

000 001 002 003 004 005 LOCAL PATTERNS GENERALIZE BETTER FOR NOVEL 006 ANOMALIES 007 008 009 010

011 **Anonymous authors**
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046
047
048
049
050
051
052
053
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-

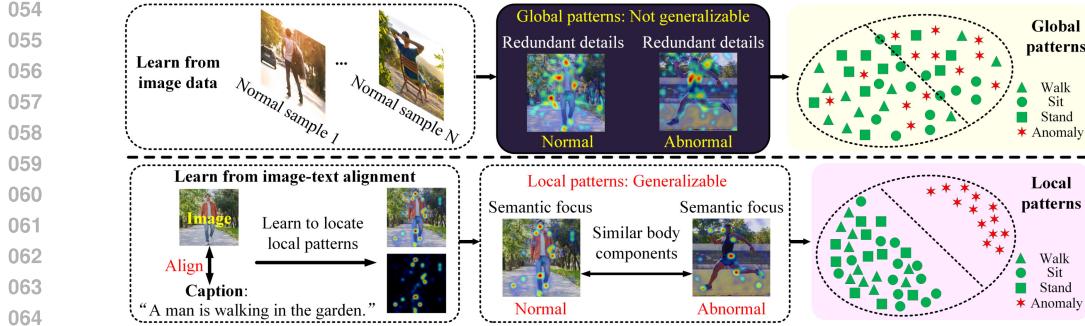


Figure 1: Top: Existing methods rely on global patterns with redundant details, which are inconsistent across visual data variations, limiting their generalization to novel samples. As a result, normal and abnormal samples are poorly distinguished. Bottom: Our method focuses on local patterns that capture 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).

108 Reconstruction-based approaches Astrid et al. (2024); Yang et al. (2023b); Fang et al. (2020); Li
 109 et al. (2020a); Gong et al. (2019); Asad et al. (2021); Abati et al. (2019); Sabokrou et al. (2018)
 110 produce increased error when encountering irregular features Ramachandra et al. (2020); Madan
 111 et al. (2023); Yu et al. (2023) that do not reside in training data. For instance, the method Zaheer
 112 et al. (2022a) learns not to reconstruct anomalies. Gong et al. (2019); Gao et al. (2022) augment
 113 encoders to improve the sensitivity of reconstruction error to anomalies. Madan et al. (2021); Chang
 114 et al. (2020); Singh et al. (2023); Yu et al. (2022b); Shi et al. (2023a) integrate multi-modal features
 115 Ding et al. (2021) while Huang et al. (2022) integrates a probabilistic decision model. Zaheer
 116 et al. (2022b) assesses the quality of reconstruction to improve stability. Prediction-based methods
 117 Luo et al. (2021b); Morais et al. (2019); Luo et al. (2021a); Liu et al. (2018); Nguyen & Meunier
 118 (2019); Zeng et al. (2021) evaluate the divergence in normal and abnormal temporal dependencies,
 119 leveraging latent spaces Zhang et al. (2020) or hybrid attention Zhang et al. (2022b).

120 To better distinguish anomalies, Lv et al. (2021); Lu et al. (2020); Liu et al. (2021); Park et al. (2020);
 121 Li et al. (2021a) combine prediction with reconstruction. Sato et al. (2023); Wu et al. (2023); Luo
 122 et al. (2019) study the distribution over normal samples and propose novel features Arad & Werman
 123 (2023). Similarly, Yan et al. (2023) proposes denoising diffusion modules. Flaborea et al. (2023)
 124 exploits the enhanced mode coverage of diffusive probabilistic models. To improve representation
 125 capacities, Chang et al. (2021); Fan et al. (2024) propose snippet-level attention. Liu et al. (2023a);
 126 Yu et al. (2022a); Purwanto et al. (2021) introduce pyramid deformation and CRFs to learn spatio-
 127 temporal dependencies Bertasius et al. (2021); Cho et al. (2022). Wang et al. (2021) combines multi-
 128 scale features to enhance prediction. Stergiou et al. (2024) combines interpolation with extrapolation
 129 for prediction. Wang et al. (2022b) proposes a self-supervised scheme with discriminative DNNs.
 130 We propose generalizable local patterns to better represent unseen samples.

131 2.2 WEAKLY SUPERVISED ANOMALY DETECTION

132 Multi-instance learning (MIL) takes videos as bags and snippets as instances, transforming video-
 133 level labels to instance-level Feng et al. (2021). The methods iteratively locate abnormal segments
 134 and fine-tune models using anomalous segments which are dissimilar to normal ones Zhang et al.
 135 (2023a). To collect abnormal segments, inter-sample distances are evaluated Lu et al. (2022);
 136 Ionescu et al. (2019) based on spatio-temporal similarities Dhiman & Vishwakarma (2020); Lv
 137 et al. (2023); Chang et al. (2020); Markovitz et al. (2020). Li et al. (2021b) proposes a probabilistic
 138 framework. Sun et al. (2020); Li et al. (2020b) build graphical representations and integrated col-
 139 lective properties in measuring similarities. Sapkota & Yu (2022) performs dynamic non-parametric
 140 clustering. To improve robustness, Zhang et al. (2023b) proposes to interpret the vulnerability of
 141 MIL. Wu & Liu (2021) introduces causal relations to enhance MIL Tian et al. (2021). Yang et al.
 142 (2023a) proposes binary network augmentation strategy. Differently, we propose generalizable rep-
 143 resentations which facilitate the measurement of similarities between seen and unseen events.

