LOCAL PATTERNS GENERALIZE BETTER FOR NOVEL ANOMALIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Video anomaly detection (VAD) aims to identify novel actions or events which are unseen during training. Existing mainstream VAD techniques typically focus on the global patterns with redundant details and struggle to generalize to unseen samples. In this paper, we propose a framework that identifies the local patterns which generalize to novel samples and models the dynamics of local patterns. The capability of extracting spatial local patterns is achieved through a two-stage process involving image-text alignment and cross-modality attention. Generalizable representations are built by focusing on semantically relevant components which can be recombined to capture the essence of novel anomalies, reducing unnecessary visual data variances. To enhance local patterns with temporal clues, we propose a State Machine Module (SMM) that utilizes earlier high-resolution textual tokens to guide the generation of precise captions for subsequent low-resolution observations. Furthermore, temporal motion estimation complements spatial local patterns to detect anomalies characterized by novel spatial distributions or distinctive dynamics. Extensive experiments on popular benchmark datasets demonstrate the achievement of state-of-the-art performance. Code is available at `https://anonymous.4open.science/r/Local-Patterns-Generalize-Better-1E30/`.

1 INTRODUCTION

Video anomaly detection (VAD) is the task of localizing from videos the events that deviate from regular patterns, such as violence, accidents and other unexpected events. Nowadays, numerous platforms such as CCTVs and UAVs play an increasingly important role in surveillance. However, given the vast volume of video data and the low probability of anomalies, it is impractical for humans to manually detect these events. Additionally, visual data variances and domain differences between normal and anomalous events hinder the effectiveness of detection methods. As a result, VAD has become a significant research topic in weakly supervised or unsupervised learning Gong et al. (2019); Shi et al. (2023b); Chalapathy et al. (2017); Lu et al. (2020); Pang et al. (2020); Lv et al. (2021); Georgescu et al. (2021a); Zaheer et al. (2020b); Ristea et al. (2021); Acsintoae et al. (2021).

Existing main-stream works Li et al. (2022c); Luo et al. (2021a); Georgescu et al. (2021a) for VAD are divided into four categories. The first category of methods detects anomalies by leveraging distinctive spatial and temporal features. These methods include prediction-based ones Luo et al. (2021a); Lv et al. (2021); Lu et al. (2020); Park et al. (2020) and reconstruction-based ones Yang et al. (2023b); Lv et al. (2023); Chang et al. (2020); Liu et al. (2021). To enhance representational capacity, some methods combine multi-grained spatio-temporal representations Zhang et al. (2024) for better discrimination, or integrate various features Georgescu et al. (2021a); Cho et al. (2023) to better align with unseen samples Liu et al. (2022b). The second category involves using Multiple Instance Learning (MIL) to iteratively identify useful data segments and fine-tune models for anomaly detection Cho et al. (2023); Wang et al. (2022a); Li et al. (2022a); Zhu et al. (2022); Liu et al. (2023c). For instance, dynamic clustering techniques adapt model representations to real-time observations Wu et al. (2022); Yang et al. (2022). Prompt-enhanced MIL Chen et al. (2024) integrates semantic priors with visual features for improved modeling of anomalies. However, the generalization ability is still insufficient because background noises lead to inconsistent representations over visual data variances, as is shown by Fig. 1. The third category Liu et al. (2023c) focuses on generating realistic anomalies to refine the decision boundary between normal and ab-

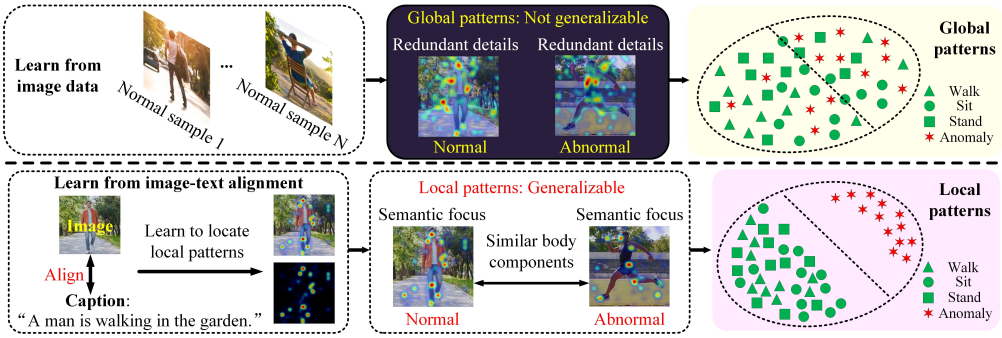

Figure 1: Top: Existing methods rely on global patterns with redundant details, which are inconsistent across visual data variations, limiting their generalization to novel samples. As a result, normal and abnormal samples are poorly distinguished. Bottom: Our method focuses on local patterns that capture semantically meaningful features such as body joints and are consistent across domains and generalize well. The spatial distributions of these local patterns highlight divergences.

normal samples. Prompt based approaches Wu et al. (2024a) have also been proposed for generating pesudo anomalies. However, the generated anomalies are based on prior assumptions which cannot cover diverse and unexpected anomalous samples in real-world cases. The fourth category Zanella et al. (2024) leverages the visual-language knowledge and reasoning capabilities Yang et al. (2024a) from large models to generate textual descriptions or produce pseudo labels for self-training Yang et al. (2024b), thereby improving the discrimination of abnormal events Micorek et al. (2024).

To generalize model representations to novel anomalies, we propose a two-stage framework for identifying local patterns. In Stage 1, image-text alignment is used to locate text-informative local patterns that are consistent across visual data variances. Stage 2 further refines the local patterns using cross-modality attention, resulting in more compact local patterns. Finally, spatial local patterns are augmented with temporal clues to better determine anomalies.

In sum, the proposed framework is composed of an Image-Text Alignment Module (ITAM) and a Cross-Modality Attention Module (CMAM) for identifying local patterns in two stages. ITAM selects text-informative regions, converting high-dimensional visual data into efficient image tokens. The tokens are converted by Temporal Sentence Generation Module (TSGM) into texts, which CMAM uses to refine the selection of image tokens as local patterns. Temporal clues enhance local patterns in two ways. TSGM generates the sentences for cross-modality attention by considering multi-moment contexts, while temporal motion estimation enriches spatial local patterns with temporal dynamics. The effectiveness is validated on multiple benchmarks, including ShanghaiTech, Ubnormal and so on. The contributions can be highlighted as follows:

- This paper proposes a novel two-stage approach to identify the local patterns that are consistent across visual data variances and generalize to novel abnormal samples. The first stage uses image-text alignment to identify semantically meaningful components, facilitating generalizable representations. Cross-modality attention further refines the components, yielding both the benefits of texts in generalization and the advantages of visual features in representing details.

- Temporal solutions are used to enhance spatial local patterns. Firstly, temporal sentence generation integrates the contexts from different moments to produce coherent descriptions of events. Additionally, temporal motion estimation complements local patterns by modeling dynamics.

- State-of-the-art performance is achieved with the proposed framework on multiple benchmarks.

2 Related Works

2.1 Unsupervised Video Anomaly Detection

Due to the unbalanced nature of surveillance videos, most training datasets are without anomaly annotations because it is expensive to label Li et al. (2022b); Liu et al. (2023b); Deng et al. (2023).

Reconstruction-based approaches Astrid et al. (2024); Yang et al. (2023b); Fang et al. (2020); Li et al. (2020a); Gong et al. (2019); Asad et al. (2021); Abati et al. (2019); Sabokrou et al. (2018) produce increased error when encountering irregular features Ramachandra et al. (2020); Madan et al. (2023); Yu et al. (2023) that do not reside in training data. For instance, the method Zaheer et al. (2022a) learns not to reconstruct anomalies. Gong et al. (2019); Gao et al. (2022) augment encoders to improve the sensitivity of reconstruction error to anomalies. Madan et al. (2021); Chang et al. (2020); Singh et al. (2023); Yu et al. (2022b); Shi et al. (2023a) integrate multi-modal features Ding et al. (2021) while Huang et al. (2022) integrates a probabilistic decision model. Zaheer et al. (2022b) assesses the quality of reconstruction to improve stability. Prediction-based methods Luo et al. (2021b); Morais et al. (2019); Luo et al. (2021a); Liu et al. (2018); Nguyen & Meunier (2019); Zeng et al. (2021) evaluate the divergence in normal and abnormal temporal dependencies, leveraging latent spaces Zhang et al. (2020) or hybrid attention Zhang et al. (2022b).

To better distinguish anomalies, Lv et al. (2021); Lu et al. (2020); Liu et al. (2021); Park et al. (2020); Li et al. (2021a) combine prediction with reconstruction. Sato et al. (2023); Wu et al. (2023); Luo et al. (2019) study the distribution over normal samples and propose novel features Arad & Werman (2023). Similarly, Yan et al. (2023) proposes denoising diffusion modules. Flaborea et al. (2023) exploits the enhanced mode coverage of diffusive probabilistic models. To improve representation capacities, Chang et al. (2021); Fan et al. (2024) propose snippet-level attention. Liu et al. (2023a); Yu et al. (2022a); Purwanto et al. (2021) introduce pyramid deformation and CRFs to learn spatio-temporal dependencies Bertasius et al. (2021); Cho et al. (2022). Wang et al. (2021) combines multi-scale features to enhance prediction. Stergiou et al. (2024) combines interpolation with extrapolation for prediction. Wang et al. (2022b) proposes a self-supervised scheme with discriminative DNNs. We propose generalizable local patterns to better represent unseen samples.

2.2 Weakly Supervised Anomaly Detection

Multi-instance learning (MIL) takes videos as bags and snippets as instances, transforming video-level labels to instance-level Feng et al. (2021). The methods iteratively locate abnormal segments and fine-tune models using anomalous segments which are dissimilar to normal ones Zhang et al. (2023a). To collect abnormal segments, inter-sample distances are evaluated Lu et al. (2022); Ionescu et al. (2019) based on spatio-temporal similarities Dhiman & Vishwakarma (2020); Lv et al. (2023); Chang et al. (2020); Markovitz et al. (2020). Li et al. (2021b) proposes a probabilistic framework. Sun et al. (2020); Li et al. (2020b) build graphical representations and integrated collective properties in measuring similarities. Sapkota & Yu (2022) performs dynamic non-parametric clustering. To improve robustness, Zhang et al. (2023b) proposes to interpret the vulnerability of MIL. Wu & Liu (2021) introduces causal relations to enhance MIL Tian et al. (2021). Yang et al. (2023a) proposes binary network augmentation strategy. Differently, we propose generalizable representations which facilitate the measurement of similarities between seen and unseen events.

2.3 Methods with Data Augmentation

To generate pseudo abnormal samples in fine-tuning, Liu et al. (2023c); Lin et al. (2022); Kim et al. (2022); Liu et al. (2022a); Astrid et al. (2021) propose pseudo abnormal snippet synthesizers which are trained on normal samples Yu et al. (2021). Zaheer et al. (2020a) employs a generator which was not fully trained to create abnormal samples. Chen et al. (2022) generates class balanced training data with a conditional GAN. Lim et al. (2018) focuses on infrequent normal samples during generation, harnessing novel sampling strategies. Besides frame-level analysis Zaheer et al. (2020b), object-level approaches Sun & Gong (2023); Ionescu et al. (2019); Luo et al. (2021a) provide fine-grained analysis. Acsintoae et al. (2022) introduces a new dataset with diverse anomalies. However, the lack in real-world modes in generated data highlights the necessity for generalizable patterns.

