Multi-Scale Latent Temporal Compression Attention and Video-Text Cross Attention Fusion for Video Action Detection

Yicheng Qiu[0009-0006-8841-8648] and Keiji Yanai(✉)[0000-0002-0431-183X]

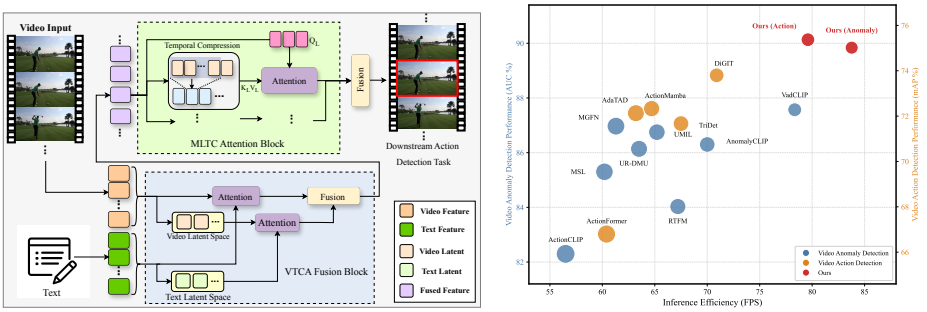
The University of Electro-Communications, Tokyo, Japan
qiu-y@mm.inf.uec.ac.jp, yanai@cs.uec.ac.jp

Abstract. In video action detection, modeling the temporal dependencies in continuous video streams is crucial. Previous methods based on deep neural networks can perform frame-level video recognition tasks. However, these methods struggle to model dynamic temporal patterns, facing challenges such as redundant frame information, difficulty balancing global and local features, and insufficient fusion of visual and textual features. To address these issues, we propose a video temporal attention module based on the Multi-Scale Latent Temporal Compression Attention, which aggregates continuous temporal sequences into feature representations across diverse temporal scales while efficiently reducing redundancy. We also introduce a Video-Text Cross Attention Fusion module that leverages cross-attention at both latent and standard feature levels, achieving fine-grained fusion between video and textual features to enhance semantic alignment and feature interaction. Experimental results across multiple video recognition benchmarks demonstrate that our proposed methods achieve superior performance compared to existing methods, validating the effectiveness and generalization capability.

Keywords: Video Action Detection · Temporal Attention · Video-Text Fusion · Anomaly Detection

1 Introduction

Video action detection is a critical task in multimodal video understanding, with a wide range of real-world applications such as intelligent surveillance [26,35,37], autonomous driving [6, 30], human behavior analysis [34, 55], and embodied intelligence [9]. Robust models are required to recognize and temporally localize predefined actions as well as abnormal events such as accidents or illegal activities in long, untrimmed video streams. Early approaches extract spatio-temporal features with 3D convolutional networks, while recent methods build on Transformer architectures, state space models [14, 38], and multimodal pre-training to model longer temporal context and to align visual content with category semantics. Although recent Multimodal Large Language Models (MLLMs) [60,63] generalize well on general video tasks, their limited capacity for long-video temporal reasoning, generative uncertainty, and heavy computational cost make



(a) Framework overview (b) Performance vs. efficiency

Fig. 1: (a) MLTC-Attention models multi-scale temporal dependencies via compress-then-attend in latent space, and VTCA-Fusion enforces hierarchical video-text alignment at both standard and latent feature levels. (b) Performance–efficiency on video anomaly detection (blue, left axis) and video action detection (orange, right axis): our method (red) attains the best accuracy at the highest inference speed (FPS) with the fewest FLOPs on both tasks.

them ill-suited for fine-grained video action detection, motivating specialized architectures with efficient temporal attention and multimodal fusion. Transformer attention [47] is widely used to model temporal context in videos, but two problems remain. First, attention over long sequences repeatedly processes a large amount of redundant frame information, which limits both feature quality and training efficiency. Existing remedies either apply pooling [44], which irreversibly discards fine-grained details, or window-based attention [27], which restricts the receptive field and misses long-range dependencies; recent efficient temporal models [4] still struggle to obtain global context and local precision at the same time. To address this issue, we propose a video temporal attention module based on Multi-Scale Latent Temporal Compression Attention, termed MLTC-Attention. The key insight is to first project features into a compact latent space and then compress the temporal dimension at multiple scales before performing attention, rather than attending to all raw features directly. This compress-then-attend design, inspired by recent sparse and compressed attention [50,61], filters temporal redundancy while larger compression scales preserve global context and smaller scales preserve local details, and a gate re-weights the scales according to the input content.

Second, multimodal learning has become a key route to better video understanding, but previous video-text methods [49, 68] align video and text features at a single representation level, which limits the depth of cross-modal interaction. To address this limitation, we propose a multimodal feature fusion module, termed Video-Text Cross Attention Fusion (VTCA-Fusion), which enforces cross-modal alignment at two complementary levels: the standard feature space captures explicit semantic correspondences, while a parallel latent space branch, inspired by latent space feature interaction [62], captures compressed cross-modal patterns. By exploiting the shared category information between video and text, the module achieves fine-grained alignment and complementarity, producing fused features that are more discriminative. As shown in Fig. 1, our framework integrates MLTC-Attention and VTCA-Fusion for video action detection. To summarize, the main contributions of this paper are as follows:

– MLTC-Attention, which compresses features in a multi-scale temporal latent space and adaptively extracts dynamic features across different temporal scales. It strikes a balance between global and local feature extraction, thereby enhancing the ability to efficiently process temporal features in video.
– VTCA-Fusion, which utilizes cross attention between video features and enhanced category text features at both the standard and the latent level. It overcomes the limitations of existing methods in fusing video and text features, and achieves fine-grained alignment and complementarity between the two modalities.
– The framework, built around the MLTC-Attention and VTCA-Fusion modules, is applicable to various video action detection tasks. Experimental validation on four benchmarks demonstrates the effectiveness of the proposed modules.

2 Related Work

2.1 Video Understanding Models

Video representation learning has progressed from 3D CNNs such as C3D [46] and I3D [3], to Transformers such as ViViT [1] and Video Swin Transformer [27], self-supervised VideoMAE [48] and InternVideo [51], and SSM-based VideoMamba [19]. On the multimodal side, CLIP [40], ImageBind [13], and LanguageBind [68] align vision with language and other modalities, Video-LLaVA [22] and InternVL [8] connect video encoders with large language models, and M2-CLIP [49] adapts multimodal models to video action recognition.

