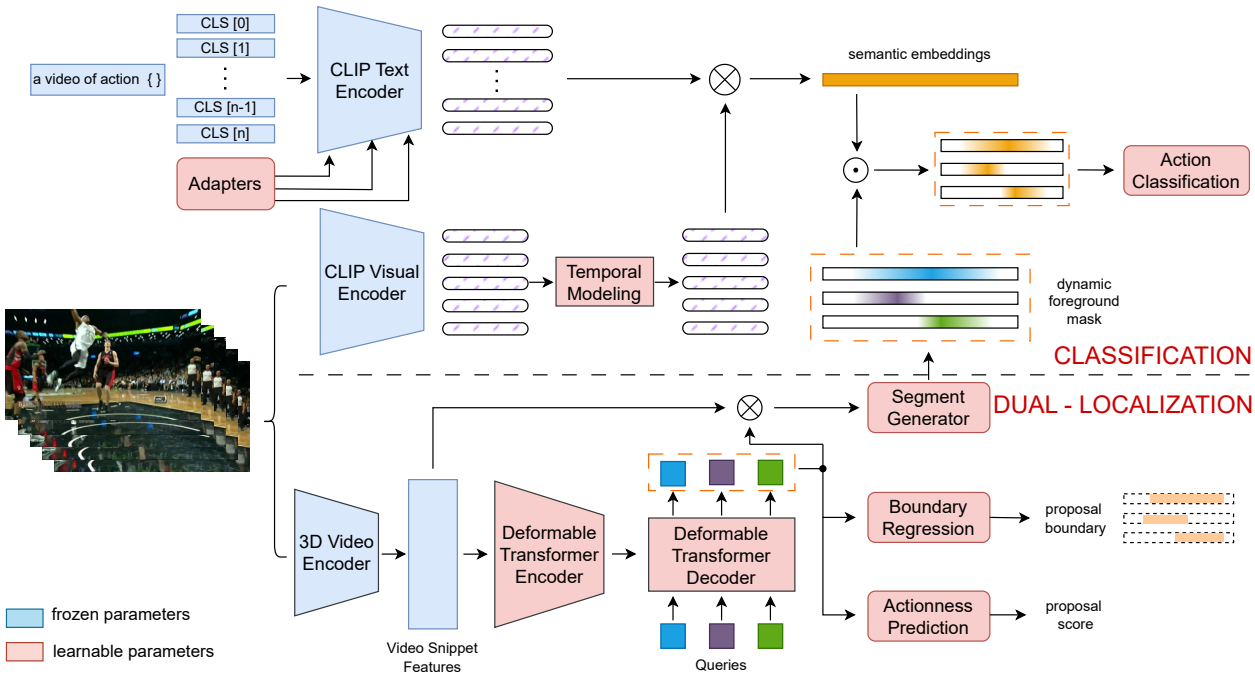

Figure 2. Overall architecture of ZEETAD. The classification module associates frame embeddings and textual categories embeddings from CLIP encoders to generate the temporal semantic embeddings. The dual-localization receives the 3D encoded video snippet features and passes them through Deformable Transformer for class-agnostic action proposal generation. Segment generator is proposed in the dual-localization module to dynamically mask the appropriate semantic embeddings for each event. Finally, the selected proposal segment of semantic representation are aggregated and classified. Adapters as a finetuning technique are attached to CLIP text encoder to increase the generalizability of zero-shot classification.

action scenarios in ZS TAD. Both Efficient-Prompt and STALE have outperformed all existing ZS TAD methods. In our approach, we harness the power of the CLIP pretrained model within the action classifier module to accurately identify unseen classes. Specifically, we adapt CLIP to produce temporal semantic embeddings, which are subsequently clustered for each action candidate and ultimately result in the assignment of the final action class.

3. Methodology

As shown in Fig.2, our framework is the unification of two sub-tasks, action proposal localization and action classification. In the upper part of Fig.2, CLIP pretrained model encodes the RGB frames and action category contents and the frame and text embeddings are subsequently multiplied to create semantic embeddings for all frames. A deformable Transformer encoder-decoder model receives the 3D encoded video features, predicting the action intervals and confidence scores. Unlike supervised TAD that assigns labels by distinguishing video features, ZS TAD indirectly classifies action by assessing the matching score of visual and textual embeddings. Hence, segment generator is designed to generate foreground mask of the semantic embeddings pertinent to the corresponding action boundary. Finally, the assembled semantic embeddings are fed to an classifier to yield the activity label.