2.4 Methods Exploring the Representation of Unseen Categories

To adapt model representations and work under changing anomalies, meta learning-based methods Lu et al. (2020); Park et al. (2020), transfer-learning based approaches Doshi & Yilmaz (2020); Perini et al. (2022), continual learning Doshi & Yilmaz (2020) and self-supervised approaches Pang et al. (2020); Degardin & Proença (2021) introduce adaptable feature representations. Attention-based methods Sultani et al. (2018); Guo et al. (2023); Li et al. (2021c); Luo et al. (2017) attend

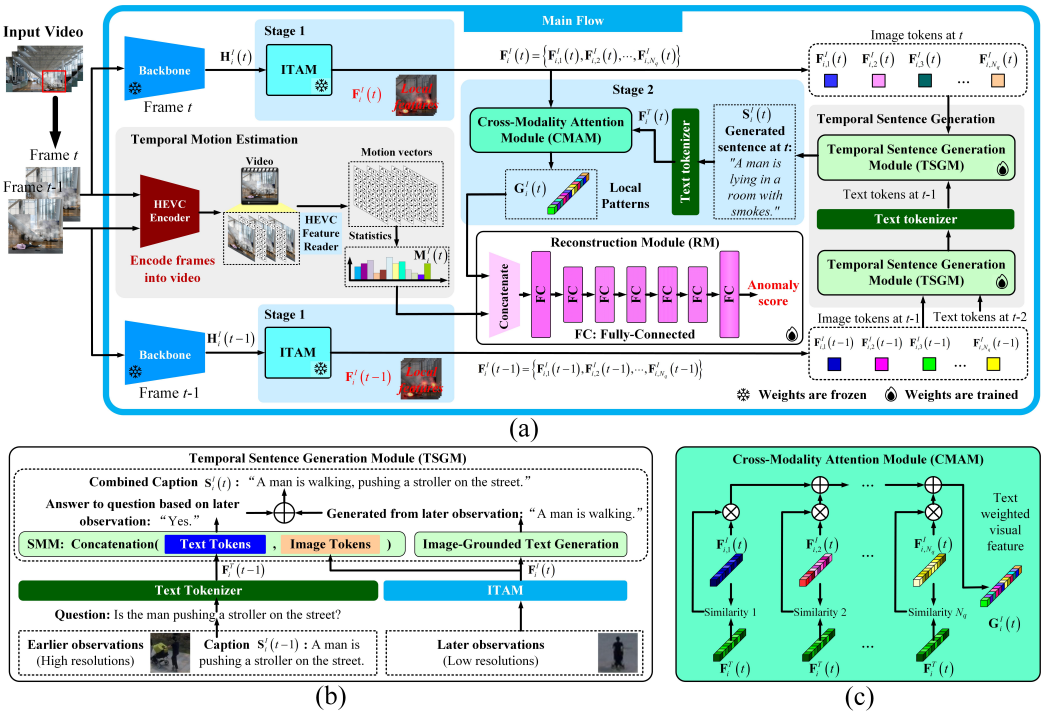

Figure 2: Structure of the method. (a) Main flow: Visual features are extracted using backbone and Image-Text Alignment Module (ITAM) which identifies caption-informative local features as image tokens. Conditioned on image tokens, the Temporal Sentence Generation Module (TSGM) generates sentences which are then combined with image tokens using the Cross-Modality Attention Module (CMAM) to highlight key patterns. HEVC Encoder estimates inter-frame motion through video compression. Spatial local patterns and motion are jointly analyzed to detect anomalies through reconstruction. (b) TSGM uses the State Machine Module (SMM) for sentence generation based on image tokens and earlier sentences. (c) CMAM is implemented based on image-text similarity.

to domain-invariant features in addressing unseen samples. To better align with anomaly detection, Georgescu et al. (2021a) integrates multiple sub-tasks. Zhou et al. (2023a) introduces hierarchical graphs for representing videos and maximizing inter-class margins. Differently, our approach locates the text-informative local patterns which generalize to unseen events.

2.5 Prompting Methods

Prompt-based approaches have been widely used in anomaly detection Du et al. (2022); Liu et al. (2023c); Sato et al. (2023). For instance, Zhou et al. (2023b) learns object-agnostic text prompts for generalized abnormality recognition. Yang et al. (2024a) proposes rule-based reasoning to achieve few-normal-shot prompting. Unlike approaches that use direct prompts, we explore local patterns which bridge the gap between images and texts in Visual-Language Models (VLMs).

3 Methodology

To represent unexpected anomalies using generalizable representations, we establish a framework capable of identifying caption-informative local patterns. The framework uses ITAM and CMAM to localize spatial local patterns in two stages, as is shown in Fig. 2(a). To augment local patterns with temporal clues, temporal sentence generation and temporal motion estimation are investigated. Firstly, TSGM models the dependencies between earlier text tokens and later image tokens, enhancing the input sentence for CMAM, as is shown in Fig. 2(b). Then inter-frame motion vectors are obtained from video compression. Finally, spatial and temporal clues are combined in the Reconstruction Module (RM) to detect anomalies. In the following parts we will discuss each module.

3.1 CROPPING OF IMAGE REGIONS

Due to the wide field of view in some frames containing numerous objects, it is difficult even for GPT-4 Achiam et al. (2023) to focus on all objects together. As a result, local regions are cropped as the first step in our pipeline. We have experimented with both YOLOv7 Wang et al. (2023) and Qwen-7B Bai et al. (2023) for cropping bounding box regions based on prompts. Specifically, "How many people are there?" and "The bounding box of the i-th object" are sequentially provided to Qwen-7B which returns corresponding boxes. The comparisons will be included in Appendix D.

3.2 STAGE 1 FOR IDENTIFYING SPATIAL LOCAL PATTERNS

This stage identifies features in cropped image regions that align with texts. The texts describe generic movement attributes (e.g., "A man is walking with swinging arms and legs"). When encountering an unseen action, such as running, the model can recombine known components like arms and legs to generate descriptive language that captures the essence of the action without explicitly naming it. As illustrated in Fig. 1, heatmaps indicate the attention on local components, highlighting similar semantic regions for shared attributes. More visualizations are in Fig. 4.

To identify the local patterns that align with texts, the frozen Image Encoder Li et al. (2023) and the image transformer of Q-Former in BLIP-2 Li et al. (2023) are employed as backbone and ITAM, respectively. The backbone outputs $\mathbf{H}_i^I(t) \in \mathbb{R}^{S_d \times V_d}$, Q-Former has an image transformer and a text transformer for aligning features from both modalities, $\mathbf{F}_i^I(t) \in \mathbb{R}^{N_q \times H_d}$ is the image transformer's output with N_q image tokens $\mathbf{F}_{i,1}^I(t), ..., \mathbf{F}_{i,N_q}^I(t)$ which inform about the captions of image region i and remain consistent over visual data variances, as will be shown by the heatmaps in Fig. 4. Detailed structures are in Appendix C. Algorithm 1 shows the workflow of Stage 1 and Stage 2.

Algorithm 1 Two-Stage Process for Identifying Spatial Local Patterns

Input: Input image, Backbone, ITAM, CMAM, TSGM and Text tokenizer
Output: Cross-modal embedding $\mathbf{G}_i^I(t)$ representing spatial local patterns
Stage 1: Image Token Extraction
    Use the backbone to extract feature maps $\mathbf{H}_i^I(t)$
    Feed $\mathbf{H}_i^I(t)$ into ITAM to obtain image tokens $\{\mathbf{F}_{i,j}^I(t)\}_{j=1}^{N_q}$ which align with texts
Stage 2: Cross-Modality Attention
    Feed image tokens $\{\mathbf{F}_{i,j}^I(t)\}_{j=1}^{N_q}$ into TSGM and obtain text tokens $\mathbf{F}_i^T(t)$
    CMAM weightedly sums $\{\mathbf{F}_{i,j}^I(t)\}_{j=1}^{N_q}$ according to their similarity with $\mathbf{F}_i^T(t)$
Return: Return weighted sum $\mathbf{G}_i^I(t)$, representing the cross-modal features

3.3 STAGE 2 FOR IDENTIFYING LOCAL PATTERNS

This stage further highlights local patterns by generating a sentence conditioned on image tokens and summing them based on their similarities to the generated sentence. Using the image tokens $\mathbf{F}_{i,1}^I(t), ..., \mathbf{F}_{i,N_q}^I(t)$ from Stage 1, TSGM generates a sentence for image region i, as is shown in Fig. 2(a). TSGM utilizes SMM for inter-frame caption augmentation and a frozen Q-Former Li et al. (2023) for image-grounded text generation. SMM determines whether previous events still reside in current frame while Q-Former captions current frame. The outputs from SMM and Q-Former are combined to form the augmented sentence $\mathbf{S}_i^I(t)$. Even with incomplete observations at t, $\mathbf{S}_i^I(t)$ can recognize previously occurring events from current frame as long as the events still reside.

The embedding $\mathbf{F}_i^T(t) \in \mathbb{R}^{S_l \times H_d}$ of $\mathbf{S}_i^I(t)$, where $S_l = 32$ denotes the maximum number of tokens in one sentence, is provided to CMAM. CMAM uses the first element in $\mathbf{F}_i^T(t)$ as query and the image tokens as keys and values for attention operations, as is illustrated in Fig. 2(c) and Eq. (1):

$$\mathbf{G}_i^I(t) = (\mathbf{F}_i^T(t)[0]\mathbf{F}_i^I(t)^\top)\mathbf{F}_i^I(t), \mathbf{G}_i^I(t) \in \mathbb{R}^{H_d} \tag{1}$$

$\mathbf{F}_i^T(t)$ is obtained by the text transformer in Q-Former Li et al. (2023) with first element $\mathbf{F}_i^T(t)[0]$ representing the whole sentence. Eq. (1) weightedly sums image tokens according to their cosine

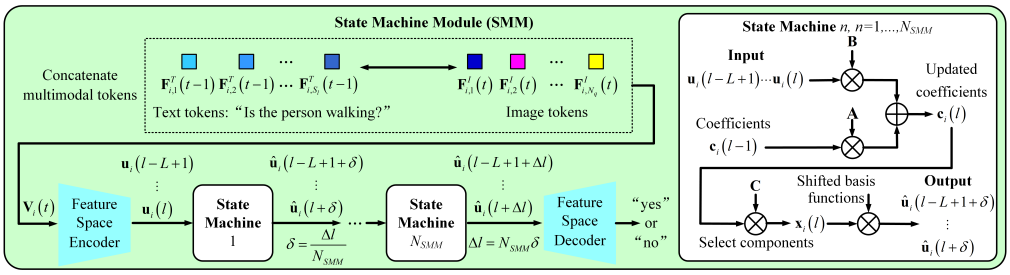

Figure 3: Structure of SMM, which stacks N_{SMM} state machines, each predicting δ ahead.

similarity to $\mathbf{F}_i^T(t)$. In this way, the advantage of image tokens in characterizing visual details and the benefits of textual features in generalizing over visual data variances are both achieved. Ablation studies will compare the performance of $\mathbf{G}_i^I(t)$ against features from single modalities.

3.4 Temporal Sentence Generation in Stage 2

In Stage 2, the generation of captions is influenced by visual data variances such as low resolutions. As is shown in Fig. 2(b), the module Li et al. (2023) for image-grounded text generation only provides a coarse caption "A man is walking" on later low-resolution observations. It is not as precise as earlier caption "A man is pushing a stroller on the street" even if they actually describe the same event. Therefore, SMM in TSGM determines whether earlier high-resolution events are represented by later image tokens. It captures inter-frame dependencies and refines sentence coherence. SMM uses earlier text tokens to generate precise captions for low-resolution observations. The objects in consecutive frames are associated using intersection over union similarity between bounding boxes.

Specifically, SMM augments image tokens $\mathbf{F}_i^I(t) = \{\mathbf{F}_{i,1}^I(t), ..., \mathbf{F}_{i,N_q}^I(t)\}$ with earlier captions $\mathbf{S}_i^I(t-1)$ which are firstly converted from declarative sentences to interrogative sentences. For instance, "The man is pushing a stroller." is changed to "Is the man pushing a stroller?" whose text tokens are $\mathbf{F}_i^T(t-1) = \{\mathbf{F}_{i,1}^T(t-1), ..., \mathbf{F}_{i,S_l}^T(t-1)\}$. Details of this conversion will be shown in Appendix G Hardeniya et al. (2016) . As is shown in Fig. 2(b), SMM combines $\mathbf{F}_i^T(t-1)$ with $\mathbf{F}_i^I(t)$ as input. The state machines in SMM evolve across the dimension of input sequence $\mathbf{V}_i(t) = [\mathbf{F}_{i,1}^T(t-1); ...; \mathbf{F}_{i,S_l}^T(t-1); \mathbf{F}_{i,1}^I(t); ...; \mathbf{F}_{i,N_q}^I(t)]^\top \in \mathbb{R}^{H_d \times (S_l+N_q)}$ where $L = S_l + N_q$ is the sequence length and each token has dimension H_d. SMM predicts a binary decision ("yes" or "no") based on the sequence, determining whether the event in $\mathbf{S}_i^I(t-1)$ is still present in $\mathbf{F}_i^I(t)$.