2.2 Video Action Detection Models

Temporal action detection has evolved from boundary-matching methods like BMN [23] and graph-based G-TAD [56] to Transformer-based approaches such as ActionFormer [64]. TriDet [42] employs multi-scale feature pyramids, TemporalMaxer [44] uses max-pooling, ActionMamba [4] and efficient temporal attention with state space models [36, 39] leverage SSM [14] for global temporal modeling, and recent works [12, 20] explore cross-layer decoupling and boundary refinement.

2.3 Video Anomaly Detection Models

MIL-based methods [43] learn anomaly scores from video-level labels and were later improved with graph convolutions [66], feature magnitude learning [45], multi-sequence Transformers [21], and dual memory units [67]. Recent methods transfer CLIP [40] knowledge to weakly supervised anomaly detection [33,53,62], use text prompts for normality guidance [59] or MIL enhancement [5], and model short and long anomalies across temporal scales [31].

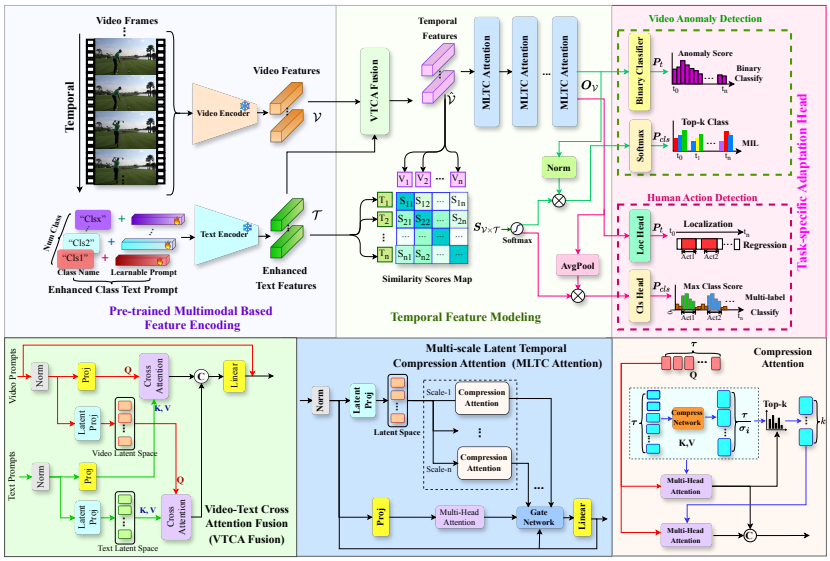


Fig. 2: Overview of the overall architecture of our proposed video action detection framework. Our method integrates the VTCA-Fusion module and MLTC-Attention Blocks to process temporal and categorical features, with the resulting predictions feeding into various downstream tasks.

3 Method

3.1 Overview

Figure 2 shows the overall framework. Given a video of τ frames, the pre-trained video encoder $f_{\mathcal{V}}(\cdot)$ of the multimodal model extracts frame features $\mathcal{V} = [v_1, \cdots, v_\tau] \in \mathbb{R}^{\tau \times d}$. For each of the c categories, the category text t_c is concatenated with the learnable prompt t_L, the Enhanced Context, and encoded by the pre-trained text encoder $f_{\mathcal{T}}(\cdot)$; the resulting text features $\mathcal{T} \in \mathbb{R}^{c \times d}$ are projected into the video feature space. The VTCA-Fusion module fuses $\mathcal{V}$ and $\mathcal{T}$ through standard and latent space cross attention, producing the fused sequence $\hat{\mathcal{V}} = [\hat{v}_1, \cdots, \hat{v}_\tau] \in \mathbb{R}^{\tau \times d}$ with the temporal resolution preserved, together with a video-text similarity map $\boldsymbol{S}_{\mathcal{V} \times \mathcal{T}} \in \mathbb{R}^{\tau \times c}$. The MLTC-Attention module then compresses and emphasizes critical temporal information of $\hat{\mathcal{V}}$ in the latent space via multi-scale compression attention, outputting temporal features $\boldsymbol{O}_{\mathcal{V}} \in \mathbb{R}^{\tau \times d}$ that capture both fine-grained details and long-range temporal context. Integrating $\boldsymbol{O}_{\mathcal{V}}$ with $\boldsymbol{S}_{\mathcal{V} \times \mathcal{T}}$ yields the frame-level temporal scores $\boldsymbol{P}_t \in \mathbb{R}^{\tau}$ and the class scores $\boldsymbol{P}_{cls} \in \mathbb{R}^{\tau \times c}$, which feed the task-specific loss functions.

3.2 Video-Text Cross Attention Fusion Module

Standard cross-modal fusion operates at a single representation level, which may overfit to surface-level correlations while missing deeper semantic structures. To overcome this, the VTCA-Fusion module performs cross attention at both standard and latent feature levels simultaneously: the standard-level branch preserves full-dimensional semantic details for explicit alignment, while the latent-level branch operates in a compressed space that encourages the model

to learn more abstract cross-modal representations. Its architecture is illustrated in the bottom-left part of Figure 2. We utilize the cross attention, where video frame features attend to text prompt features. Specifically, the video features $\mathcal{V} \in \mathbb{R}^{\tau \times d}$ are treated as queries, while the enhanced text prompt embeddings $\mathcal{T} \in \mathbb{R}^{c \times d}$ serve as both keys and values within a multi-head attention. We denote $\boldsymbol{Q}_{\mathcal{V}} \in \mathbb{R}^{\tau \times d}, \boldsymbol{K}_{\mathcal{T}} \in \mathbb{R}^{c \times d}, \boldsymbol{V}_{\mathcal{T}} \in \mathbb{R}^{c \times d}$ represent the input for standard multi-head attention, and $\boldsymbol{Q}^l_{\mathcal{V}} \in \mathbb{R}^{\tau \times d_l}, \boldsymbol{K}^l_{\mathcal{T}} \in \mathbb{R}^{c \times d_l}, \boldsymbol{V}^l_{\mathcal{T}} \in \mathbb{R}^{c \times d_l}$ represent the input for attention of latent space features from the Latent Projection, $d_l < d \in \mathbb{Z}^+$ denotes the output feature dimension in the latent space, which reduces the dimension to compress the feature information, thereby enabling the interaction between video and enhanced text features within the latent space. The attention outputs are computed separately, as follows:

$$\boldsymbol{Attn}_{\mathcal{V}\times\mathcal{T}} = \mathbf{softmax}\left(\frac{\boldsymbol{Q}_{\mathcal{V}}\boldsymbol{K}_{\mathcal{T}}^{\top}}{\sqrt{d}}\right)\boldsymbol{V}_{\mathcal{T}}, \quad \boldsymbol{Attn}^l_{\mathcal{V}\times\mathcal{T}} = \mathbf{softmax}\left(\frac{\boldsymbol{Q}^l_{\mathcal{V}}\boldsymbol{K}^{l^{\top}}_{\mathcal{T}}}{\sqrt{d_l}}\right)\boldsymbol{V}^l_{\mathcal{T}}, \tag{1}$$