3.1. Problem Definition

Our work focuses on the problem of ZS TAD. Given a training set of untrimmed videos $\mathcal{D}_{train} = \{\mathcal{V}_i\}_{i=1}^{n}$, we have an input set of RGB frames $X_i = \{f_t\}_{t=1}^{T}$, where $\mathcal{T}$ is the number of video snippets from input sequence. As a TAD task, annotation is demonstrated as $Y_i = \{s_k, e_k, c_{sk}\}_{k=1}^{K_i}$ where K_i is the amount of action events, s and e are respectively the start and end of each event and c_s is activity category. The testing set $\mathcal{D}_{test}$ shares the same data structure of $\mathcal{D}_{train}$; regarding the ZS scenarios, the activity classes in $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$ are non-overlapping $\mathcal{C}_{train} \cap \mathcal{C}_{test} = \emptyset$. Our proposed method aims to predict all video segments in an open-set scenario guided by word embeddings of $\mathcal{C}_{test}$.

3.2. Semantic Representation

We adopt CLIP [40] pretrained Vision-Language model to set up the ZS action recognition module. To fit CLIP into ZS TAD, our idea is to fuse each frame embedding with entire set of textual category embeddings, creating the se-

mantic embedding for every frames. The semantic representation brings the probabilities of every classes at each temporal unit. Multiple action proposals in a video input are assigned with labels via segmenting and aggregating related semantic embeddings. Specifically, video frame embeddings $\mathcal{F}^{rgb} \in \mathbb{R}^{T\times d}$ are acquired by passing middle RGB frames of video snippets through CLIP visual encoder $\Phi_{CLIP-v}(.)$. To inject the temporal context into unattached frame embeddings, we apply a temporal Transformer [43] on $\mathcal{F}^{rgb}$. The temporal modeling module includes layers of Residual Attention Blocks constructed from Multi-head self-attention, Layer Norm, QuickGELU and MLP:

$$\mathcal{F}^{rgb-t} = \Phi_{TEMP}(\{\mathcal{F}^{rgb}(1), ..., \mathcal{F}^{rgb}(T)\}) \quad (1)$$

In terms of text embeddings $\mathcal{F}^{text} \in \mathbb{R}^{C\times d}$, a set of action categories are prepended with prompt template "a video of action" and then passed through the CLIP language encoder $\Phi_{CLIP-t}(.)$. For ZS video action recognition task, a mean pooling is commonly followed by $\mathcal{F}^{rgb-t}$ to get the aggregated video embedding, which is then used to find the highest matching score with $\mathcal{F}^{text}$. This cannot be naively applied in TAD because there are multiple events in individual video input. Therefore, we generate temporal semantic representation $\mathcal{S} \in \mathbb{R}^{T\times C}$ made up of visual and text features:

$$\mathcal{S} = \mathcal{F}^{rgb-t} \cdot (\mathcal{F}^{text})^T \quad (2)$$

To recognize the category of an action proposal, we find the corresponding region on semantic embedddings and deduce the activity label from them. The embedding selection and aggregation on semantic representation are supported by dual-localization module. We analyze it in the next section.

3.3. Dual Localization

The backbone for dual-localization framework is motivated by Deformable DETR [59], an encoder-decoder object detection model based on the Transformer. While the predecessor DETR [6] takes long time to converge and its attention modules fails to generate high-resolution image feature maps, Deformable DETR requires remarkably less time to train and more importantly, achieves better performance at detecting small objects. This characteristic brings advantage in coping with the vagueness in action boundaries, temporal redundancy in long videos and short action detection. Initially, RGB frames are sampled into length $\mathcal{T}$ and pretrained 3D video encoder (e.g., I3D [7] or TSP [2]) extracts the video snippet features $\mathcal{F}^{3D} \in \mathbb{R}^{T\times l}$. A 1D convolutional is followed by to match video feature dimension with CLIP feature dimension. The encoder with $\mathcal{L}_E$ Transformer layers models the relations among video features via deformable attention modules and returns feature sequence $\mathcal{F}^{enc} \in \mathbb{R}^{T\times d}$ carrying temporal context. The decoding network consists of $\mathcal{L}_D$ deformable cross attention layers and receives the encoded feature sequence $\mathcal{F}^{enc}$ (served as key) and $\mathcal{N}_q$ learnable embedding queries $q \in \mathbb{R}^d$. For each query q, the decoder outputs an embedding $\mathcal{F}^{dec} \in \mathbb{R}^d$, which are later utilized by three prediction heads.