$\mathbf{V}_i(t)$ is deemed as the combination of H_d 1-dimensional signals each with length L. The dependencies in sequences are represented using O length-L Legendre polynomials Arfken et al. (2011) $[g_o(1), ..., g_o(L)], o \in [1, O]$, as will be shown in Fig. 6 of Appendix A. The input tensor $\mathbf{V}_i(t)$ is approximated by the weighted sums of the O fixed polynomials. For simplicity, index t is omitted in the following parts which conduct analysis along the column dimension of input tensor at any moment t. The Feature Space Encoder produces $\mathbf{U}_i(t) = [\mathbf{u}_i(l-L+1); ...; \mathbf{u}_i(l)]^\top \in \mathbb{R}^{O \times L}$, where $\mathbf{u}_i(l') = [u_{i,1}(l'), ..., u_{i,O}(l')]$ for $l' \in (l-L, l]$. Here, l varies along the column dimension of $\mathbf{V}_i(t)$ and $\mathbf{U}_i(t)$, $(l-L, l]$ is the window of columns which are encoded by $\mathbf{c}_i(l)$ together.

To better model multi-modal sequence of H_d-dimensional signals, N_{SMM} state machines are stacked in SMM, each predicting $\Delta l / N_{SMM}$ ahead, as is shown in Fig. 3. The advantages will be shown in ablation studies. Eq. (2) shows the representation of $\mathbf{U}_i(t)$ with basis functions:

$$u_{i,o}(l') = c_{i,o}(l) g_o(l' - l + L), o \in [1, O], l' \in (l-L, l] \tag{2}$$

In SMM, a state vector $\mathbf{c}_i(l) = [c_{i,1}(l); ...; c_{i,O}(l)]^\top$ with O weights encoding the dependencies between texts and visual tokens in $\mathbf{V}_i(t)$, the dependencies are decomposed onto weighted Legendre basis functions. State vector evolution informs about the prediction ("yes" or "no").

$$\mathbf{c}_i(l+1) = \mathbf{A}\mathbf{c}_i(l) + \mathbf{B} \sum\nolimits_{o=1}^{O} u_{i,o}(l+1) \tag{3}$$

where $\mathbf{A} = \mathbf{A}(O, L) \in \mathbb{R}^{O\times O}$ and $\mathbf{B} = \mathbf{B}(O, L) \in \mathbb{R}^{O\times 1}$ are derived from Legendre polynomials Gu et al. (2020). As O grows, more diversified basis functions can represent more complex dependencies. Assume that $\mathbf{c}_i(l)$ encodes $\mathbf{u}_i(l-L+1), ..., \mathbf{u}_i(l)$ based on which $\mathbf{u}_i(l+1)$ is predicted. $\mathbf{u}_i(l+1)$ denotes "yes" or "no". $\mathbf{c}_i(l+1)$ encodes $\mathbf{u}_i(l-L+2), ..., \mathbf{u}_i(l+1)$. Eq. (3) will be derived in Appendix A. In SMM shown by Fig. 3, the transformation $\mathbf{x}_i(l) = \mathbf{C}\mathbf{c}_i(l)$, where $\mathbf{C} \in \mathbb{R}^{O\times O}$ is learnable, highlights important components, Eq. (3) is transformed to

$$\mathbf{x}_i(l) = \mathbf{C}\mathbf{A}^{L-1}\mathbf{B}\sum\nolimits_{o=1}^{O} u_{i,o}(l-L+1) + ... + \mathbf{C}\mathbf{B}\sum\nolimits_{o=1}^{O} u_{i,o}(l) \tag{4}$$

Finally, the elements of $\mathbf{x}_i(l)$ are multiplied with shifted basis functions $[g_o(1+\delta), ..., g_o(L+\delta)], o \in [1, O], \Delta l = 1, \delta = \Delta l/N_{SMM}$, producing shifted weighted basis functions:

$$\hat{u}_{i,o}(l'+\delta) = x_{i,o}(l)g_o(l'-l+L+\delta), o \in [1, O], l' \in (l-L, l] \tag{5}$$

The Feature Space Decoder projects $\hat{\mathbf{u}}_i(l+\delta) = [\hat{u}_{i,1}(l+\delta), ..., \hat{u}_{i,O}(l+\delta)]$ onto a prediction ("yes" or "no"), as is shown by Fig. 3. Cross entropy loss is employed. In each batch, B_s images correspond to B_s declarative sentences which are converted into B_s questions. The tokens of each image are concatenated with those of each corresponding question before feeding into SMM.

$$L_{SMM} = -\sum_{i=0}^{B_s-1}\sum_{j=0}^{B_s-1} y_{i,j} log(\frac{Sim(P(i,j), Emb("yes"))}{Sim(P(i,j), Emb("yes")) + Sim(P(i,j), Emb("no"))}) \tag{6}$$

where ground truth $y_{i,j}$ takes 1 when the Qwen-Chat model Bai et al. (2023) receives question j together with image i and returns "yes", else $y_{i,j}$ takes 0. $Sim(P(i,j), Emb("yes"))$ is the cosine similarity between the embedding $P(i,j)$ of SMM's output and the embedding of "yes".

3.5 Temporal Motion Estimation and Spatio-Temporal Anomaly Detection

To enhance the spatial local patterns obtained from Stage 2, this paper proposes to encode frames into H.265 (HEVC) videos using FFmpeg Zeng et al. (2016). As is illustrated in Fig. 2(a), motion vectors from encoded videos are extracted, each motion vector is associated with a 8×8 macroblock. The orientation of each motion vector is computed as $atan2(y, x)$ and quantized into $D_m = 8$ equispaced bins, x and y are the horizontal and vertical components. The average magnitudes of motion vectors in these bins produce a D_m-dimensional histogram $\mathbf{M}_i^I(t)$ representing region i.

To detect anomalies with anomalous local patterns or irregular dynamics, the Reconstruction Module (RM) with 7 fully-connected layers is trained on normal spatial and temporal data. As is shown in Fig. 2(a), the first layer takes in the concatenation of local patterns $\mathbf{G}_i^I(t)$ and dynamics $\mathbf{M}_i^I(t)$, it maps $H_d + D_m$ input channels to D_h output channels while the last layer maps D_h input channels to $H_d + D_m$ output channels. The 5 hidden layers have D_h input channels and D_h output channels. The reconstructions of spatial and temporal features are conducted together, facilitating the reconstruction of each one to depend on the other. Reconstruction error determines anomaly scores.

4 Experiments and Results

This section compares the proposed method with state-of-the-art ones and presents ablation studies.

4.1 Experimental Setup

Datasets Experiments are conducted on seven datasets. The training sets of ShanghaiTech, Avenue and UCSD Ped2 contain only normal events and anomalies reside in test data. (1) **ShanghaiTech** dataset Liu et al. (2018) includes 330 training videos and 107 test videos. Among the two versions of ShanghaiTech dataset Liu et al. (2018) and Zhong et al. (2019); Li et al. (2022a); Zanella et al. (2023), the latter includes abnormal behaviors in both training set and test set. As our approach is unsupervised, we use the first version. (2) **CUHK Avenue** dataset Lu et al. (2013) involves 16

Table 1: Performance (AUC, %) on the benchmarks. ST, Ave, UB, Ped2 and NWPU represent ShanghaiTech, CUHK Avenue, Ubnormal, UCSD Ped2 and NWPU Campus, respectively. Macro-AUC and micro-AUC Reiss & Hoshen (2022) are evaluated.

Algorithm	Year	ST	Ave	UB	Ped2	NWPU
Georgescu et al. (2021b)	2021	89.3 / 82.7	92.3 / 90.4	- / 61.3	99.7 / 98.7	-
Acsintoae et al. (2021)	2021	90.5 / -	93.2 / -	-	-	-
Cai et al. (2021)	2021	- / 73.7	- / 86.6	-	- / 96.6	- / 64.5
Reiss & Hoshen (2022)	2022	89.6 / 85.9	96.2 / 93.3	-	**99.9 / 99.1**	-
Zhong et al. (2022)	2022	- / 74.5	- / 89.0	-	- / 98.1	-
Zhang et al. (2022a)	2022	- / 80.3	- / 80.5	-	- / 92.9	-
Lu et al. (2022)	2022	85.9 / 77.6	88.6 / 87.4	-	-	- / 62.2
Acsintoae et al. (2022)	2022	90.5 / 83.7	93.2 / 93.0	-	-	-
Liu et al. (2023c)	2023	91.4 / 85.0	93.9 / 93.6	-	-	-
Cao et al. (2023)	2023	- / 79.2	- / 86.8	-	-	- / 68.2
Hirschorn et al. (2023)	2023	- / 85.9	-	- / 79.2	-	-
Arad & Werman (2023)	2023	- / 85.9	- / 93.5	-	- / 99.1	-
Sun & Gong (2023)	2023	- / 83.4	- / 93.7	-	- / 98.1	-
Liu et al. (2023a)	2023	- / 78.8	- / 92.8	-	- / 99.7	-
Yu et al. (2022a)	2023	- / 72.6	- / 90.7	-	- / 97.2	-
Zhang et al. (2024)	2024	93.0 / 87.5	94.5 / 94.3	-	-	72.2 / 70.1
Micorek et al. (2024)	2024	91.5 / 86.7	96.1 / 94.3	85.5 / 72.8	99.9 / 99.7	-
Astrid et al. (2024)	2024	- / 71.39	- / 82.14	-	- / 94.05	-
Yang et al. (2024a)	2024	- / 85.2	- /89.7	- / 71.9	- / 97.9	-
Proposed Method	2024	**92.7 / 88.9**	**94.9 / 94.5**	**86.8 / 81.5**	99.8 / 99.1	**73.5 / 71.6**

training videos and 21 test videos. (3) **Ubnormal** dataset Acsintoae et al. (2022) is divided into a training set with 268 videos, a validation set with 64 videos, and a test set with 211 videos. (4) **NWPU Campus** dataset Cao et al. (2023) comprises 43 scenes, 28 classes of anomalies and 16 hours of video footage. (5) **UCSD Ped2** dataset Li et al. (2014) contains 16 normal training videos and 12 test videos. (6) **UCF Crime** dataset Sultani et al. (2018) includes 1610 training videos in which 800 contain only normal behaviors. The test set includes 290 videos in which 140 include anomalies. (7) **XD Violence** Wu et al. (2020) includes 4754 videos where 2349 are non-violent and 2405 are violent. There are 3954 training videos and 800 test videos where 500 are violent.

Evaluation Metrics Following previous literature Markovitz et al. (2020), Area under Curve (AUC, %) is adopted for evaluation. Differently, the accuracy on XD-Violence dataset is measured using precision-recall curve and the corresponding Average Precision (AP, %) Panariello et al. (2022).

Implementation Details To capture more contexts, bounding boxes are expanded by 50% on both sides horizontally and vertically. The benefits of box expansion will be shown in Table 4 of Appendix D. For image region i at t, the output of backbone and ITAM are $\mathbf{H}_i^I(t) \in \mathbb{R}^{S_d \times V_d}$ and $\mathbf{F}_i^I(t) \in \mathbb{R}^{N_q \times H_d}$ which satisfy $S_d = 257$, $V_d = 1408$, $N_q = 32$, $H_d = 768$. Each of the N_q image tokens has embedding size H_d. Following BLIP-2 Li et al. (2023), the backbone has "ViT-L/14" structure in Radford et al. (2021). The text tokenizer in Fig. 2(a) will be detailed in Appendix C. For sentences with fewer than S_l tokens, $\mathbf{F}_i^T(t)$ is padded with zeros. RM has $D_h = 512$ in intermediate layers.

SMM, with $N_{SMM} = 3$ state machines, is trained on the COCO-Caption dataset Lin et al. (2014). Table 3 shows the influences of N_{SMM}. The Feature Space Encoder (H_d input channels, $O = 64$ output channels) and Feature Space Decoder (O input channels, H_d output channels) are learnable fully-connected layers, the weights in $\mathbf{C} \in \mathbb{R}^{O \times O}$ are also learnable. All weights are initialized with distribution $N(0, 0.02)$. Training spans 20 epoches with initial learning rate 5×10^{-5} and decay 0.99. RM takes concatenated $\mathbf{G}_i^I(t)$ and $\mathbf{M}_i^I(t)$ as input, with ReLU activations. It is trained using Adam optimizer with learning rate 10^{-3} for 10 epoches, using MSE loss. Implementations are based on Pytorch Pytorch (2018) and a NVIDIA A100 GPU. RM is trained on benchmark videos without anomalies. The influences of RM's number of layers will be shown in Appendix F. The evaluations on operational efficiency will be detailed in Appendix H.

Table 2: Performance on UCF-Crime (micro-AUC, %) and XD-Violence (AP, %). UCF and XD represent UCF-Crime and XD-Violence, respectively.