In Eq. 1, $\boldsymbol{Attn}_{\mathcal{V}\times\mathcal{T}} \in \mathbb{R}^{\tau \times d}$ represents the result of standard multi-head attention, where $\boldsymbol{Q}_{\mathcal{V}}\boldsymbol{K}_{\mathcal{T}}^{\top}$ yields the similarity $\boldsymbol{S}_{\mathcal{V}\times\mathcal{T}}$ between the video and enhanced text features, and $\boldsymbol{Attn}^l_{\mathcal{V}\times\mathcal{T}} \in \mathbb{R}^{\tau \times d_l}$ represents the computation result of multi-head attention in latent space. Finally, we construct an output projection weight tensor $\boldsymbol{W}_o \in \mathbb{R}^{(d+d_l)\times d}$ to fuse the outputs of the standard attention and latent space attention, the final result of VTCA-Fusion module is computed, as follows:

$$\hat{\mathcal{V}} = \mathcal{V} + \boldsymbol{W}_o(\boldsymbol{Attn}_{\mathcal{V}\times\mathcal{T}} \| \boldsymbol{Attn}^l_{\mathcal{V}\times\mathcal{T}}). \tag{2}$$

where $\|$ denotes the operation of concatenating tensors. Through this computation, we obtain the temporal feature $\hat{\mathcal{V}} \in \mathbb{R}^{\tau \times d}$ that fuses the video and text features.

3.3 Multi-Scale Latent Temporal Compression Attention Block

Standard self-attention [47] over the full temporal sequence incurs $O(\tau^2)$ complexity, while pooling-based methods [44] reduce cost but irreversibly discard fine-grained information. Window-based attention [27] restricts the receptive field and cannot model long-range dependencies. Our MLTC addresses this by first projecting features into a lower-dimensional latent space to reduce per-token cost, then compressing the temporal dimension at multiple scales so that larger scales capture global context while smaller scales preserve local dynamics. Attending to the compressed keys and values reduces the cost of each compression branch from $O(\tau^2)$ to $O(\tau^2/\sigma_i)$, while a single global self-attention branch is retained to preserve the full receptive field. The architecture of the MLTC is shown in the bottom-center and bottom-right parts of Figure 2. The overall input to the module is the temporal feature $\mathcal{X}_{\mathcal{V}} \in \mathbb{R}^{\tau \times d}$. First, the Latent Projection maps $\mathcal{X}_{\mathcal{V}}$ into the latent space, giving $\boldsymbol{Q}^l_t, \boldsymbol{K}^l_t, \boldsymbol{V}^l_t \in \mathbb{R}^{\tau \times \frac{d}{2}}$ as the inputs of the multi-scale latent compression attention. In the bottom-right part of Figure 2, the scales set $[\textit{scale-1}, \cdots, \textit{scale-}n]$ denote the compression sizes along the temporal dimension, with each scale corresponding to a Compression Attention module.

The total temporal dimension of the Compression Attention input is τ; $\boldsymbol{K}_t^l$ and $\boldsymbol{V}_t^l$ are each transformed along the temporal dimension by their own Compress Network of *scale-i*, as follows:

$$\boldsymbol{K}_{t_i}^l = \boldsymbol{W}_{\boldsymbol{i,K}}^{\boldsymbol{out}} \mathbf{GELU}(\boldsymbol{W}_{\boldsymbol{i,K}}^{\boldsymbol{in}} \boldsymbol{K}_t^l), \quad \boldsymbol{V}_{t_i}^l = \boldsymbol{W}_{\boldsymbol{i,V}}^{\boldsymbol{out}} \mathbf{GELU}(\boldsymbol{W}_{\boldsymbol{i,V}}^{\boldsymbol{in}} \boldsymbol{V}_t^l), \tag{3}$$

where $\boldsymbol{W}_{\boldsymbol{i,*}}^{\boldsymbol{in}} \in \mathbb{R}^{\frac{\tau}{\sigma_i} \times \tau}$, $\boldsymbol{W}_{\boldsymbol{i,*}}^{\boldsymbol{out}} \in \mathbb{R}^{\frac{\tau}{\sigma_i} \times \frac{\tau}{\sigma_i}}$, $* \in \{K, V\}$, denote the input and output weights of the two Compress Networks, which compress the temporal dimension from τ to $\frac{\tau}{\sigma_i}$; i indexes the scales set $[\textit{scale-1}, \cdots, \textit{scale-n}]$, $\sigma_i \in \mathbb{Z}^+$ is the compression size of *scale-i*, and **GELU** is the activation function. The outputs $\boldsymbol{K}_{t_i}^l, \boldsymbol{V}_{t_i}^l \in \mathbb{R}^{\frac{\tau}{\sigma_i} \times \frac{d}{2}}$ are the inputs of the i-th Compression Attention module. In the Compression Attention module, the computation is first performed based on the input tensors $\boldsymbol{Q}_t^l$ with the compressed $\boldsymbol{K}_{t_i}^l$ and $\boldsymbol{V}_{t_i}^l$ as follows:

$$\mathcal{S}_i = \mathbf{softmax}\left(\frac{\boldsymbol{Q}_t^l \boldsymbol{K}_{t_i}^{l^\top}}{\sqrt{d/2}}\right), \quad \boldsymbol{Attn}_i^{cp} = \mathcal{S}_i \boldsymbol{V}_{t_i}^l, \tag{4}$$

where $\mathcal{S}_i \in \mathbb{R}^{\tau \times \frac{\tau}{\sigma_i}}$ represents the similarity scores tensor between $\boldsymbol{Q}_t^l$ and $\boldsymbol{K}_{t_i}^l$. $Attn_i^{cp} \in \mathbb{R}^{\tau \times \frac{d}{2}}$ represents the computation result of the attention, *cp* means the compressed $\boldsymbol{K}_{t_i}^l, \boldsymbol{V}_{t_i}^l$. Secondly, to capture the critical temporal information from $\boldsymbol{K}_{t_i}^l$ and $\boldsymbol{V}_{t_i}^l$, we perform top-k selection based on $\mathcal{S}_i$, as follows:

$$\mathcal{I}_i = \mathbf{Top\text{-}K}\left(\mathcal{S}_i(t), k\right), \tag{5}$$

$$\boldsymbol{K}_{t_i}^{\mathcal{I}_i} \boldsymbol{V}_{t_i}^{\mathcal{I}_i} = \Big\|_{j \in \mathcal{I}_i} \boldsymbol{K}_t^l \boldsymbol{V}_t^l [j\sigma_i : (j+1)\sigma_i], \tag{6}$$