Boundary regression head predicts the normalized action middle point and duration of an activity. The two variables are computed by applying a feed-forward network (FFN) then a sigmoid function into the output embedding:

$$\hat{Y}_b = \{m, d\} = \text{sigmoid}(FFN(\mathcal{F}^{dec})) \quad (3)$$

Actionness prediction head: Conventionally, image object detection model uses classification score to rank the duplicate queries and pick the best bounding boxes. Since 3D encoded snippet feature does not comprehend strong discriminative power and adjacent frames of an action event share high similarity with the main ones, classification score is not a reliable ranking indicator for localization quality. Hence, we adopt the actionness score from TadTR [31] to support proposal selection. In detail, the ROIAlign [14] along with predicted boundaries B is applied on the encoder output features $\mathcal{F}^{enc}$ to extract the small feature map within the action interval:

$$f^{RoI} = RoI(\mathcal{F}^{enc}, B) \quad (4)$$

Subsequently, the aligned feature f^{RoI} is fed to an FFN with sigmoid activation to regress the actionness score. Actionness score is effective because it guides the model to be more sensitive to the local features [31].

Segment generator head is the continuation of section 3.2. To classify an event, we should attentively accumulate relevant semantic embeddings based on the action boundary. Using the start and end timestamps straight from the boundary regression head as binary segmentation is not optimal. For instance, the action "high jump" and "long jump" both comprise of striding and jumping. Equal attentions to these sub-actions could lead to classification uncertainty. Based on this observation and motivated by the mask formulation of MaskFormer [10], we propose the proposal segment generator module. The semantic mask $\mathcal{M} \in \mathbb{R}^T$ is created by multiplying the video snippet feature $\mathcal{F}^{3D}$ with the prediction embedding $\mathcal{F}^{dec}$ via dot product:

$$\mathcal{M} = \text{sigmoid}(\mathcal{F}^{3D} \cdot \mathcal{F}^{dec}) \quad (5)$$

The semantic mask is kept flexible ($\mathcal{M} \in [0, 1]$) instead of applying thresholding to binarize them. The region mask helps localize the events on the semantic embeddings and put more concentration on semantic units that provide decent classification clues. The final class of a action proposal is determined by follow:

$$\hat{Y}_c = \arg\max \frac{1}{T} \sum_{t=1}^{T} (\mathcal{M}_{(t)} \odot S_{(t)}) \quad (6)$$

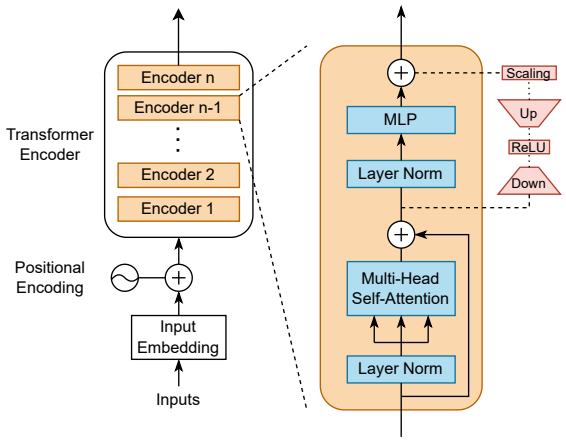

Figure 3. Adapter-based finetuning mechanism in our ZEETAD.

3.4. Vision-Language Model Finetuning

Fully finetuning on large-scale datasets has been well-known to instill the knowledge to downstream tasks. Considering the massive scale of the Transformer model, full finetuning takes tremendous amount of time and computation. Furthermore, full finetuning tends to overfit the downstream task and loses the generalization acquired in the large-scale pretraining stage, which is extremely harmful to ZS tasks. To address the above challenges, existing works [17, 21, 58] focused on parameter-efficient finetuning technique that adjusts minimum amount of learnable parameters and keep the entire backbone frozen. In this paper, we leverage AdaptFormer [9] to implement our Adapter-based finetune the pretrained CLIP text encoder. As displayed in Fig.3, our Adapter-based finetune replaces the original MLP sub-layers in Transformer blocks with AdaptMLP, which is a bottleneck module parallel to the original one. During finetuning, only the added parameters are optimized and the entire encoder is frozen. The additional modules only occupied for 1.46% of model parameters.