Algorithm	UCF	XD	Algorithm	UCF	XD
Joo et al. (2023)	87.58	82.19	Wu et al. (2024a)	86.40	76.03
Chen et al. (2023)	86.98	80.11	Chen et al. (2024)	86.83	88.21
Pu et al. (2023)	86.76	85.59	Yang et al. (2024b)	87.79	83.68
Tan et al. (2024)	86.71	82.10	Wu et al. (2024b)	88.02	84.51
Zanella et al. (2024)	80.28	85.36	Proposed Method	**88.83**	**86.96**

Table 3: Ablations of components using Micro-AUC. TME and TSG are short for Temporal Motion Estimation and Temporal Sentence Generation, respectively. N_q is the number of image tokens.

Setting	Stage 1	Stage 2	N_q	TME	TSG	N_{SMM}	ST	Ave	UB	Ped2
1	×	×	32	×	×	-	71.9	85.1	72.8	79.3
2	✓	×	32	×	×	-	80.3	86.5	73.6	90.7
3	✓	✓	32	×	✓	3	86.4	88.7	79.3	96.8
4	✓	w/o image tokens	32	×	✓	3	79.2	87.8	72.1	95.4
5	✓	w/o text tokens	32	×	✓	3	80.7	89.6	74.5	95.9
6	✓	✓	32	✓	✓	3	88.9	94.5	81.5	99.1
7	✓	✓	32	✓	w/o SMM	3	87.6	93.0	80.1	98.5
8	×	×	32	✓	×	-	84.1	86.2	77.6	94.5
9	✓	✓	32	✓	✓	1	88.6	94.4	79.8	99.1
10	✓	✓	32	✓	✓	5	88.9	94.5	81.5	99.1

4.2 Comparisons with Baselines

To demonstrate the superiority, the proposed approach is compared with existing ones, including LLM-based baselines Yang et al. (2024a), for detecting anomalies. Significant improvements are observed in Table 1. Such improvements are attributed to the identification of spatial local patterns and dynamics. Results on non-human objects are shown in Table 2 with UCF-Crime and XD-Violence datasets, suggesting that local patterns can generalize to different object types.

4.3 Ablation Studies

Ablation on Stage 1 and Stage 2 In Setting 1 of Table 3, the reconstruction error of backbone features $\mathbf{H}_i^I(t)$ is used to detect anomalies. Setting 2 and Setting 3 show the utilization of Stage 1 and both stages for reconstruction, respectively. The comparison shows that Stages 1 and 2 both play crucial roles in selecting text-informative local patterns, as will be illustrated in Fig. 4.

Ablation on ITAM's Structure To demonstrate that the primary contributor to generalization is image-text alignment instead of pre-existing models, we conduct an ablation study by varying the structure and training data of ITAM. Detailed results and analysis are provided in Appendix E.

Ablation on Cross-Modality Attention Setting 4 in Table 3 replaces cross-modality feature $\mathbf{G}_i^I(t)$ in Setting 3 with textual feature $\mathbf{F}_i^T(t)$ of sentence from TSGM, using reconstruction error on $\mathbf{F}_i^T(t)$ to determine anomalies. Setting 5 discards text tokens, only using the reconstruction error on $\mathbf{F}_i^I(t)$. Fig. 5 also shows that combining visual and textual features outperforms using a single modality.

Ablation on Temporal Motion Estimation The improvement of Setting 6 over Setting 3 demonstrates that temporal dynamics complements local patterns in detecting anomalies. Setting 8 shows the performance of only using reconstruction error on dynamics $\mathbf{M}_i^I(t)$ for anomaly detection.

Ablation on the SMM in Temporal Sentence Generation in Stage 2 The comparison between Setting 6 and Setting 7 shows that if TSGM only uses Q-Former Li et al. (2023) for image-grounded text generation without SMM to incorporate previous captions, performance drops. As a result, the mixture of image tokens and text tokens from different moments contributes to more informative sentences. More ablations on SMM will be shown in Appendix B.

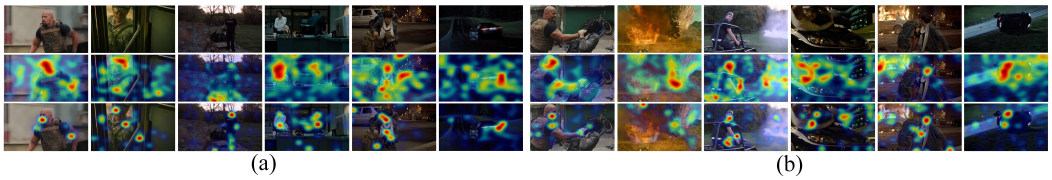

Figure 4: Heatmaps of local patterns in two stages. (a) Normal events. (b) Abnormal events. Both (a) and (b) follow the same row arrangement: the first row contains input images, the second row shows the features from ITAM, and the third row shows the local patterns selected by CMAM.

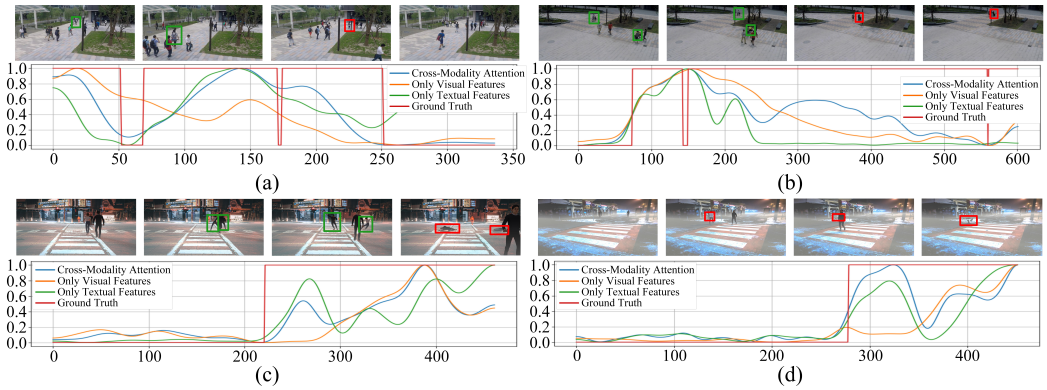

Figure 5: Anomaly scores obtained using image features, text features and combined ones. Cross-modality attention detects anomalies even under occlusion and low resolution. Green and red boxes show the anomalies detected with single modality and cross-modality attention, respectively.

Ablation on SMM's Structure SMM stacks N_{smm} state machines, each predicting a future period of $\Delta l/N_{smm}$. The stacking mechanism achieves the full prediction Δl. Setting 6, 9 and 10 in Table 3 show that $N_{smm} = 3$ outperforms $N_{smm} = 1$. The task for each state machine becomes simpler because each one focuses on short-term dependencies. Predicting a long period Δl requires capturing both short- and long-term dependencies. A single state machine struggles to handle these varying dependencies effectively, especially in our case with non-linear multi-modal dependencies.

Moreover, the ablation on the number of image tokens N_q will be involved in Appendix E.

4.4 Subjective Results on Local Patterns

Fig. 4 subjectively shows local patterns. The second rows of Fig. 4(a) and (b) highlight the patterns for $\mathbf{F}_i^I(t)$ while the third rows display those for $\mathbf{G}_i^I(t)$. The heatmaps, generated using Grad-CAM Selvaraju et al. (2017), show that local patterns span similar semantic regions across normal and abnormal events. Cross-modality attention refines these patterns to focus on semantically relevant components, enhancing generalization. More visualizations will be presented in Appendix I.

5 Discussion and Conclusion

Limitations: The limitation of our work lies in the reliance on object detectors, because the direct processing of an image with many objects using VLM can result in context being ignored. Please refer to Appendix J for more potential directions of improvement.

Conclusions: In this paper, we establish a framework for video anomaly detection by locating local patterns through image-text alignment and cross-modality attention. At the core of the framework is identifying the text-informative local patterns that generalize to novel anomalies, ensuring consistent representations across novel visual data. Additionally, temporal sentence generation and motion estimation augment cross-modality attention and complement spatial local patterns, respectively. Extensive experiments show that the framework surpasses existing state-of-the-art methods.

REFERENCES

Davide Abati, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Latent space autoregression for novelty detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 481–490. IEEE, 2019.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Ubnormal: New benchmark for supervised open-set video anomaly detection. *arXiv preprint arXiv:2111.08644*, 2021.

Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Ubnormal: New benchmark for supervised open-set video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20143–20153, 2022.

Yoav Arad and Michael Werman. Beyond the benchmark: Detecting diverse anomalies in videos. *arXiv preprint arXiv:2310.01904*, 2023.

George B Arfken, Hans J Weber, and Frank E Harris. *Mathematical methods for physicists: a comprehensive guide*. Academic press, 2011.

Mujtaba Asad, Jie Yang, Enmei Tu, Liming Chen, and Xiangjian He. Anomaly3d: Video anomaly detection based on 3d-normality clusters. *Journal of Visual Communication and Image Representation*, 75:103047, 2021.

Marcella Astrid, Muhammad Zaigham Zaheer, and Seung-Ik Lee. Synthetic temporal anomaly guided end-to-end video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 207–214, 2021.

Marcella Astrid, Muhammad Zaigham Zaheer, and Seung-Ik Lee. Constricting normal latent space for anomaly detection with normal-only training data. *arXiv preprint arXiv:2403.16270*, 2024.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, pp. 4, 2021.

Ruichu Cai, Hao Zhang, Wen Liu, Shenghua Gao, and Zhifeng Hao. Appearance-motion memory consistency network for video anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 938–946, 2021.

Congqi Cao, Yue Lu, Peng Wang, and Yanning Zhang. A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20392–20401, 2023.

Raghavendra Chalapathy, A. K. Menon, and S. Chawla. Robust, deep and inductive anomaly detection. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 36–51. Springer, 2017.

Shuning Chang, Yanchao Li, Shengmei Shen, Jiashi Feng, and Zhiying Zhou. Contrastive attention for video anomaly detection. *IEEE Transactions on Multimedia*, 24:4067–4076, 2021.

Yunpeng Chang, Zhigang Tu, Wei Xie, and Junsong Yuan. Clustering driven deep autoencoder for video anomaly detection. In *European Conference on Computer Vision*, pp. 329–345. Springer, 2020.

Junxi Chen, Liang Li, Li Su, Zheng-jun Zha, and Qingming Huang. Prompt-enhanced multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18319–18329, 2024.

Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 387–395, 2023.

Zhi Chen, Jiang Duan, Li Kang, and Guoping Qiu. Supervised anomaly detection via conditional generative adversarial network and ensemble active learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7781–7798, 2022.

MyeongAh Cho, Taeoh Kim, Woo Jin Kim, Suhwan Cho, and Sangyoun Lee. Unsupervised video anomaly detection via normalizing flows with implicit latent features. *Pattern Recognition*, 129: 108703, 2022.

MyeongAh Cho, Minjung Kim, Sangwon Hwang, Chaewon Park, Kyungjae Lee, and Sangyoun Lee. Look around for anomalies: Weakly-supervised anomaly detection via context-motion relational learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12137–12146, 2023.

Bruno Degardin and Hugo Proença. Iterative weak/self-supervised classification framework for abnormal events detection. *Pattern Recognition Letters*, 145:50–57, 2021.

Andong Deng, Taojiannan Yang, and Chen Chen. A large-scale study of spatiotemporal representation learning with a new benchmark on action recognition. pp. 20519–20531, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Chhavi Dhiman and Dinesh Kumar Vishwakarma. View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics. *IEEE Transactions on Image Processing*, 29:3835–3844, 2020.

Mingyu Ding, Zhenfang Chen, Tao Du, Ping Luo, Josh Tenenbaum, and Chuang Gan. Dynamic visual reasoning by learning differentiable physics models from video and language. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Keval Doshi and Yasin Yilmaz. Continual learning for anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 254–255, 2020.

Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14084–14093, 2022.

Yidan Fan, Yongxin Yu, Wenhuan Lu, and Yahong Han. Weakly-supervised video anomaly detection with snippet anomalous attention. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

Zhiwen Fang, Joey Tianyi Zhou, Yang Xiao, Yanan Li, and Feng Yang. Multi-encoder towards effective anomaly detection in videos. *IEEE Transactions on Multimedia*, 23:4106–4116, 2020.

Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14009–14018, 2021.

Alessandro Flaborea, Luca Collorone, Guido Maria D'Amely Di Melendugno, Stefano D'Arrigo, Bardh Prenkaj, and Fabio Galasso. Multimodal motion conditioned diffusion model for skeleton-based video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10318–10329, 2023.

Jie Gao, Bineng Zhong, and Yan Chen. Robust tracking via learning model update with unsupervised anomaly detection philosophy. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12742–12752. IEEE, 2021a.