where $\mathcal{I}_i \in \mathbb{R}^k$ denotes the indices of the top-k most relevant temporal segments in $\boldsymbol{K}_{t_i}^l, \boldsymbol{V}_{t_i}^l$ with $\boldsymbol{Q}_t^l$. Following the top-k selection strategy in [43], we set $k = \max(1, \lfloor \frac{\tau}{\sigma_i}/8 \rfloor)$, which selects the top 1/8 of the compressed blocks at each scale with a minimum of 1. The $\boldsymbol{K}_{t_i}^{\mathcal{I}_i} \in \mathbb{R}^{\sigma_i k \times \frac{d}{2}}, \boldsymbol{V}_{t_i}^{\mathcal{I}_i} \in \mathbb{R}^{\sigma_i k \times \frac{d}{2}}$ represent the key and value pairs retrieved via $\mathcal{I}_i$, and they together with $\boldsymbol{Q}_t^l$, form the input tensors for computing the attention of critical temporal segments, as follows:

$$\boldsymbol{Attn}_i^{tk} = \mathbf{softmax}\left(\frac{\boldsymbol{Q}_t^l \boldsymbol{K}_{t_i}^{\mathcal{I}_i^\top}}{\sqrt{d/2}}\right) \boldsymbol{V}_{t_i}^{\mathcal{I}_i}, \tag{7}$$

where $Attn_i^{tk} \in \mathbb{R}^{\tau \times \frac{d}{2}}$ represents the attention result obtained from the top-k filtering method, enabling the model to enhance feature learning for critical temporal segments. The results $Attn_i^{cp}, Attn_i^{tk}$ obtained through the above methods are concatenated, as follows:

$$\boldsymbol{Attn}_i = \boldsymbol{Attn}_i^{cp} \| \boldsymbol{Attn}_i^{tk}, \tag{8}$$

where $Attn_i \in \mathbb{R}^{\tau \times d}$ is the complete i-th Compression Attention output result. We also apply a Global Attention $Attn_{global}$, which performs standard self-attention [47] over the entire video sequence $\mathcal{X}_\mathcal{V}$ to capture long-range temporal

dependencies and holistic context. Finally, we introduce a gate network with weights applied to each output to fuse Compression Attention across various scales and Global Attention, with the process described as follows:

$$\boldsymbol{O}_{\mathcal{V}} = \boldsymbol{W}_{gate} \odot (Attn_{global} \| \underbrace{Attn_1 \| \cdots \| Attn_n}_{n}), \tag{9}$$

where $\boldsymbol{W}_{gate} \in \mathbb{R}^{\tau \times (n+1)}$ denotes the gating weights used to allocate weights to each output, and $\boldsymbol{O}_{\mathcal{V}} \in \mathbb{R}^{\tau \times d}$ indicates the fused output. It is further processed by a standard LayerNorm and FFN layer to obtain the final representation of the MLTC-Attention module. Based on temporal feature output $\boldsymbol{O}_{\mathcal{V}}$ and the Video-Text Similarity map $\boldsymbol{S}_{\mathcal{V} \times \mathcal{T}}$, the post-processing step is performed to generate the final temporal and categorical predictions, the $\boldsymbol{P}_t \in \mathbb{R}^{\tau}$ represents the temporal prediction scores along the temporal dimension, and the $\boldsymbol{P}_{cls} \in \mathbb{R}^{\tau \times c}$ provides the class-level scores across time.

3.4 Loss Functions

Loss of Video Anomaly Detection. Following the standard weakly supervised setting [43, 53, 67], only video-level annotations are available for training: given the video-level label $Y \in \{0, 1\}$, 0 for normal and 1 for abnormal, we broadcast it along the temporal dimension as $\boldsymbol{y}^i = Y$ for $i = 1, \cdots, \tau$, so that no frame-level annotation is used. A binary cross-entropy loss $\mathcal{L}_{BCE}^{a\text{-}n}$ is computed between the predicted scores $\boldsymbol{P}_t$ and $\boldsymbol{y}$. Since an abnormal video still contains normal frames, this broadcast label is inevitably noisy for abnormal videos; the top-k lowest-scoring frames of an abnormal video are therefore additionally selected as normal instances $\boldsymbol{P}_{k\text{-}n} \in \mathbb{R}^k$ with loss $\mathcal{L}_{BCE}^{k\text{-}n}$, which lets the model learn normal features from abnormal videos and mitigates the label noise introduced by the broadcast. Following MIL [43], the mean of the top-k scores of each category in $\boldsymbol{P}_{cls}$ along the temporal dimension gives the predicted logits $\tilde{\boldsymbol{P}}_{cls} \in \mathbb{R}^c$, trained with a multi-class cross-entropy $\mathcal{L}_{CE}^{cls}$ against the category labels $\boldsymbol{y}_{cls} \in \mathbb{R}^c$:

$$\begin{aligned}
&\mathcal{L}_{BCE}^{a\text{-}n} = -\frac{1}{\tau}\sum_{i=1}^{\tau}[\boldsymbol{y}^i \log(\boldsymbol{P}_t^i) + (1-\boldsymbol{y}^i)\log(1-\boldsymbol{P}_t^i)], \quad \mathcal{L}_{BCE}^{k\text{-}n} = -\frac{1}{k}\sum_{i \in k}\log(1-\boldsymbol{P}_{k\text{-}n}^i), \\
&\mathcal{L}_{CE}^{cls} = -\sum_{i=1}^{c}\boldsymbol{y}_{cls}^i \log(\tilde{\boldsymbol{P}}_{cls}^i), \quad \mathcal{L}_{a\text{-}n} = \mathcal{L}_{BCE}^{a\text{-}n} + \mathcal{L}_{BCE}^{k\text{-}n} + \mathcal{L}_{CE}^{cls},
\end{aligned} \tag{10}$$

where the total loss $\mathcal{L}_{a\text{-}n}$ preserves both the binary anomaly classification and the multi-category loss commonly used in MIL.

Loss of Temporal Action Detection. We use Focal Loss [24] for action classification, $\mathcal{L}_{cls}$, and Distance-IoU [65] for temporal segment regression, $\mathcal{L}_{iou}$:

$$\mathcal{L}_{act} = \frac{1}{\tau^+}\sum_{i=1}^{\tau}[\mathcal{L}_{cls}(\boldsymbol{P}_{cls}^i, \boldsymbol{y}_{cls}^i) + \lambda \mathcal{L}_{iou}(\boldsymbol{P}_t^i, \boldsymbol{y}_t^i)], \tag{11}$$

where $\boldsymbol{y}_{cls}, \boldsymbol{y}_t \in \mathbb{R}^\tau$ are the labels for action categories and temporal segments, τ^+ denotes the number of positive samples assigned during training, and Soft Non-Maximum Suppression is applied as post-processing at inference, and λ is the weight of $\mathcal{L}_{iou}$.