3.5. Training and Inference

Training: Following [59], we include boundary refinement on each decoding layer and set dropout rate as zero within the transformer. Bipartite matching is used to find the lowest matching cost among targets to their corresponding predictions (one-to-one mapping) and the loss is computed on the matching pairs. While the matching cost regards the classification probabilities and the resemblance between target and output bounding boxes, the total training loss is defined as follow:

$$\mathcal{L} = \lambda_1.\mathcal{L}_{cls} + \lambda_2.\mathcal{L}_{bbox} + \lambda_3.\mathcal{L}_{actionness} \quad (7)$$

To maximize the alignment between the frame sequence embeddings and corresponding text embeddings, cross-entropy loss with temperature parameter τ is used:

$$\mathcal{L}_{cls} = -\sum_i \log \frac{\exp(f_i \cdot t_j/\tau)}{\sum_j \exp(f_i \cdot t_j/\tau)} \quad (8)$$

$\mathcal{L}_{bbox}$ regresses the midpoint and the duration of the action proposal via L1 loss and the generalized IoU loss:

$$\mathcal{L}_{bbox} = \alpha_1.||b_{gt} - b_i||_1 + \alpha_2.gIoU(b_{gt}, b_i) \quad (9)$$

Actionness score is supervised by the offset of the IoU between the predicted bounding box and its closest groundtruth segment. The higher overlap rate is equivalent to higher actionness score a, and vice versa. $\mathcal{L}_{actionness}$ is a L1 loss that minimizes the ranking loss of all action proposals:

$$\mathcal{L}_{actionness} = \sum ||a_i - IoU(b_{gt}, b_i)||_1 \quad (10)$$

Inference. For each video input, we have $\mathcal{N}_q$ proposal boundaries, labels, classification scores and actionness scores. The proposal labels are obtained by applying $argmax$ on the classification head's logits. The final proposal confidence score is obtained by multiplying classification scores and actionness scores. We gather the predictions in top-k order according to the confidence score. At last, Soft-NMS [3] is applied to remove duplicate and low-quality predictions.

4. Experimental Results

4.1. Dataset and Metrics

We conduct experiments on THUMOS14 [15] and ActivityNet-1.3 [4] datasets. THUMOS14 collects videos from 20 sports action classes, containing 200 and 213 videos for training and testing, respectively. ActivityNet-1.3 contains 200 classes of daily activities with the total of 19994 videos. We follow previous literature [20] to adopt two validation splits (75% seen training classes - 25% novel testing classes and 50% seen training classes - 50% novel testing classes) for zero-shot scenarios. To ensure the legitimate generalization, the final results are averaged on 10 random splits. We compare to only existing methods that comply with this evaluation scheme. Following existing TAD and ZS TAD methods, the mean average precision (mAP) at different IOU thresholds is reported as main evaluation metrics.

4.2. Implementation Details

Similar to other TAD methods, available 3D video features are used as the input of the localization module. The selection of video features is centered around the best performance on action boundary localization. On THUMOS14, we utilize the two stream I3D [7] video encoder

Table 1. Performance comparison between our ZEETAD with SOTA ZS TAD methods on ActivityNet-1.3 and THUMOS14. mAPs at different IOU thresholds of 0.3, 0.4, 0.5, 0.6, 0.7 and averaging (AVG) are reported.

Train-Test split	Method	THUMOS14						ActivityNet-1.3			
		0.3	0.4	0.5	0.6	0.7	**AVG**	0.5	0.75	0.95	**AVG**
75%-25%	B-II [35]	28.5	20.3	17.1	10.5	6.9	16.6	32.6	18.5	5.8	19.6
	B-I [35]	33.0	25.5	18.3	11.6	5.7	18.8	35.6	20.4	2.1	20.2
	Eff-Prompt [20]	39.7	31.6	23.0	14.9	7.5	23.3	37.6	22.9	3.8	23.1
	STALE [35]	40.5	32.3	23.5	15.3	7.6	23.8	38.2	25.2	**6.0**	24.9
	ZEETAD	**61.4**	**53.9**	**44.7**	**34.5**	**20.5**	**43.2**	**51.0**	**33.4**	5.9	**32.5**
50%-50%	B-II [35]	21.0	16.4	11.2	6.3	3.2	11.6	25.3	13.0	3.7	12.9
	B-I [35]	27.2	21.3	15.3	9.7	4.8	15.7	28.0	16.4	1.2	16.0
	Eff-Prompt [20]	37.2	29.6	21.6	14.0	7.2	21.9	32.0	19.3	2.9	19.6
	STALE [35]	38.3	30.7	21.2	13.8	7.0	22.2	32.1	20.7	**5.9**	20.5
	ZEETAD	**45.2**	**38.8**	**30.8**	**22.5**	**13.7**	**30.2**	**39.2**	**25.7**	3.1	**24.9**