Mariana Iuliana Georgescu, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. A background-agnostic framework with adversarial training for abnormal event detection in video. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):4505–4523, 2021b.

Dong Gong, L. Liu, V. Le, B. Saha, M.R. Mansour, S. Venkatesh, and A.V. Den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of IEEE International Conference on Computer Vision*, pp. 1705–1714. IEEE, 2019.

Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. Hippo: Recurrent memory with optimal polynomial projections. *Advances in neural information processing systems*, 33: 1474–1487, 2020.

Chongye Guo, Hongbo Wang, Yingjie Xia, and Guorui Feng. Learning appearance-motion synergy via memory-guided event prediction for video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

Nitin Hardeniya, Jacob Perkins, Deepti Chopra, Nisheeth Joshi, and Iti Mathur. *Natural language processing: python and NLTK*. Packt Publishing Ltd, 2016.

Or Hirschorn, Shai Avidan, and Shai Avidan. Normalizing flows for human pose anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13545–13554, 2023.

Xin Huang, Yutao Hu, Xiaoyan Luo, Jungong Han, Baochang Zhang, and Xianbin Cao. Boosting variational inference with margin learning for few-shot scene-adaptive anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7842–7851, 2019.

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)*, pp. 3230–3234. IEEE, 2023.

Siwon Kim, Kukjin Choi, Hyun-Soo Choi, Byunghan Lee, and Sungroh Yoon. Towards a rigorous evaluation of time-series anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7194–7201. AAAI Press, 2022.

Bo Li, Sam Leroux, and Pieter Simoens. Decoupled appearance and motion learning for efficient anomaly detection in surveillance video. *Computer Vision and Image Understanding*, 210: 103249, 2021a.

Guoqiu Li, Guanxiong Cai, Xingyu Zeng, and Rui Zhao. Scale-aware spatio-temporal relation learning for video anomaly detection. In *European Conference on Computer Vision*, pp. 333–350. Springer, 2022a.

Jing Li, Qingwang Huang, Yingjun Du, Xiantong Zhen, Shengyong Chen, and Ling Shao. Variational abnormal behavior detection with motion consistency. *IEEE Transactions on Image Processing*, 31:275–286, 2021b.

Junnan Li, Caiming Xiong, and Steven CH Hoi. Learning from noisy data with robust representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9485–9494, 2021c.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Nanjun Li, Faliang Chang, and Chunsheng Liu. Spatial-temporal cascade autoencoder for video anomaly detection in crowded scenes. *IEEE Transactions on Multimedia*, 23:203–215, 2020a.

Nanjun Li, Faliang Chang, and Chunsheng Liu. A self-trained spatial graph convolutional network for unsupervised human-related anomalous event detection in complex scenes. *IEEE Transactions on Cognitive and Developmental Systems*, 2022b.

Shuo Li, Fang Liu, and Licheng Jiao. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24. AAAI Press, 2022c.

Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):18–32, 2014.

Xuelong Li, Mulin Chen, and Qi Wang. Quantifying and detecting collective motion in crowd scenes. *IEEE Transactions on Image Processing*, 29:5571–5583, 2020b.

Swee Kiat Lim, Yi Loo, Ngoc-Trung Tran, Ngai-Man Cheung, Gemma Roig, and Yuval Elovici. Doping: Generative data augmentation for unsupervised anomaly detection with gan. In *2018 IEEE international conference on data mining (ICDM)*, pp. 1122–1127. IEEE, 2018.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Xiangru Lin, Yuyang Chen, Guanbin Li, and Yizhou Yu. A causal inference look at unsupervised video anomaly detection. In *Proceedings of the Thirty-sixth AAAI Conference on Artificial Intelligence*. AAAI Press, 2022.

Boyang Liu, Pang-Ning Tan, and Jiayu Zhou. Unsupervised anomaly detection by robust density estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 2022a.

Wen Liu, W. Luo, D. Lian, and S. Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6536–6545. IEEE, 2018.

Wenrui Liu, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Diversity-measurable anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12147–12156, 2023a.

Yang Liu, Dingkang Yang, Yan Wang, Jing Liu, Jun Liu, Azzedine Boukerche, Peng Sun, and Liang Song. Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models. *ACM Computing Surveys*, 2023b.

Yuejiang Liu, Riccardo Cadei, Jonas Schweizer, Sherwin Bahmani, and Alexandre Alahi. Towards robust and adaptive motion forecasting: A causal representation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17081–17092. IEEE, 2022b.

Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13588–13597, 2021.

Zuhao Liu, Xiao-Ming Wu, Dian Zheng, Kun-Yu Lin, and Wei-Shi Zheng. Generating anomalies for video anomaly detection with prompt-based feature mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24500–24510, 2023c.

Cewu Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2720–2727. IEEE, 2013.

Yiwei Lu, Frank Yu, Mahesh Kumar Krishna Reddy, and Yang Wang. Few-shot scene-adaptive anomaly detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pp. 125–141. Springer, 2020.

Yue Lu, Congqi Cao, Yifan Zhang, and Yanning Zhang. Learnable locality-sensitive hashing for video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 33 (2):963–976, 2022.

Weixin Luo, W. Liu, and S. Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of IEEE International Conference on Computer Vision*, pp. 341–349. IEEE, 2017.

Weixin Luo, Wen Liu, Dongze Lian, Jinhui Tang, Lixin Duan, Xi Peng, and Shenghua Gao. Video anomaly detection with sparse coding inspired deep neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):1070–1084, 2019.

Weixin Luo, Wen Liu, and Shenghua Gao. Normal graph: Spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection. *Neurocomputing*, 444:332–337, 2021a.

Weixin Luo, Wen Liu, Dongze Lian, and Shenghua Gao. Future frame prediction network for video anomaly detection. *IEEE transactions on pattern analysis and machine intelligence*, 44(11): 7505–7520, 2021b.

Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Learning normal dynamics in videos with meta prototype network. *arXiv preprint arXiv:2104.06689*, 2021.

Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. Unbiased multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8022–8031, 2023.

Neelu Madan, Arya Farkhondeh, Kamal Nasrollahi, Sergio Escalera, and Thomas B Moeslund. Temporal cues from socially unacceptable trajectories for anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2150–2158, 2021.

Neelu Madan, Nicolae-Cătălin Ristea, Radu Tudor Ionescu, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B Moeslund, and Mubarak Shah. Self-supervised masked convolutional transformer block for anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Amir Markovitz, Gilad Sharir, Itamar Friedman, Lihi Zelnik-Manor, and Shai Avidan. Graph embedded pose clustering for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10539–10547. IEEE, 2020.

Jakub Micorek, Horst Possegger, Dominik Narnhofer, Horst Bischof, and Mateusz Kozinski. Mulde: Multiscale log-density estimation via denoising score matching for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18868–18877, 2024.

Romero Morais, V. Le, T. Tran, B. Saha, M. Mansour, and S. Venkatesh. Learning regularity in skeleton trajectories for anomaly detection in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11996–12004. IEEE, 2019.

Trong Nguyen Nguyen and J. Meunier. Anomaly detection in video sequence with appearance motion correspondence. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1273–1283. IEEE, 2019.

Aniello Panariello, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Consistency-based self-supervised learning for temporal anomaly localization. In *European Conference on Computer Vision*, pp. 338–349. Springer, 2022.

Guansong Pang, Cheng Yan, Chunhua Shen, Anton van den Hengel, and Xiao Bai. Self-trained deep ordinal regression for end-to-end video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12173–12182. IEEE, 2020.

Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14372–14381. IEEE, 2020.

Lorenzo Perini, Vincent Vercruyssen, and Jesse Davis. Transferring the contamination factor between anomaly detection domains by shape similarity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 4128–4136. AAAI Press, 2022.

Yujiang Pu, Xiaoyu Wu, and Shengjin Wang. Learning prompt-enhanced context features for weakly-supervised video anomaly detection. *arXiv preprint arXiv:2306.14451*, 2023.

Didik Purwanto, Yie-Tarng Chen, and Wen-Hsien Fang. Dance with self-attention: A new look of conditional random fields on anomaly detection in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 173–183, 2021.

Automatic Differentiation In Pytorch. Pytorch, 2018.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Bharathkumar Ramachandra, Michael J Jones, and Ranga Raju Vatsavai. A survey of single-scene video anomaly detection. *IEEE transactions on pattern analysis and machine intelligence*, 44(5): 2293–2312, 2020.

Tal Reiss and Yedid Hoshen. Attribute-based representations for accurate and interpretable video anomaly detection. *arXiv preprint arXiv:2212.00789*, 2022.

Nicolae-Catalin Ristea, Neelu Madan, Radu Tudor Ionescu, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B Moeslund, and Mubarak Shah. Self-supervised predictive convolutional attentive block for anomaly detection. *arXiv preprint arXiv:2111.09099*, 2021.

Mohammad Sabokrou, M. Khalooei, M. Fathy, and E. Adeli. Adversarially learned one-class classifier for novelty detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3379–3388. IEEE, 2018.

Hitesh Sapkota and Qi Yu. Bayesian nonparametric submodular video partition for robust anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3212–3221, 2022.

Fumiaki Sato, Ryo Hachiuma, and Taiki Sekii. Prompt-guided zero-shot anomaly action recognition using pretrained deep skeleton features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6471–6480, 2023.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Chenrui Shi, Che Sun, Yuwei Wu, and Yunde Jia. Video anomaly detection via sequentially learning multiple pretext tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10330–10340, 2023a.

Haoyue Shi, Le Wang, Sanping Zhou, Gang Hua, and Wei Tang. Abnormal ratios guided multi-phase self-training for weakly-supervised video anomaly detection. *IEEE Transactions on Multimedia*, 2023b.

Ashish Singh, Michael J Jones, and Erik G Learned-Miller. Eval: Explainable video anomaly localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18717–18726, 2023.

Alexandros Stergiou, Brent De Weerdt, and Nikos Deligiannis. Holistic representation learning for multitask trajectory anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6729–6739, 2024.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6479–6488, 2018.

Che Sun, Yunde Jia, Yao Hu, and Yuwei Wu. Scene-aware context reasoning for unsupervised abnormal event detection in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 184–192, 2020.

Shengyang Sun and Xiaojin Gong. Hierarchical semantic contrast for scene-aware video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22846–22856, 2023.

Weijun Tan, Qi Yao, and Jingfeng Liu. Overlooked video classification in weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 202–210, 2024.

Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4975–4986, 2021.

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.

Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7464–7475, 2023.

Guodong Wang, Yunhong Wang, Jie Qin, Dongming Zhang, Xiuguo Bao, and Di Huang. Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. *arXiv preprint arXiv:2207.10172*, 2022a.

Siqi Wang, Yijie Zeng, Guang Yu, Zhen Cheng, Xinwang Liu, Sihang Zhou, En Zhu, Marius Kloft, Jianping Yin, and Qing Liao. E3 outlier: a self-supervised framework for unsupervised deep outlier detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):2952–2969, 2022b.

Xuanzhao Wang, Zhengping Che, Bo Jiang, Ning Xiao, Ke Yang, Jian Tang, Jieping Ye, Jingyu Wang, and Qi Qi. Robust unsupervised video anomaly detection by multipath frame prediction. *IEEE transactions on neural networks and learning systems*, 33(6):2301–2312, 2021.

Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. Self-supervised sparse representation for video anomaly detection. In *European Conference on Computer Vision*, pp. 729–745. Springer, 2022.

Peihao Wu, Wenqian Wang, Faliang Chang, Chunsheng Liu, and Bin Wang. Dss-net: Dynamic self-supervised network for video anomaly detection. *IEEE Transactions on Multimedia*, 2023.

Peng Wu and Jing Liu. Learning causal temporal relation and feature discrimination for anomaly detection. *IEEE Transactions on Image Processing*, 30:3513–3527, 2021.

Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pp. 322–339. Springer, 2020.

Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, and Yanning Zhang. Open-vocabulary video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18297–18307, 2024a.

Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6074–6082, 2024b.

Cheng Yan, Shiyu Zhang, Yang Liu, Guansong Pang, and Wenjun Wang. Feature prediction diffusion model for video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5527–5537, 2023.

Yuchen Yang, Kwonjoon Lee, Behzad Dariush, Yinzhi Cao, and Shao-Yuan Lo. Follow the rules: Reasoning for video anomaly detection with large language models. *arXiv preprint arXiv:2407.10299*, 2024a.

Zhen Yang, Yuanfang Guo, Junfu Wang, Di Huang, Xiuguo Bao, and Yunhong Wang. Towards video anomaly detection in the real world: A binarization embedded weakly-supervised network. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023a.