4 Experiments

4.1 Datasets and Evaluation Metrics

For video anomaly detection we use UCF-Crime [43], over 1,900 surveillance videos covering 13 real-world abnormal events together with normal activities, and XD-Violence [52], about 4,700 videos from movies and social media covering 6 categories of abnormal events. For temporal action detection we use THUMOS14 [16], 413 untrimmed videos with temporal annotations over 20 action categories, and ActivityNet [2], 19,994 videos with temporal annotations over 200 action categories. Following common practice, we report the frame-level AUC of the ROC curve on UCF-Crime and the frame-level Average Precision (AP) on XD-Violence, which better reflects the precision–recall balance under class imbalance; for action detection we report the mean Average Precision (mAP) averaged over multiple temporal IoU (tIoU) thresholds between predicted and ground-truth segments.

4.2 Implementation Details

We implemented our method on a single NVIDIA L40S GPU. Both pre-trained encoders $f_{\mathcal{V}}(\cdot)$ and $f_{\mathcal{T}}(\cdot)$ are frozen; only the projection layers, the learnable prompts, VTCA-Fusion, MLTC-Attention and the task head are trained, so all text-encoder variants share the same protocol and schedule. During the training process, we use the AdamW [28] optimizer and employ a scheduler that integrates linear warm-up and cosine annealing to adjust the learning rate, which is set to 1e-5. The batch size was set to 64. For video data input, we utilized pretrained encoder weights to extract video features, and the sequence length of video feature input is 512 frames. Each frame was resized to a resolution of 224×224 pixels. For the text input with the enhanced text prompt design, where each textual class label is embedded with 10 learnable prompt sequence length. The maximum text sequence length is set to 79, following the configuration of CLIP [40]. The model was trained for 20 epochs.

4.3 Evaluation and Comparison

Results of Video Anomaly Detection. Table 1 compares our method with previous state-of-the-art methods on UCF-Crime and XD-Violence. Our method achieves the highest AUC on UCF-Crime and, without incorporating any audio features, the highest AP on XD-Violence. Compared with recent LLM-based approaches, LAVAD [63], EventVAD [41], and VERA [60], our method obtains

Table 1: Comparison of frame-level performance between our method and previous methods on the UCF-Crime and XD-Violence benchmarks. †: results reproduced by us with the listed vision encoder.

Method	Vision Encoder	UCF-Crime AUC(%)	XD-Violence AP(%)
LAVAD [63]	ImageBind [13]	80.28	62.01
DMU† [67]	I3D [3]	86.14	81.46
AnomalyCLIP [62]	ViT-B16 [11]	86.30	77.40
VERA [60]	InternVL2 [8]	86.55	70.54
UMIL [29]	ViT-B32 [11]	86.75	–
PE-MIL [5]	I3D [3]	86.83	–
MGFN [7]	VideoSwin [27]	86.97	79.31
OE-CTST [31]	I3D [3]	86.99	81.78
CLIP-TSA [17]	ViT-B16 [11]	–	82.15
EventVAD [41]	ViT-B16 [11]	87.51	64.04
VadCLIP† [53]	I3D [3]	87.57	82.48
TPWNG [59]	ViT-B16 [11]	87.79	83.68
MLTC+VTCA+CLIP [40]	I3D [3]	88.68	83.44
MLTC+VTCA+CLIP [40]	ViT-B16 [11]	88.63	84.63
MLTC+VTCA+LangBind [68]	ViT-B16 [11]	89.17	84.89
MLTC+VTCA+DeBERTa [15]	ViT-B16 [11]	**89.84**	**86.53**

Table 2: Comparison on the THUMOS14 and ActivityNet-1.3 with the mAP of tIoU. †: results reproduced by us with the listed vision encoder.

Model	Vision Encoder	THUMOS14						ActivityNet-1.3			
		0.3	0.4	0.5	0.6	0.7	Avg	0.5	0.75	0.95	Avg
ActionFormer [64]	I3D [3]	82.10	77.80	71.00	59.40	43.90	66.80	55.50	38.12	8.78	37.05
TemporalMaxer [44]	InternVidv2 [51]	85.23	81.27	74.17	63.70	49.14	70.70	60.99	40.53	9.79	38.59
TriDet [42]	InternVidv2 [51]	85.81	81.78	75.61	64.64	50.50	71.67	56.40	38.74	9.60	37.60
Dyfadet [58]	VideoMAEv2 [48]	85.99	81.66	76.32	64.46	50.08	71.70	58.19	39.30	8.63	38.62
AdaTAD† [25]	InternVidv2 [51]	86.75	82.25	76.10	65.28	50.29	72.13	59.99	39.53	9.79	38.69
CLTDR-GMG [20]	InternVidv2 [51]	87.00	84.00	78.40	67.90	54.00	74.30	55.00	38.00	8.60	37.10
ActionMamba [4]	InternVidv2 [51]	87.05	82.66	76.51	65.56	49.87	72.33	62.20	43.14	9.57	41.73
BDRC-Net [12]	InternVidv2 [51]	88.60	83.80	78.00	66.80	51.20	73.70	61.40	42.10	10.20	41.10
DiGIT [18]	InternVidv2 [51]	87.60	84.20	77.60	67.30	52.50	73.80	62.00	43.10	11.30	42.00
MLTC+VTCA+CLIP [40]	InternVidv2 [51]	88.95	85.12	78.40	68.02	52.51	74.60	63.81	45.34	10.92	43.59
MLTC+VTCA+LangBind [68]	InternVidv2 [51]	88.99	85.22	78.73	68.30	52.78	74.80	63.84	45.35	11.17	43.82
MLTC+VTCA+DeBERTa [15]	InternVidv2 [51]	**89.50**	**85.38**	**78.82**	**69.05**	**54.15**	**75.37**	**63.87**	**45.43**	**11.36**	**43.90**

Table 3: Ablation for module components and for the latent projection in each module.

Module					UCF-C	THUMOS
MLTC	Top-k	VTCA	Latent (MLTC)	Latent (VTCA)	AUC	mAP
		✓		✓	86.00	72.53
✓		✓	✓	✓	88.27	73.75
✓	✓		✓		86.65	73.23
✓	✓	✓			88.19	73.68
✓	✓	✓	✓		89.09	74.92
✓	✓	✓		✓	88.92	74.66
✓	✓	✓	✓	✓	89.84	75.37

Table 4: Ablation results for alternative module design.