while on ActivityNet-1.3,TSP features [2] is adopted. In terms of ActivityNet-1.3, we resize the video features to standard length $\mathcal{T}$ of 100 by linear interpolation. For THUMOS14, we follow previous papers [25] to divide the long video features into windows of length 128 with overlap rate of 0.75 for both training and testing. Regarding the ZS classification module, the model version of CLIP model is ViT-B/16. The input frames are picked from the middle frame of a snippet and thus, the length of the CLIP video features are equal to length $\mathcal{T}$ of the 3D video encoded features. We set $\mathcal{L}_E = 4$, $\mathcal{L}_D = 4$, and $\mathcal{N}_q = 10$ on ActivityNet and $\mathcal{N}_q = 40$ on THUMOS. All parameters of the network including add-in adapters are learnable except the CLIP vision encoder and text encoder. Adam optimizer with learning rate of 10^{-4} and batch size of 16 are set as the model hyperparameters. ZEETAD is trained for 30 epochs on both datasets. The threshold for Soft-NMS is 0.3.

4.3. Main Result

Table 1 displays results of ZEETAD and other existing methods on open-set scenarios of TAD. We also include two baseline experiments (B-I and B-II) from STALE [35]. Overall, we achieve the state-of-the-art results at nearly all IoU thresholds on two datasets as well as two ZS settings. Specifically, on THUMOS14, in terms of average mAP, we surpass the second-best method, STALE, by 19.4% on 75%-25% data split and by 8.0% on 50%-50% data split. The large gain margins are also perceived on ActivityNet-1.3. ZEETAD achieves average mAPs of 32.5% and 24.9% on the mentioned dataset. The reported performance validates the effectiveness of our approach towards the problem of temporal action detection generalization.

4.4. Ablation Study

We implement experiments on THUMOS14 dataset and evaluate the options of constituent components and model structures in this section.

Vision-Language model finetuning: Different methods of CLIP model finetuning are investigated in Table 2. The average mAPs for not applying finetuning techniques are 37.2% and 27.1% on 75%-25% and 50%-50%, respectively. We implement Text Prompt Tuning based on CoOp [58] and use the appended learnable context length of 16. We observe the performance drop in both dataset splits compared to the one not using TPT. The performance gap between TPT and baseline in 50% setting is bigger than the 75% setting, which could be an overfitting problem. According to [57], the learned context is prone to overfitting on the seen classes and cannot be generalizable to unseen classes. By integrating adapter [9], we improve the average mAP by 6% and 3.1% on two data settings.

Encoded feature utilization: To demonstrate the input problem of STALE, our network takes in uni-visual feature encoder backbone for both localizer and classifier. Experimented on 75%-25% setting, Table 3 describes the placement of two types of visual features (CLIP and I3D) for localization and classification modules. We can easily observe in Table 3 that using encoded features with no regard to their dedicated purpose results in big drop in performance. CLIP visual encoder have no capability to convey the temporal context and using online temporal learning model is trivial compared to 3D pretrained video encoder like I3D. This could be derived from the fact that using I3D on both branches yields better results.

Component effectiveness: We assess the importance of supporting components proposed in our method in Table 4. Although the model is still able to elaborate the ZS TAD task if these components are omitted, their improvement

Table 2. The effectiveness of different finetuning methods on two data split settings on THUMOS14 dataset. mAPs at different IOU thresholds of 0.3, 0.4, 0.5, 0.6, 0.7 and averaging (AVG) are reported.