Zhiwei Yang, Peng Wu, Jing Liu, and Xiaotao Liu. Dynamic local aggregation network with adaptive clusterer for anomaly detection. In *European Conference on Computer Vision*, pp. 404–421. Springer, 2022.

Zhiwei Yang, Jing Liu, Zhaoyang Wu, Peng Wu, and Xiaotao Liu. Video event restoration based on keyframes for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14592–14601, 2023b.

Zhiwei Yang, Jing Liu, and Peng Wu. Text prompt with normality guidance for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18899–18908, 2024b.

Fei Yu, Mo Zhang, Hexin Dong, Sheng Hu, Bin Dong, and Li Zhang. Dast: Unsupervised domain adaptation in semantic segmentation based on discriminator attention and self-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10754–10762, 2021.

Guang Yu, Siqi Wang, Zhiping Cai, Xinwang Liu, Chuanfu Xu, and Chengkun Wu. Deep anomaly discovery from unlabeled videos via normality advantage and self-paced refinement. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 13987–13998, 2022a.

Jiashuo Yu, Jinyu Liu, Ying Cheng, Rui Feng, and Yuejie Zhang. Modality-aware contrastive instance learning with self-distillation for weakly-supervised audio-visual violence detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 6278–6287, 2022b.

Shoubin Yu, Zhongyin Zhao, Haoshu Fang, Andong Deng, Haisheng Su, Dongliang Wang, Weihao Gan, Cewu Lu, and Wei Wu. Regularity learning via explicit distribution modeling for skeletal video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

M Zaigham Zaheer, Arif Mahmood, M Haris Khan, Mattia Segu, Fisher Yu, and Seung-Ik Lee. Generative cooperative learning for unsupervised video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14744–14754, 2022a.

Muhammad Zaigham Zaheer, Jin-ha Lee, Marcella Astrid, and Seung-Ik Lee. Old is gold: Redefining the adversarially learned one-class classifier training paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14183–14193. IEEE, 2020a.

Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In *European Conference on Computer Vision*, pp. 358–376. Springer, 2020b.

Muhammad Zaigham Zaheer, Jin-Ha Lee, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. Stabilizing adversarially learned one-class novelty detection using pseudo anomalies. *IEEE Transactions on Image Processing*, 31:5963–5975, 2022b.

Luca Zanella, Benedetta Liberatori, Willi Menapace, Fabio Poiesi, Yiming Wang, and Elisa Ricci. Delving into clip latent space for video anomaly recognition. *arXiv preprint arXiv:2310.02835*, 2023.

Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing large language models for training-free video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18527–18536, 2024.

Hao Zeng, Zhiyong Zhang, and Lulin Shi. Research and implementation of video codec based on ffmpeg. In *2016 international conference on network and information systems for computers (ICNISC)*, pp. 184–188. IEEE, 2016.

Xianlin Zeng, Yalong Jiang, Wenrui Ding, Hongguang Li, Yafeng Hao, and Zifeng Qiu. A hierarchical spatio-temporal graph convolutional neural network for anomaly detection in videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

Chen Zhang, Guorong Li, Yuankai Qi, Shuhui Wang, Laiyun Qing, Qingming Huang, and Ming-Hsuan Yang. Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16271–16280, 2023a.

Menghao Zhang, Jingyu Wang, Qi Qi, Haifeng Sun, Zirui Zhuang, Pengfei Ren, Ruilong Ma, and Jianxin Liao. Multi-scale video anomaly detection by multi-grained spatio-temporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17385–17394, 2024.

Sijia Zhang, Maoguo Gong, Yu Xie, A Kai Qin, Hao Li, Yuan Gao, and Yew-Soon Ong. Influence-aware attention networks for anomaly detection in surveillance videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8):5427–5437, 2022a.

Xinfeng Zhang, Jinpeng Fang, Baoqing Yang, Shuhan Chen, and Bin Li. Hybrid attention and motion constraint for anomaly detection in crowded scenes. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022b.

Yu Zhang, Xiushan Nie, Rundong He, Meng Chen, and Yilong Yin. Normality learning in multi-space for video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(9):3694–3706, 2020.

Yu-Xuan Zhang, Hua Meng, Xue-Mei Cao, Zhengchun Zhou, Mei Yang, and Avik Ranjan Adhikary. Interpreting vulnerabilities of multi-instance learning to adversarial perturbations. *Pattern Recognition*, pp. 109725, 2023b.

Jia-Xing Zhong, Nannan Li, Weijie Kong, Thomas H Li Shan Liu, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1237–1246. IEEE, 2019.

Yuanhong Zhong, Xia Chen, Yongting Hu, Panliang Tang, and Fan Ren. Bidirectional spatio-temporal feature learning with multiscale evaluation for video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8285–8296, 2022.

Hang Zhou, Junqing Yu, and Wei Yang. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. *arXiv preprint arXiv:2302.05160*, 2023a.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. *arXiv preprint arXiv:2310.18961*, 2023b.

Yuansheng Zhu, Wentao Bao, and Qi Yu. Towards open set video anomaly detection. In *European Conference on Computer Vision*, pp. 395–412. Springer, 2022.

A DETAILS ABOUT SMM

We model the sequence $\mathbf{V}_i(t) = [\mathbf{F}^T_{i,1}(t-1); ...; \mathbf{F}^T_{i,S_l}(t-1); \mathbf{F}^I_{i,1}(t); ...; \mathbf{F}^I_{i,N_q}(t)]^\top \in \mathbb{R}^{H_d \times (S_l+N_q)}, L = S_l + N_q$ of object i consisting of both S_l columns denoting the sentence from $t-1$ and N_q columns denoting image tokens at t. Matrix $\mathbf{V}_i(t)$ is projected to $\mathbf{U}_i(t) \in \mathbb{R}^{O \times L}$ with the Feature Space Encoder which is a fully-connected layer in Fig. 3. The l'-th column in $\mathbf{U}_i(t)$ is $[u_{i,1}(l'), ..., u_{i,O}(l')]^\top, l' \in (l-L, l]$. l varies along the column dimension of $\mathbf{V}_i(t)$ and $\mathbf{U}_i(t)$, $(l-L, l]$ is the window of columns which are encoded by $\mathbf{c}_i(l)$ together. For simplicity, we ignore index t in the following parts which conduct analysis along the column dimension of input tensor at any moment t.

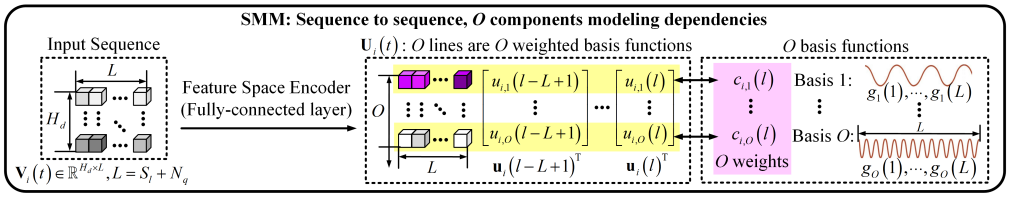

Figure 6: Basis functions in state machines for modeling sequences.

Fig. 6 shows the basis functions in state machines for modeling sequences. The O rows in $\mathbf{U}_i(t)$ are the weighted version of the O length-L Legendre polynomials Arfken et al. (2011) $[g_o(1), ..., g_o(L)], o \in [1, O]$. Specifically, the o-th row can be represented as:

$$[u_{i,o}(l-L+1), ..., u_{i,o}(l)] = c_{i,o}(l)[g_o(1), ..., g_o(L)], o \in [1, O] \tag{7}$$

In SMM, a state vector $\mathbf{c}_i(l) \in \mathbb{R}^O$ with O weights $c_{i,1}(t), ..., c_{i,O}(t)$ are the weights of polynomials and encode the dependencies among the columns in $\mathbf{V}_i(t)$.

As can be seen from Fig. 6, $\mathbf{c}_i(l_1)$ encodes the dependencies among $l_1 - L + 1, ..., l_1$ columns, $\mathbf{c}_i(l_2)$ encodes the dependencies among $l_2 - L + 1, ..., l_2$ columns. According to Gu et al. (2020), the dynamics of a 1-dimensional sequence $f_i(l)$ across a period can be represented by $\mathbf{c}_i(l) \in \mathbb{R}^O$, satisfying $[f_i(l-L+1), ..., f_i(l)] = \sum_{o=1}^{O} c_{i,o}(l)[g_o(1), ..., g_o(L)], f_i(l') \in \mathbb{R}, l' \in [l-L+1, l]$. The transitions from $\mathbf{c}_i(l_1)$ to $\mathbf{c}_i(l_2)$ facilitates the prediction in the sequence:

$$\frac{d}{dl}\mathbf{c}_i(l) = \mathbf{A}_{HiPPO}\mathbf{c}_i(l) + \mathbf{B}_{HiPPO}f_i(l) \tag{8}$$

By combining Eq. (7) with Eq. (8), we can obtain:

$$\frac{d}{dt}\mathbf{c}_i(l) = \mathbf{A}_{HiPPO}\mathbf{c}_i(l) + \mathbf{B}_{HiPPO}\sum\nolimits_{o=1}^{O} u_{i,o}(l) \tag{9}$$

The matrices $\mathbf{A}_{HiPPO}$ and $\mathbf{B}_{HiPPO}$ are defined in Gu et al. (2020) with $o, h \in [1, O]$:

$$\mathbf{A}_{HiPPO}(o,h) = \begin{cases} -\frac{(2o+1)^{0.5}(2h+1)^{0.5}}{L} & \text{if } o > h, \\ 0 & \text{if } o < h, \\ -\frac{o+1}{L} & \text{if } o = h. \end{cases} \tag{10}$$

$$\mathbf{B}_{HiPPO}(o) = -\frac{(2o+1)^{0.5}}{L} \tag{11}$$

To discretize Eq. (9), we obtain

$$\lim_{\Delta \to 0} \frac{\mathbf{c}_i(l+\Delta) - \mathbf{c}_i(l)}{\Delta} = \lim_{\Delta \to 0} (\frac{\mathbf{A}_{HiPPO}\mathbf{c}_i(l) + \mathbf{B}_{HiPPO} \sum_{o=1}^{O} u_{i,o}(l)}{2} + \frac{\mathbf{A}_{HiPPO}\mathbf{c}_i(l+\Delta) + \mathbf{B}_{HiPPO} \sum_{o=1}^{O} u_{i,o}(l+\Delta)}{2}), \Delta = 1 \tag{12}$$

which can be transformed to

$$\mathbf{c}_i(l) = \frac{\mathbf{I} + \frac{\Delta}{2}\mathbf{A}_{HiPPO}}{\mathbf{I} - \frac{\Delta}{2}\mathbf{A}_{HiPPO}} \mathbf{c}_i(l-1) + \frac{\Delta \mathbf{B}_{HiPPO}}{\mathbf{I} - \frac{\Delta}{2}\mathbf{A}_{HiPPO}} \sum\nolimits_{o=1}^{O} u_{i,o}(l) \tag{13}$$

which simplifies to

$$\mathbf{c}_i(l) = \mathbf{A}\mathbf{c}_i(l-1) + \mathbf{B} \sum\nolimits_{o=1}^{O} u_{i,o}(l) \tag{14}$$

As a result, $\mathbf{A} = \mathbf{A}_{Legendre}(O, L) \in \mathbb{R}^{O \times O}$ and $\mathbf{B} = \mathbf{B}_{Legendre}(O, L) \in \mathbb{R}^{O \times 1}$ are determined by Legendre bases.

B Subjective Results of Temporal Sentence Generation with SMM

In Fig. 7, the green bounding boxes indicate the anomalies that can be detected by directly applying the Q-Former, as described in Li et al. (2023), for image-grounded text generation in TSGM. The red bounding boxes show the cases where only with the combination of SMM and Q-Former in TSGM can the anomalies be detected. The curves show anomaly scores. Under poor observational conditions like occlusions and low resolutions, SMM complements the Q-Former in TSGM to effectively detect abnormal events.

C Structures of Modules for Image-Text Alignment and Image-Grounded Text Generation

ITAM is the image transformer of Q-Former Li et al. (2023), as is shown in Fig. 8. It outputs $\mathbf{F}_i^I(t) \in \mathbb{R}^{N_q \times H_d}$ which is aligned with the output from text transformer Li et al. (2023) during training to learn extracting text-aligned features. The text tokenizer in Fig. 2 is part of the text transformer of Q-Former Li et al. (2023). ITAM and text tokenizer are frozen in our work.