Attention Modules		UCF-C	THUMOS
Temporal Attn	V&T Fusion	AUC	mAP
Self-Attn [47]	VTCA	86.00	72.53
Soft-Attn [57]	VTCA	86.14	72.64
MS-TCT [10]	VTCA	88.01	73.44
MLTC	Cross-Attn [32]	88.19	73.23
MLTC	VTCA	89.84	75.37

clearly higher accuracy on both datasets without relying on large language models, indicating that LLM-based approaches show no clear advantage on fine-grained temporal localization while requiring substantially higher computational cost, whereas our targeted attention and fusion designs are more effective for fine-grained temporal anomaly detection.

Results of Temporal Action Detection. Table 2 reports the results on THUMOS14 and ActivityNet-1.3. On THUMOS14, our method achieves the best average mAP over tIoU thresholds from 0.3 to 0.7 and consistently maintains the best mAP across all thresholds; on ActivityNet-1.3, it also achieves the best overall performance and continues to lead at every tIoU threshold. Our method consistently outperforms both earlier Transformer- and SSM-based detectors and the most recent state-of-the-art methods on both datasets.

Computational Efficiency Analysis. Table 7 compares the computational cost and inference speed of our MLTC module with previous methods under the same input resolution and hardware settings, on a single NVIDIA L40S GPU with 512-frame inputs at 224×224 resolution, excluding I/O and prepro-

Table 5: Ablation for different temporal compression scales.

Scales	UCF-C	THUMOS
	AUC	mAP
w/o	86.00	72.53
[4,16,32]	88.77	74.32
[8,16,64]	89.05	74.57
[8,32,64]	89.84	75.37
[16,64,128]	88.91	74.65

Table 6: Ablation for text prompt types.

Prompt Types	UCF-C	THUMOS
	AUC	mAP
w/o prompt	86.65	73.23
Random	86.72	73.31
Generic	86.84	73.45
Class-Only	88.74	74.62
Learnable	89.84	75.37

Table 7: Efficiency comparison with prior methods.

Method	FPS	FLOPs
MGFN [7]	61.30	57.46G
AnomCLIP [62]	69.98	50.4G
STPrompt [54]	79.19	49.6G
VadCLIP [53]	78.34	44.8G
Ours	83.77	42.1G

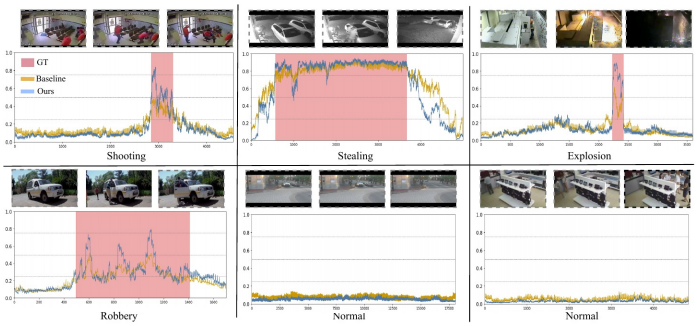


Fig. 3: Qualitative results of our proposed methods on Video Anomaly Detection tasks.

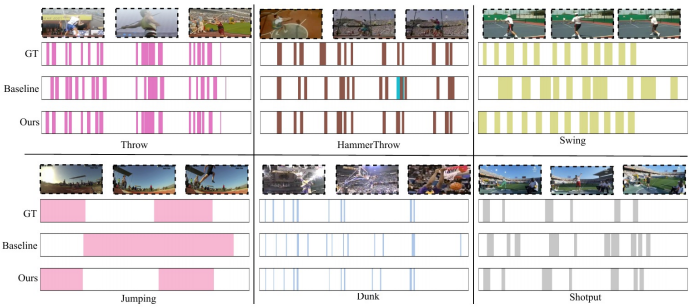


Fig. 4: Qualitative results of our proposed methods on temporal action detection tasks.

cessing overhead. Our method achieves the highest throughput with the fewest FLOPs among all compared methods while also being the most accurate, reducing the computational cost compared with prior methods while simultaneously improving accuracy. By projecting features into a compact latent space and compressing temporal redundancy before the attention computation, our compress-then-attend paradigm achieves a better accuracy–efficiency trade-off than prior methods that increase complexity for marginal gains.

4.4 Qualitative Analyses

We present visualization examples of temporal predictions produced by our model. Figure 3 shows Video Anomaly Detection results, where the curves are predicted anomaly scores over time and the red shaded regions are ground-truth (GT) anomaly intervals, and Figure 4 shows temporal action detection results, where colored segments are predicted action categories and temporal intervals; compared with the baseline, our method yields more precise temporal predictions on both tasks. Figure 5 visualizes the category-wise temporal attention: for

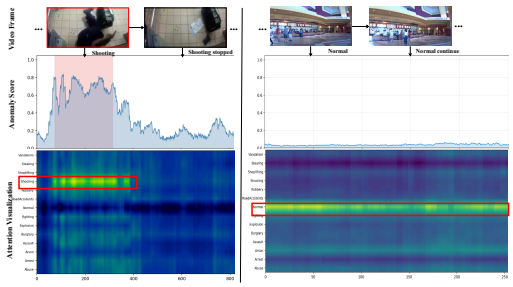


Fig. 5: Visualization of internal representation category-wise temporal feature.

anomaly videos the model correctly activates on the corresponding category during the GT interval, while for normal videos "Normal" dominates the response, reflecting the joint effect of MLTC-Attention, which captures fine-grained dynamics across multiple scales, and VTCA-Fusion, which aligns video features with category-specific textual semantics.

4.5 Ablation Studies

To better analyze the contribution of each component in our model, we conduct ablation studies on the representative benchmarks UCF-Crime [43] and THUMOS14 [16].

Contribution of Our Proposed Modules. As shown in Table 3, we ablate each module within the overall architecture and evaluate on representative datasets for each task. First, for the MLTC-Attention module, we replace it with standard attention. The results on both UCF-Crime and THUMOS14 datasets show a noticeable performance decline, demonstrating the superiority of MLTC-Attention in capturing temporal features. We also conducted experiments on the effect of the Top-k selection, which retains the top $1/8$ of the compressed blocks at each scale, at least one block, following strategies in previous works [42]. Its removal led to a slight performance decrease on both datasets, further validating the effectiveness of this filtering mechanism in focusing on the most informative temporal segments. When the VTCA-Fusion module is removed, the correlation between video and text features weakens, resulting in a performance drop on both tasks. Finally, the latent component, which applies latent space attention within both the MLTC-Attention and VTCA-Fusion modules, is crucial for understanding key semantic features in latent space; its removal degrades performance significantly.