	75%-25%						**50%-50%**					
	0.3	0.4	0.5	0.6	0.7	**AVG**	0.3	0.4	0.5	0.6	0.7	**AVG**
w/o Finetuning	52.3	46.6	38.9	29.2	19.1	37.2	39.2	34.2	28.1	20.6	13.2	27.1
Text Prompt Tuning	53.5	47.0	37.5	27.0	16.9	36.4	37.1	31.5	25.12	18.4	11.6	24.7
Adapter	**61.4**	**53.9**	**44.7**	**34.5**	**20.5**	**43.2**	**45.2**	**38.8**	**30.8**	**22.5**	**13.7**	**30.2**

Table 3. The effect of encoded feature utilization for localization and classification modules on 75%-25% setting on THUMOS14 dataset. mAPs at different IOU thresholds of 0.3, 0.5, 0.7 and averaging (AVG) are reported.

Localizer	Classifier	0.3	0.5	0.7	**AVG**
CLIP	CLIP	47.3	29.7	11.5	29.7
I3D	I3D	43.4	33.0	16.9	31.5
I3D	CLIP	**61.4**	**44.7**	**20.5**	**43.2**

cannot be denied. Action proposal confidence score is fused from classification score and actionness score. Without actionness prediction, the average mAP is reduced by 0.9%. Temporal modeling module helps inject temporal context into CLIP encoded frame features, creating a temporal-coherent semantic representation. Directly generating semantic embeddings from frame embeddings decreases the performance from 43.2% to 41.7%. Segment generator is in charge of providing the soft masks of semantic embeddings corresponding to action proposals. Should it be excluded, we need to segment the semantic representation via action boundaries, which is analogous to hard (binary) mask prediction. The average mAP is declined to 33.4% if segment generator is not employed, making it the key component in our network.

Model structure: To ascertain the advantage of one-stage ZS TAD over two-stage counterpart, ZEETAD is compared to its variant in Table 5. To convert ZEETAD to two-stage framework, the localization branch is trained separately and the predicted action proposals are given to gather the semantic embeddings of the classification branch. This setup resembles Efficient-Prompt [20]. We obtain the final

Table 4. Ablation study of proposed components on 75%-25% setting on THUMOS14 dataset. mAPs at different IOU thresholds of 0.3, 0.5, 0.7 and averaging (AVG) are reported. Actionness Prediction, Temporal Modeling and Segment Generator are denoted as AP, TM and SG, respectively.

	0.3	0.5	0.7	**AVG**
ZEETAD	61.4	44.7	20.5	43.2
- AP	62.6 (+1.2)	43.9 (-0.8)	19.7 (-0.8)	42.3 (**-0.9**)
- TM	59.6 (-1.8)	43.4 (-1.3)	20.3 (-0.2)	41.7 (**-1.5**)
- SG	47.6 (-13.8)	34.5 (-10.2)	17.0 (-3.5)	33.4 (**-9.8**)

Table 5. Analysis of model structure on 75%-25% setting on THUMOS14 dataset. mAPs at different IOU thresholds of 0.3, 0.4, 0.5, 0.6, 0.7 and averaging (AVG) are reported.

	0.3	0.4	0.5	0.6	0.7	**AVG**
Two-stage	40.3	36.1	28.9	21.4	14.2	28.2
One-stage	**61.4**	**53.9**	**44.7**	**34.5**	**20.5**	**43.2**

average result of 28.2%, which is much lower than the result of one-stage setup. Compared to the result of [20] on the same setting, we still achieve better performance. This could be attributed to the stronger class-agnostic proposal detection backbone (Deformable DETR [59]) and also the more efficient deep prompt tuning technique.

5. Conclusion

In this work, we propose the Transformer-based end-to-end framework for Zero-shot Temporal Action Detection model, titled ZEETAD. The model is designed as a one-stage TAD method that unite the localization and classification tasks. Large-scale Vision-Language pretrained model plays an important role to empower the zero-shot classification capability. ZEETAD revises the CLIP zero-shot mechanism in image recognition and transforms it into video action proposal classification. Concurrently, the localization branch enhances the action recognition rate by segmenting the appropriate semantic embeddings with strong discriminative features. In addition, an efficient finetuning method known as adapter is integrated to CLIP text encoder. Adapter helps augment the text embeddings so that they could increase the matching score with their corresponding video embeddings. The experimental results on THUMOS14 and ActivityNet-1.3 verified the effectiveness of our approach and ZEETAD significantly outperforms existing methods.

Acknowledgment

This material is based upon work supported by the National Science Foundation (NSF) under Award No OIA-1946391 RII Track-1, NSF 1920920 RII Track 2 FEC, NSF 2223793 EFRI BRAID, NSF 2119691 AI SUSTEIN, NSF 2236302.