Image Transformer To select from $\mathbf{H}_i^I(t)$ the caption-informative local patterns, this module is built with self-attention layers, cross-attention layers and feed-forward layers, as is shown by Fig. 8. Firstly, N_q learnable query embeddings attend to each other in self-attention layers before interacting with $\mathbf{H}_i^I(t)$ through cross-attention layers. Each query embedding has dimension H_d. The Image-attention Module involves 6 sequential transformer layers each of which includes one self-attention layer, one cross-attention layer and one feed-forward layer. $\mathbf{H}_i^I(t)$ acts as a static input to the cross-attention layers across all transformer layers. The transformer layers sequentially refine the understanding and integration of $\mathbf{H}_i^I(t)$ with learned queries. Each self-attention layer is implemented according to Vaswani et al. (2017) with 12 heads, producing output $\mathbf{Q}_i^I(t) \in \mathbb{R}^{N_q \times H_d}$. Each cross-attention layer has 12 heads with $\mathbf{H}_i^I(t)$ functioning as key and value, it performs feature fusion by combining $\mathbf{H}_i^I(t)$ with $\mathbf{Q}_i^I(t)$ to $\mathbf{Z}_i^I(t) \in \mathbb{R}^{N_q \times H_d}$. $\mathbf{Z}_i^I(t)$ is projected by fully-connected feed-forward layers to $\mathbf{F}_i^I(t) \in \mathbb{R}^{N_q \times H_d}$.

Text Transformer To encode captions, the module is built with self-attention layers and feed-forward layers, as is shown by Fig. 8. The self-attention layers and feed-forward layers are shared by Image-attention Module and Text-attention Module. In self-attention modules, the text tokens $\mathbf{E}_j^I(t) = [\mathbf{E}_{j,1}^I(t), \mathbf{E}_{j,2}^I(t), ..., \mathbf{E}_{j,S_l}^I(t)] \in \mathbb{R}^{S_l \times H_d}$ in a sentence with maximum length S_l attend to each other. $S_l = N_q$ and $\mathbf{E}_{j,1}^I(t) \in \mathbb{R}^{H_d}, \mathbf{E}_{j,2}^I(t) \in \mathbb{R}^{H_d}, ..., \mathbf{E}_{j,S_l}^I(t) \in \mathbb{R}^{H_d}$.

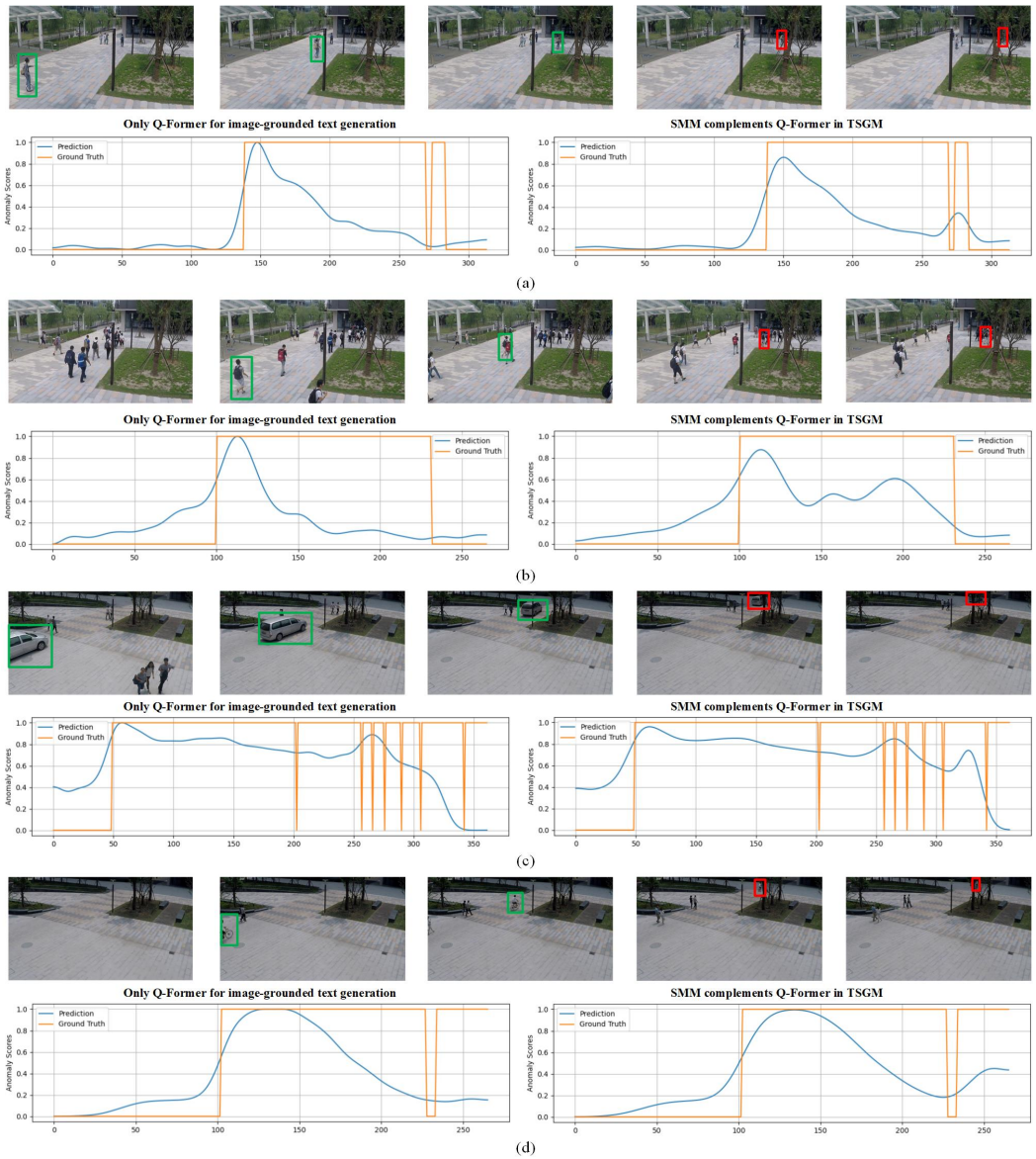

Figure 7: Demonstration of the effectiveness of SMM in TSGM. (a) The man is riding a unicycle but viewed under low resolutions. (b) The man is running but viewed under low resolutions. (c) The vehicle is viewed under occlusions. (d) The man is riding a bicycle but viewed under low resolutions.

To shorten the embeddings of an entire sequence, we follow Devlin et al. (2018) by prepending special token [CLS] to the start of input sequence for aggregating information based on the fact that all tokens attend to each other. Due to the fact that the first token informs about the whole sequence, we only keep the first element $\mathbf{F}_i^T(t)[0] \in \mathbb{R}^{H_d}$ of text transformer's output $\mathbf{F}_i^T(t) \in \mathbb{R}^{S_l \times H_d}$.

Image-Grounded Text Generation Module Conditioned on visual features $\mathbf{F}_i^I(t)$, the module iteratively generates new text tokens until the full sentence with maximum length S_l is produced. Following Radford et al. (2018) Devlin et al. (2018) where token "[BOS]" signals the start of text generation, we initialize the sentence to be "[BOS]" following by $S_l - 1$ zero placeholders. In each iteration, a new text token is generated and replaces one zero placeholder, as is shown in Fig. 9.

In the $k - th$ iteration, the input sequence has previously generated tokens $\mathbf{S}_{i,1}(t), \mathbf{S}_{i,2}(t), ..., \mathbf{S}_{i,k}(t)$ followed by $S_l - k$ zero placeholders, producing embeddings $\mathbf{E}_{i,1}^T(t), \mathbf{E}_{i,2}^T(t), ..., \mathbf{E}_{i,S_l}^T(t)$. The visual embeddings are concatenated with text tokens, producing

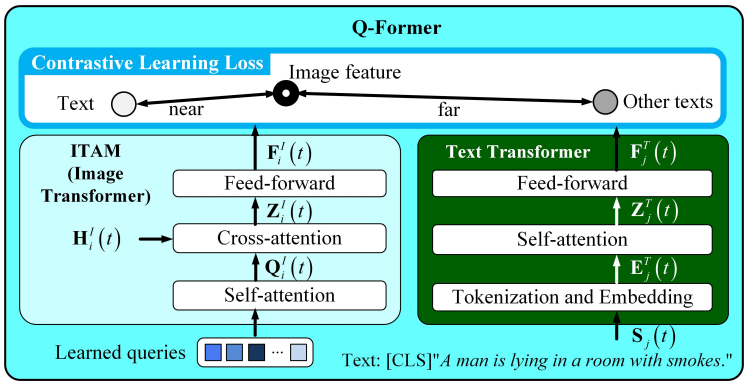

Figure 8: ITAM is the image transformer in Q-Former Li et al. (2023), it identifies the local features through aligning the visual features from image transformer with textual features from text transformer.

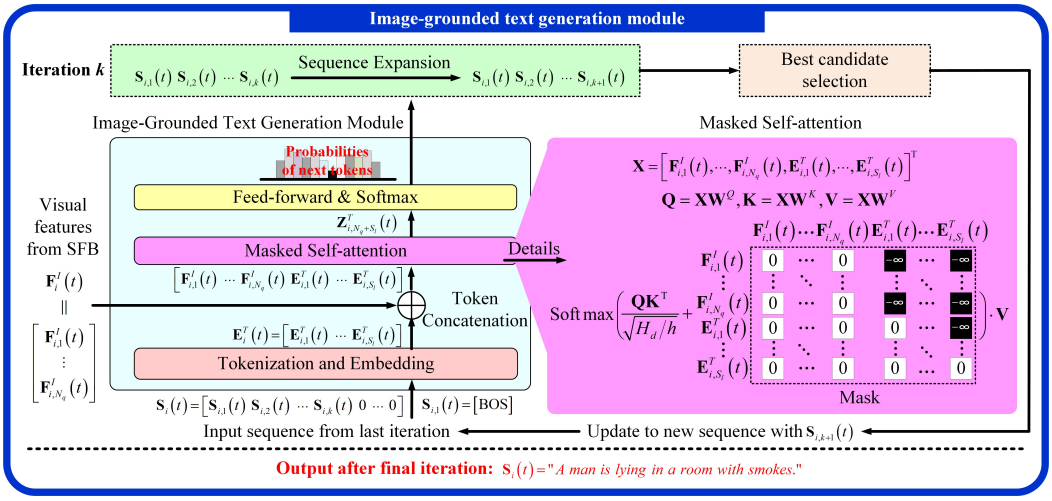

Figure 9: Structure of the module for image-grounded text generation, the module is part of the Q-Former Li et al. (2023).

$\mathbf{X} = [\mathbf{F}^I_{i,1}(t), \mathbf{F}^I_{i,2}(t), ..., \mathbf{F}^I_{i,N_q}(t), \mathbf{E}^T_{i,1}(t), \mathbf{E}^T_{i,2}(t), ..., \mathbf{E}^T_{i,S_l}(t)]^T$ as the input to self-attention layer. As is shown in Fig. 9, the mask in self-attention layer enables visual tokens to attend to each other, and facilitates each of the S_l text tokens attend to all visual tokens and earlier text tokens. Specifically, provided query, key and values $\mathbf{Q} = \mathbf{XW^Q}$, $\mathbf{K} = \mathbf{XW^K}$ and $\mathbf{V} = \mathbf{XW^V}$ with $\mathbf{W^Q}$, $\mathbf{W^K}$ and $\mathbf{W^V}$ being learnable weights, self-attention is implemented by

$$\mathbf{Z}^T_i(t) = Softmax(\frac{\mathbf{QK^T}}{\sqrt{H_d/h}} + \mathbf{M})\mathbf{V} \tag{15}$$

where the values in mask $\mathbf{M}$ are shown by black and white rectangles in Fig. 9. $h = 12$ denotes the number of heads. $\mathbf{Z}^T_i(t) = [\mathbf{Z}^T_{i,1}(t), ..., \mathbf{Z}^T_{i,N_q+S_l}(t)]^T \in \mathbb{R}^{(N_q+S_l)\times H_d}$. Only the last token $\mathbf{Z}^T_{i,N_q+S_l}(t)$ is fed into feed-forward layer because the last token is informative about the complete sequence. The feed-forward layer has H_d input channels and $N_{vocabulary}$ output channels, producing $N_{vocabulary} = 30,523$ probabilities indicating the likelihood of candidate tokens. $N_{vocabulary}$ is vocabulary size, according to BERT tokenizer Devlin et al. (2018). The best candidate $\mathbf{S}_{i,k+1}(t)$ is appended to the end of sequence $\mathbf{S}_{i,1}(t), \mathbf{S}_{i,2}(t), ..., \mathbf{S}_{i,k}(t)$ before beginning the next iteration. The iterations terminate upon generating the whole sequence $\mathbf{S}_i(t)$ with length S_l. This module is trained with cross-entropy loss.