Alternative of Our Proposed Modules. We replace MLTC-Attention or VTCA-Fusion with several previously used temporal and cross-modal attention modules. As shown in Table 4, substituting MLTC with Self-Attention, Soft-Attention, or MS-TCT consistently lowers performance, indicating that multi-scale temporal compression is crucial for modeling long-range dependencies. Similarly, replacing VTCA with standard Cross-Attention also degrades results, demonstrating that our multi-level video-text fusion achieves more effective cross-modal alignment.

Multi-scale Compression factors. To analyze the effect of different temporal compression scales on the MLTC-Attention module, we conducted an ablation study by varying scales. As shown in the Table 5, the w/o denotes the absence of compression scales, corresponding to the standard attention. The smaller compression set [4, 16, 32] preserves fine-grained temporal details but introduces redundant computations, while the larger compression set [16, 64, 128] overly compresses the temporal dimension, leading to the loss of local dynamic characteristics. The configuration [8, 32, 64] achieves a better balance between feature compression and redundancy reduction. Intuitively, the larger scales 32 and 64 capture global anomaly regions across the temporal extent, while the smaller scale 8 refines local action boundaries, and their combination achieves sharper activation within ground-truth intervals.

Prompt Types. As shown in Table 6, removing textual input, denoted w/o in the table, clearly degrades performance, indicating that text provides useful semantic priors. Random and generic prompts introduce only weak guidance, as they lack category-aligned semantics. Using Class-Only prompts brings some class information but remains limited. In contrast, the Learnable prompt achieves the best results, showing that learnable prompt tokens effectively capture category-specific semantics and significantly improve multimodal feature learning.

5 Conclusion

We tackle key problems in multimodal video action detection by our novel solutions. First, we propose the MLTC-Attention module, which utilizes a multi-scale temporal latent space compression attention. It is designed to adaptively capture important dynamic features across different temporal scales, effectively reducing the computational redundancy and inefficiency. Second, we present the VTCA-Fusion module, which employs cross attention between video content and enhanced multi-category text. This enables precise alignment and complementary integration of visual and textual features, addressing limitations in existing multimodal feature integration methods. Extensive experiments on temporal action detection and video anomaly detection benchmarks demonstrate that our methods consistently improve accuracy and achieve superior overall performance.

Acknowledgments. This work was supported by JST CRONOS JPMJCS24K4 and JSPS KAKENHI 26K23875.

Disclosure of Interests. The authors declare no competing interests.

References

1. Arnab, A., Dehghani, M., Heigold, G., et al.: ViViT: A video vision transformer. In: ICCV. pp. 6836–6846 (2021)
2. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: ActivityNet: A large-scale video benchmark for human activity understanding. In: CVPR. pp. 961–970 (2015)

3. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR. pp. 6299–6308 (2017)
4. Chen, G., Huang, Y., Xu, J., et al.: Video Mamba suite: State space model as a versatile alternative for video understanding. IJCV **134**(1), 20 (2025)
5. Chen, J., Li, L., Su, L., et al.: Prompt-enhanced multiple instance learning for weakly supervised video anomaly detection. In: CVPR. pp. 18319–18329 (2024)
6. Chen, L., Ma, N., Wang, P., et al.: Survey of pedestrian action recognition techniques for autonomous driving. Tsinghua Science and Technology **25**(4), 458–470 (2020)
7. Chen, Y., Liu, Z., Zhang, B., et al.: MGFN: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In: AAAI. vol. 37, pp. 387–395 (2023)
8. Chen, Z., Wu, J., Wang, W., et al.: InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: CVPR. pp. 24185–24198 (2024)
9. Cheng, S., Fang, K., Yu, Y., et al.: VidEgoThink: Assessing egocentric video understanding capabilities for embodied ai. arXiv preprint arXiv:2410.11623 (2024)
10. Dai, R., Das, S., Kahatapitiya, K., et al.: MS-TCT: Multi-scale temporal convtransformer for action detection. In: CVPR. pp. 20041–20051 (2022)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)
12. Fang, Z., Yu, J., Hong, R.: Boundary discretization and reliable classification network for temporal action detection. IEEE TMM (2025)
13. Girdhar, R., El-Nouby, A., Liu, Z., et al.: ImageBind: One embedding space to bind them all. In: CVPR. pp. 15180–15190 (2023)
14. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. In: First Conference on Language Modeling (COLM) (2024)
15. He, P., Gao, J., Chen, W.: DeBERTaV3: Improving DeBERTa using electra-style pre-training with gradient-disentangled embedding sharing. arXiv preprint arXiv:2111.09543 (2021)
16. Jiang, Y.G., Liu, J., Roshan Zamir, A., et al.: THUMOS challenge: Action recognition with a large number of classes (2014)
17. Joo, H.K., Vo, K., Yamazaki, K., Le, N.: CLIP-TSA: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 3230–3234 (2023)
18. Kim, H.J., Lee, Y., Hong, J.H., Lee, S.W.: DiGIT: Multi-dilated gated encoder and central-adjacent region integrated decoder for temporal action detection transformer. In: CVPR. pp. 24286–24296 (2025)
19. Li, K., Li, X., Wang, Y., et al.: VideoMamba: State space model for efficient video understanding. In: ECCV. pp. 237–255 (2024)
20. Li, Q., Liu, D., Kong, J., et al.: Temporal action localization with cross layer task decoupling and refinement. In: AAAI. vol. 39, pp. 4878–4886 (2025)
21. Li, S., Liu, F., Jiao, L.: Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In: AAAI. vol. 36, pp. 1395–1403 (2022)
22. Lin, B., Ye, Y., Zhu, B., et al.: Video-LLaVA: Learning united visual representation by alignment before projection. In: EMNLP. pp. 5971–5984 (2024)
23. Lin, T., Liu, X., Li, X., et al.: BMN: Boundary-matching network for temporal action proposal generation. In: ICCV. pp. 3889–3898 (2019)
24. Lin, T.Y., Goyal, P., Girshick, R., et al.: Focal loss for dense object detection. In: ICCV. pp. 2980–2988 (2017)
25. Liu, S., Zhang, C.L., Zhao, C., Ghanem, B.: End-to-end temporal action detection with 1B parameters across 1000 frames. In: CVPR. pp. 18591–18601 (2024)
26. Liu, Y., Ju, B., Yang, D., et al.: Memory-enhanced spatial-temporal encoding framework for industrial anomaly detection system. Expert Systems with Applications **250**, 123718 (2024)
27. Liu, Z., Ning, J., Cao, Y., et al.: Video Swin transformer. In: CVPR. pp. 3202–3211 (2022)
28. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
29. Lv, H., Yue, Z., Sun, Q., et al.: Unbiased multiple instance learning for weakly supervised video anomaly detection. In: CVPR. pp. 8022–8031 (2023)
30. Ma, Y., Yuan, L., Abdelraouf, A., et al.: M2DAR: Multi-view multi-scale driver action recognition with vision transformer. In: CVPR. pp. 5286–5293 (2023)
31. Majhi, S., Dai, R., Kong, Q., et al.: OE-CTST: Outlier-embedded cross temporal scale transformer for weakly-supervised video anomaly detection. In: WACV. pp. 3218–3228 (2024)
32. Ouenniche, K., Tapu, R., Zaharia, T.: Vision-text cross-modal fusion for accurate video captioning. IEEE Access **11**, 115477–115492 (2023)
33. Pu, Y., Wu, X., Yang, L., Wang, S.: Learning prompt-enhanced context features for weakly-supervised video anomaly detection. IEEE TIP (2024)
34. Qiu, Y., Niu, L., Sha, F.: Multipath 3D-Conv encoder and temporal-sequence decision for repetitive-action counting. Expert Systems with Applications **249**, 123760 (2024)
35. Qiu, Y., Sha, F., Niu, L.: Dka-yolo: Enhanced small object detection via dilation kernel aggregation convolution modules. IEEE Access **12**, 187353–187366 (2024)