References

[1] Ziad Al-Halah, Makarand Tapaswi, and Rainer Stiefelhagen. Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning. In *CVPR*, pages 5975–5984, 2016. 3

[2] Humam Alwassel, Silvio Giancola, and Bernard Ghanem. Tsp: Temporally-sensitive pretraining of video encoders for localization tasks. In *ICCV*, pages 3173–3183, 2021. 5, 7

[3] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *ICCV*, pages 5561–5569, 2017. 6

[4] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015. 6

[5] Zhaowei Cai, Gukyeong Kwon, Avinash Ravichandran, Erhan Bas, Zhuowen Tu, Rahul Bhotika, and Stefano Soatto. X-detr: A versatile architecture for instance-wise vision-language tasks. In *ECCV*, pages 290–308. Springer, 2022. 1

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020. 5

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 5, 6

[8] Guo Chen, Yin-Dong Zheng, Limin Wang, and Tong Lu. Dcan: improving temporal action detection via dual context aggregation. In *AAAI*, volume 36, pages 248–257, 2022. 3

[9] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *NeurIPS*, 35:16664–16678, 2022. 3, 6, 7

[10] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *NeurIPS*, 34:17864–17875, 2021. 5

[11] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *CVPR*, pages 11583–11592, 2022. 1, 3

[12] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. 1

[13] Yoav Goldberg and Omer Levy. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014. 3

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 5

[15] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017. 6

[16] Chao Jia, Yinfei Yang, et al. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916. PMLR, 2021. 1, 3

[17] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, pages 709–727. Springer, 2022. 3, 6

[18] Shibo Jie and Zhi-Hong Deng. Convolutional bypasses are better vision transformer adapters. *arXiv preprint arXiv:2207.07039*, 2022. 3

[19] Hyekang Kevin Joo, Khoa Vo, Kashu Yamazaki, and Ngan Le. Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 3230–3234. IEEE, 2023. 1

[20] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *ECCV*, pages 105–124. Springer, 2022. 1, 2, 3, 6, 7, 8

[21] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, pages 19113–19122, 2023. 6

[22] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE TPAMI*, 36(3):453–465, 2013. 3

[23] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 3

[24] Siyuan Li, Tobias Fischer, Lei Ke, Henghui Ding, Martin Danelljan, and Fisher Yu. Ovtrack: Open-vocabulary multiple object tracking. In *CVPR*, pages 5567–5577, 2023. 3

[25] Chuming Lin, Chengming Xu, Donghao Luo, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. Learning salient boundary feature for anchor-free temporal action localization. In *CVPR*, pages 3320–3329, 2021. 3, 7

[26] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *ICCV*, pages 3889–3898, 2019. 3

[27] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *ECCV*, pages 3–19, 2018. 3

[28] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. In *ECCV*, pages 388–404. Springer, 2022. 1

[29] Quande Liu, Youpeng Wen, and other. Open-world semantic segmentation via contrasting and clustering vision-language embedding. In *ECCV*, pages 275–292. Springer, 2022. 1

[30] Shilong Liu, Zhaoyang Zeng, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 1

[31] Xiaolong Liu, Qimeng Wang, Yao Hu, Xu Tang, Shiwei Zhang, Song Bai, and Xiang Bai. End-to-end temporal action detection with transformer. *IEEE Transactions on Image Processing*, 31:5427–5441, 2022. 5

[32] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *ICCV*, pages 2125–2134, 2021. 3

[33] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 3

[34] Sayak Nag, Orpaz Goldstein, and Amit K Roy-Chowdhury. Semantics guided contrastive learning of transformers for zero-shot temporal activity detection. In *WACV*, pages 6243–6253, 2023. 3

[35] Sauradip Nag, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Zero-shot temporal action detection via vision-language prompting. In *ECCV*, pages 681–697. Springer, 2022. 2, 3, 7

[36] Toan Nguyen, Minh Nhat Vu, Baoru Huang, Tuan Van Vo, Vy Truong, Ngan Le, Thieu Vo, Bac Le, and Anh Nguyen. Language-conditioned affordance-pose detection in 3d point clouds. *arXiv preprint arXiv:2309.10911*, 2023. 1