Table 4: Performance (AUC, %) on the benchmarks. ST, Ave, UB, Ped2 and NWPU represent ShanghaiTech, CUHK Avenue, Ubnormal, UCSD Ped2 and NWPU Campus, respectively. Macro-AUC and micro-AUC Reiss & Hoshen (2022) are evaluated.

Algorithm	ST	Ave	UB	Ped2	NWPU
Ours with VLM based detector	92.7 / 88.9	94.9 / 94.5	86.8 / 81.5	99.8 / 99.1	73.5 / 71.6
Ours with VLM based detector (w/o box expansion)	92.3 / 88.4	93.6 / 93.2	86.6 / 81.2	99.7 / 99.0	73.0 / 71.2
Ours with Yolo detector Wang et al. (2023)	92.8 / 89.0	94.9 / 94.5	86.9 / 81.5	99.8 / 99.1	73.7 / 71.7
Ours with Yolo detector (w/o box expansion)	92.3 / 88.5	93.7 / 93.3	86.6 / 81.3	99.7 / 99.0	73.0 / 71.2
Sliding windows	80.8 / 77.6	80.5 / 79.9	75.4 / 72.9	88.6 / 87.8	66.5 / 64.4
Ours with VLM based detector, RM with 5 layers	91.3 / 87.4	92.8 / 92.1	85.1 / 80.4	99.0 / 98.6	72.0 / 70.1
Ours with VLM based detector, RM with 9 layers	91.4 / 87.3	92.9 / 92.3	85.4 / 80.6	99.4 / 98.9	72.2 / 70.3

D Ablation Study on the Method for Object Detection

Table 4 shows the comparison between using VLM Bai et al. (2023), YOLOWang et al. (2023) and sliding windows for object detection. Specifically, window sizes are fixed as follows: 224 for ShanghaiTech, 320 for CUHK Avenue, 320 for Ubnormal, 60 for UCSD Ped2, and 224 for NWPU Campus, with the aim of including largest objects. The results indicate that effective object detection is crucial for accurate performance. Furthermore, the comparisons between the settings with and without bounding box expansions show that bounding box expansions contribute to capturing more contextual information, benefiting performance.

E Ablation Study on ITAM's Structure

Table 5 shows the influences of ITAM's structures and training data on performance. Setting 1 is the default setting with "Str. 1" and "D. 1". "Str. 1" denotes the structure Li et al. (2023) shown in Section 3.2 and "D. 1" denotes the training data of BLIP-2 Li et al. (2023). In "Str. 1", the image transformer for feature extraction has 6 transformer layers each of which includes one self-attention layer, one cross-attention layer and one feed-forward layer. Both of the self-attention layer and the cross-attention layer have 12 heads. In "Str. 2", the numbers of heads are changed to 6 with other settings fixed. In "Str. 3", the number of sequential transformer layers is changed to 3 with other hyperparameters unchanged. "D. 2" refers to the configuration where ITAM is trained on the benchmark's training set in each experiment. These training sets include only normal events. The captioning labels on benchmarks' training data are generated by running the pre-trained BLIP-2 model Li et al. (2023) on the normal videos. It can be seen that the structure and data variations do not significantly influence performance as long as image-text alignment is conducted. More importantly, ITAM can be trained using normal data and detect unseen anomalies.

Setting 5 and 6 show that the number of image tokens N_q does not significantly influence performance. Setting 7 shows that if SMM is trained using the captioning labels from dataset Lin et al. (2014) and without requiring Qwen-Chat, performance is not influenced. For instance, if the captioning label of an image is "The man is running" which prompts SMM to output "yes", then we randomly sample another sentence with a different meaning, such as "The man is fighting", which causes SMM to output "no". Implementations are based on NLTK library Hardeniya et al. (2016).

F Ablation Study on the Number of layers in RM

Table 4 compares the performance of our RM with 7 layers to configurations with 5 layers and 9 layers, respectively. It can be seen that 7 is a better choice.

Table 5: Ablations of ITAM's structure using Micro-AUC. TME and TSG are short for Temporal Motion Estimation and Temporal Sentence Generation, respectively. N_q is the number of image tokens.

Setting	Stage 1	Stage 2	N_q	TME	TSG	N_{SMM}	ST	Ave	UB	Ped2
1	Str. 1, D. 1	✓	32	✓	✓	3	88.9	94.5	81.5	99.1
2	Str. 1, D. 2	✓	32	✓	✓	3	88.9	94.5	81.5	99.1
3	Str. 2, D. 1	✓	32	✓	✓	3	88.6	94.1	81.3	99.1
4	Str. 3, D. 1	✓	32	✓	✓	3	88.7	94.4	81.2	99.1
5	Str. 1, D. 1	✓	64	✓	✓	3	88.8	94.5	81.5	99.1
6	Str. 1, D. 1	✓	128	✓	✓	3	88.9	94.6	81.5	99.1
7	Str. 1, D. 1	✓	32	✓	SMM w/o Qwen	3	88.9	94.5	81.5	99.1

G Procedures for Generating Questions in TSGM

As is shown in Fig. 2(b), TSGM firstly converts the declarative sentence "The man is pushing a stroller on the street." to an interrogative sentence "Is the man pushing a strollor on the street?" The conversion is based on nltk library Hardeniya et al. (2016) and the procedures are shown in Algorithm 2:

Algorithm 2 Algorithm for Converting Declarative Sentences to Interrogative Sentences

1: Input sentence: $\mathbf{D} \leftarrow \mathbf{S}_i^I(t-1) =$ 'The man is pushing a stroller on the street.'
2: Tokenization: $\mathbf{D} \rightarrow$ ['The', 'man', 'is', 'pushing', 'a', 'stroller', 'on', 'the', 'street', '.']
3: Locate first verb: $\mathbf{D}_{firstverb}$ = 'is'
4: Divide sentence using first verb: $\mathbf{D} \rightarrow \mathbf{D}_1 + \mathbf{D}_{firstverb} + \mathbf{D}_2$, $\mathbf{D}_1 =$ 'The man', $\mathbf{D}_2 =$ 'pushing a stroller on the street'
5: Change the order of parts: $\mathbf{Q} \leftarrow \mathbf{D}_{firstverb} + \mathbf{D}_1 + \mathbf{D}_2$
6: **return** $\mathbf{Q}$

H Operational Efficiency

All experiments are conducted on an NVIDIA A100 GPU and an Intel(R) Xeon(R) Gold 6248R CPU. For object detection, we have employed both YOLOv7 detector Wang et al. (2023) and another detector based on Qwen-VL-7B Bai et al. (2023). As is shown by Table 4, both detectors achieve similar accuracy. In terms of inference speed, the YOLO detector Wang et al. (2023) processes each frame in 1.5 milliseconds, whereas Qwen-VL-7B Bai et al. (2023) requires 5.2 seconds per frame. Consequently, we evaluate the operational efficiency of the proposed framework using the YOLO detector. The inference times of all components in the proposed framework are measured and summarized in Table 6. With all components considered, the proposed method achieves an average frame rate of 12 FPS with an average of 5 objects per frame. The average number of 5 objects is based on the findings of Wang et al. (2022a).

In the future, we aim to explore the methods that utilize a fixed number of bounding boxes per frame to maintain a constant inference time, even with an increased number of objects. Additionally, we will investigate parsing multiple objects within a single bounding box to maintain a fixed number of bounding boxes per frame.

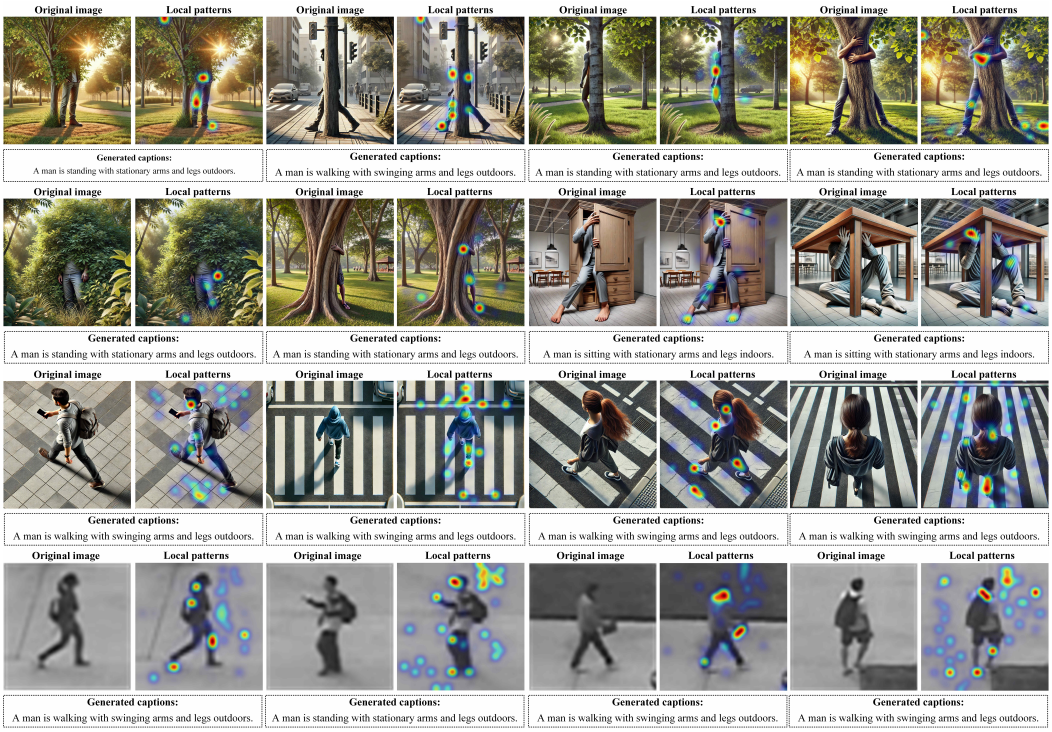

Figure 10: Visualization of the heatmaps of local patterns under occlusions, viewpoint changes and low resolutions.

Table 6: Runtime and memory consumption of different modules in the proposed framework, runtime is measured in milliseconds(ms). Inference is conducted with batch size 256.

Modules	Object detection (YOLO)	Backbone and ITAM	TSGM (SMM)	TSGM (IGTG)	CMAM	TME	RM
Runtime	2.5	10.1	1.5	5.9	0.0078	2.6	0.28
GPU Memory (Gigabytes)	2.78	18.88	0.63	13.54	0.04	0.0	0.55

Table 7 compares the proposed approach with baseline LLM-based AnomalyRuler Yang et al. (2024a). AnomalyRuler involves a VLM Processing stage with CogVLM-17B and a LLM Reasoning stage with GPT-4, consuming 192.56 ms and 504.79 ms per frame on NVIDIA A100 GPU, respectively .

Table 7: Comparison between the proposed method and LLM-based anomaly detector Yang et al. (2024a). Runtime measured in milliseconds(ms), performance measured in AUC (%).

Methods	Runtime per frame	Performance on Shanghaitech	Performance on Avenue
Ours	83.95	88.9	94.5
AnomalyRuler Yang et al. (2024a)	697.35	85.2	89.7

I VISUALIZATION OF LOCAL PATTERNS UNDER OCCLUSIONS AND VIEWPOINT CHANGES

In real-world surveillance videos, occlusions, viewpoint variations and low-resolution conditions are common. Fig. 10 shows some examples of the local patterns identified by image-text alignment and cross-modality attention. The local patterns capture semantically meaningful features such as body joints which are consistent across the variations. The compact representations ignore redundant details and contribute to generalizable embeddings.

J FUTURE WORK

One limitation of the current framework is the reliance on object detectors. Currently, the performance of current Vision-Language Models (VLMs) is limited by their fields of view. For example, when processing an image with a large scene, a vision-language model tends to overlook many details, highlighting the necessity of object detectors that facilitate the processing of local regions independently. Table 4 shows that object detectors significantly outperform sliding windows, the poor performance of the latter may result from an incorrect strategy. As a result, we will try more efficient and effective ways to parse events in complex scenes and images with large fields of view where many objects reside. Specifically, we will explore the integration of object detectors in an end-to-end large model. In simpler scenes with fewer objects, an input image is embedded with fewer vision tokens. As scenes become more complex, more objects are involved, then an input image is encoded with an increased number of vision tokens each of which describes one or more objects. Besides, we will explore ways to improve efficiency.