36. Qiu, Y., Sha, F., Niu, L.: Efficient temporal attention with state space model for temporal action localization. In: ICONIP. pp. 183–197 (2024)
37. Qiu, Y., Sha, F., Niu, L., Zhang, G.: Fire anomaly detection based on low-rank adaption fine-tuning and localization using gradient filtering. Applied Soft Computing p. 112782 (2025)
38. Qiu, Y., Yanai, K.: DST-Mamba: Spatial–temporal adapter with boundary-aware Mamba for video temporal action detection. Big Data and Cognitive Computing **10**(7), 231 (2026)
39. Qiu, Y., Yanai, K.: Efficient spatial-temporal focal adapter with SSM for temporal action detection. arXiv preprint arXiv:2604.09164 (2026)
40. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
41. Shao, Y., He, H., Li, S., et al.: EventVAD: Training-free event-aware video anomaly detection. In: ACM MM. pp. 2586–2595 (2025)
42. Shi, D., Zhong, Y., Cao, Q., et al.: TriDet: Temporal action detection with relative boundary modeling. In: CVPR. pp. 18857–18866 (2023)
43. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: CVPR. pp. 6479–6488 (2018)
44. Tang, T.N., Kim, K., Sohn, K.: TemporalMaxer: Maximize temporal context with only max pooling for temporal action localization. arXiv preprint arXiv:2303.09055 (2023)
45. Tian, Y., Pang, G., Chen, Y., et al.: Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In: ICCV. pp. 4975–4986 (2021)
46. Tran, D., Bourdev, L., Fergus, R., et al.: Learning spatiotemporal features with 3D convolutional networks. In: ICCV. pp. 4489–4497 (2015)
47. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. NeurIPS **30** (2017)
48. Wang, L., Huang, B., Zhao, Z., et al.: VideoMAE v2: Scaling video masked autoencoders with dual masking. In: CVPR. pp. 14549–14560 (2023)
49. Wang, M., Xing, J., Jiang, B., et al.: A multimodal, multi-task adapting framework for video action recognition. In: AAAI. vol. 38, pp. 5517–5525 (2024)
50. Wang, S., Li, B.Z., Khabsa, M., Fang, H., Ma, H.: Linformer: Self-attention with linear complexity. arXiv preprint arXiv:2006.04768 (2020)
51. Wang, Y., Li, K., Li, X., et al.: InternVideo2: Scaling video foundation models for multimodal video understanding. arXiv preprint arXiv:2403.15377 (2024)
52. Wu, P., Liu, J., Shi, Y., et al.: Not only look, but also listen: Learning multimodal violence detection under weak supervision. In: ECCV. pp. 322–339 (2020)
53. Wu, P., Zhou, X., Pang, G., et al.: VadCLIP: Adapting vision-language models for weakly supervised video anomaly detection. In: AAAI. vol. 38, pp. 6074–6082 (2024)
54. Wu, P., Zhou, X., Pang, G., et al.: Weakly supervised video anomaly detection and localization with spatio-temporal prompts. In: ACM MM. pp. 9301–9310 (2024)
55. Xarles, A., Escalera, S., Moeslund, T.B., Clapés, A.: ASTRA: An action spotting transformer for soccer videos. In: Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports. pp. 93–102 (2023)
56. Xu, M., Zhao, C., Rojas, D.S., et al.: G-TAD: Sub-graph localization for temporal action detection. In: CVPR. pp. 10156–10165 (2020)
57. Yan, S., Smith, J.S., Lu, W., Zhang, B.: Hierarchical multi-scale attention networks for action recognition. Signal Processing: Image Communication **61**, 73–84 (2018)
58. Yang, L., Zheng, Z., Han, Y., et al.: DyFADet: Dynamic feature aggregation for temporal action detection. In: ECCV. pp. 305–322 (2024)
59. Yang, Z., Liu, J., Wu, P.: Text prompt with normality guidance for weakly supervised video anomaly detection. In: CVPR. pp. 18899–18908 (2024)
60. Ye, M., Liu, W., He, P.: VERA: Explainable video anomaly detection via verbalized learning of vision-language models. In: CVPR. pp. 8679–8688 (2025)
61. Yuan, J., Gao, H., Dai, D., et al.: Native sparse attention: Hardware-aligned and natively trainable sparse attention. In: ACL. pp. 4644–4664 (2025)
62. Zanella, L., Liberatori, B., Menapace, W., et al.: Delving into CLIP latent space for video anomaly recognition. Computer Vision and Image Understanding **249**, 104163 (2024)
63. Zanella, L., Menapace, W., Mancini, M., et al.: Harnessing large language models for training-free video anomaly detection. In: CVPR. pp. 18527–18536 (2024)
64. Zhang, C.L., Wu, J., Li, Y.: ActionFormer: Localizing moments of actions with transformers. In: ECCV. pp. 492–510 (2022)
65. Zheng, Z., Wang, P., Liu, W., et al.: Distance-IoU loss: Faster and better learning for bounding box regression. In: AAAI. vol. 34, pp. 12993–13000 (2020)
66. Zhong, J.X., Li, N., Kong, W., et al.: Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In: CVPR. pp. 1237–1246 (2019)
67. Zhou, H., Yu, J., Yang, W.: Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In: AAAI. vol. 37, pp. 3769–3777 (2023)
68. Zhu, B., Lin, B., Ning, M., et al.: LanguageBind: Extending video-language pretraining to n-modality by language-based semantic alignment. arXiv preprint arXiv:2310.01852 (2023)