[37] Tien-Phat Nguyen, Trong-Thang Pham, Tri Nguyen, Hieu Le, Dung Nguyen, Hau Lam, Phong Nguyen, Jennifer Fowler, Minh-Triet Tran, and Ngan Le. Embryosformer: Deformable transformer and collaborative encoding-decoding for embryos stage development classification. In *CVPR*, pages 1981–1990, 2023. 3

[38] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014. 3

[39] Trong Thang Pham, Jacob Brecheisen, Anh Nguyen, Hien Nguyen, and Ngan Le. Decoding radiologists intense focus for accurate cxr diagnoses: A controllable and interpretable ai system. *arXiv preprint arXiv:2309.13550*, 2023. 1, 3

[40] Alec Radford, Jong Wook Kim, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1, 3, 4

[41] Kim Hoang Tran, Tien-Phat Nguyen, Anh Duy Le Dinh, Pha Nguyen, Thinh Phan, Khoa Luu, Donald Adjeroh, and Ngan Hoang Le. Z-gmot: Zero-shot generic multiple object tracking. *arXiv preprint arXiv:2305.17648*, 2023. 1, 3

[42] Minh Tran, Khoa Vo, Kashu Yamazaki, Arthur Fernandes, Michael Kidd, and Ngan Le. Aisformer: Amodal instance segmentation with transformer. *arXiv preprint arXiv:2210.06323*, 2022. 3

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 3, 5

[44] Khoa Vo, Hyekang Joo, Kashu Yamazaki, Sang Truong, Kris Kitani, Minh-Triet Tran, and Ngan Le. Aei: Actors-environment interaction with adaptive attention for temporal action proposals generation. *arXiv preprint arXiv:2110.11474*, 2021. 3

[45] Khoa Vo, Sang Truong, Kashu Yamazaki, Bhiksha Raj, Minh-Triet Tran, and Ngan Le. Aoe-net: Entities interactions modeling with adaptive attention mechanism for temporal action proposals generation. *IJCV*, 131(1):302–323, 2023. 3

[46] Khoa Vo, Kashu Yamazaki, Phong X Nguyen, Phat Nguyen, Khoa Luu, and Ngan Le. Contextual explainable video representation: Human perception-based understanding. In *2022 56th Asilomar Conference on Signals, Systems, and Computers*, pages 1326–1333. IEEE, 2022. 3

[47] Khoa Vo, Kashu Yamazaki, Sang Truong, Minh-Triet Tran, Akihiro Sugimoto, and Ngan Le. Abn: Agent-aware boundary networks for temporal action proposal generation. *IEEE Access*, 9:126431–126445, 2021. 3

[48] Viet-Khoa Vo-Ho, Ngan Le, Kashu Kamazaki, Akihiro Sugimoto, and Minh-Triet Tran. Agent-environment network for temporal action proposal generation. In *ICASSP*, pages 2160–2164. IEEE, 2021. 3

[49] Wenhao Wu, Xiaohan Wang, Haipeng Luo, Jingdong Wang, Yi Yang, and Wanli Ouyang. Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models. In *CVPR*, pages 6620–6630, 2023. 1

[50] Kashu Yamazaki, Sang Truong, Khoa Vo, Michael Kidd, Chase Rainwater, Khoa Luu, and Ngan Le. Vlcap: Vision-language with contrastive learning for coherent video paragraph captioning. In *ICIP*, pages 3656–3661, 2022. 1, 3

[51] Kashu Yamazaki, Khoa Vo, Quang Sang Truong, Bhiksha Raj, and Ngan Le. Vltint: visual-linguistic transformer-in-transformer for coherent video paragraph captioning. In *AAAI*, volume 37, pages 3081–3090, 2023. 1, 3

[52] Jianwei Yang, Chunyuan Li, et al. Unified contrastive learning in image-text-label space. In *CVPR*, pages 19163–19173, 2022. 1

[53] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Unified vision and language prompt learning. *arXiv preprint arXiv:2210.07225*, 2022. 2

[54] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *ECCV*, pages 492–510. Springer, 2022. 3

[55] Lingling Zhang, Xiaojun Chang, Jun Liu, Minnan Luo, Sen Wang, Zongyuan Ge, and Alexander Hauptmann. Zstad: Zero-shot temporal activity detection. In *CVPR*, pages 879–888, 2020. 3

[56] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. 1

[57] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022. 7

[58] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 6, 7

[59] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 5, 6, 8

[60] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *CVPR*, pages 15116–15127, 2023. 1