144 2.3 METHODS WITH DATA AUGMENTATION

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

154 2.4 METHODS EXPLORING THE REPRESENTATION OF UNSEEN CATEGORIES

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

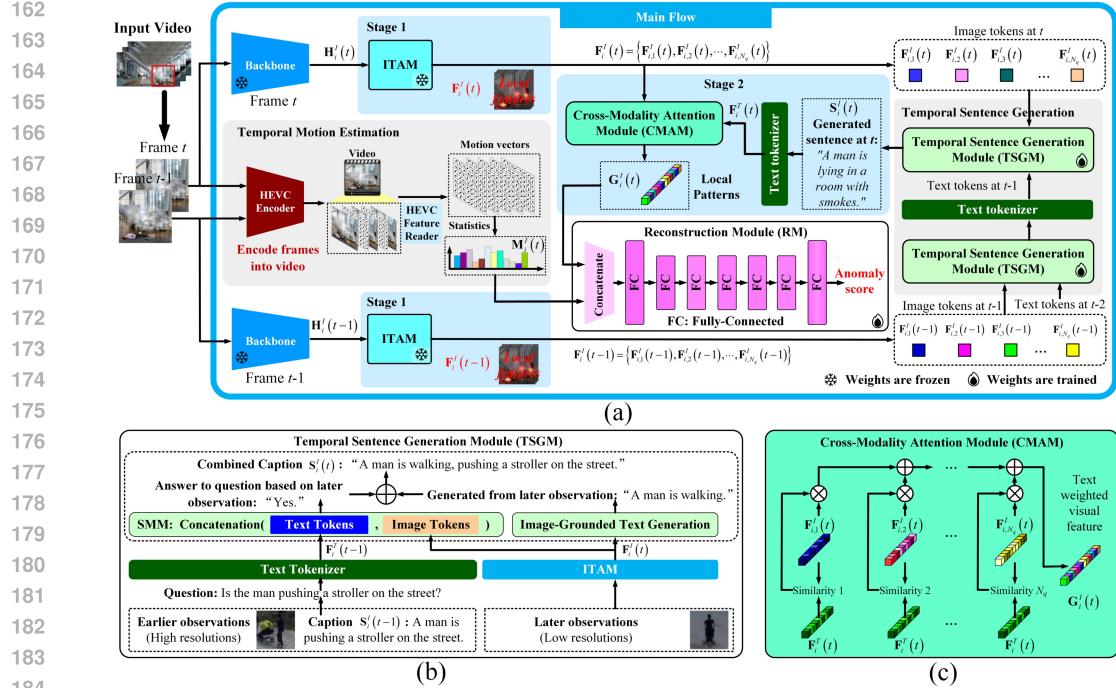


Figure 2: Structure of the method. (a) Main flow: Visual features are extracted using backbone and Image-Text Alignment Module (ITAM) which identifies caption-informative local features as image tokens. Conditioned on image tokens, the Temporal Sentence Generation Module (TSGM) generates sentences which 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.

216 3.1 CROPPING OF IMAGE REGIONS
217

218 Due to the wide field of view in some frames containing numerous objects, it is difficult even for
 219 GPT-4 Achiam et al. (2023) to focus on all objects together. As a result, local regions are cropped
 220 as the first step in our pipeline. We have experimented with both YOLOv7 Wang et al. (2023)
 221 and Qwen-7B Bai et al. (2023) for cropping bounding box regions based on prompts. Specifically,
 222 "How many people are there?" and "The bounding box of the i -th object" are sequentially provided
 223 to Qwen-7B which returns corresponding boxes. The comparisons will be included in Appendix D.
 224

225 3.2 STAGE 1 FOR IDENTIFYING SPATIAL LOCAL PATTERNS
226

227 This stage identifies features in cropped image regions that align with texts. The texts describe
 228 generic movement attributes (e.g., "A man is walking with swinging arms and legs"). When en-
 229计ering an unseen action, such as running, the model can recombine known components like
 230 arms and legs to generate descriptive language that captures the essence of the action without ex-
 231 plicitly naming it. As illustrated in Fig. 1, heatmaps indicate the attention on local components,
 232 highlighting similar semantic regions for shared attributes. More visualizations are in Fig. 4.

233 To identify the local patterns that align with texts, the frozen Image Encoder Li et al. (2023) and the
 234 image transformer of Q-Former in BLIP-2 Li et al. (2023) are employed as backbone and ITAM,
 235 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
 236 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
 237 output with N_q image tokens $\mathbf{F}_{i,1}^I(t), \dots, \mathbf{F}_{i,N_q}^I(t)$ which inform about the captions of image region
 238 i and remain consistent over visual data variances, as will be shown by the heatmaps in Fig. 4. Detailed
 239 structures are in Appendix C. Algorithm 1 shows the workflow of Stage 1 and Stage 2.

240 **Algorithm 1** Two-Stage Process for Identifying Spatial Local Patterns
241

-
- 1: **Input:** Input image, Backbone, ITAM, CMAM, TSGM and Text tokenizer
 - 2: **Output:** Cross-modal embedding $\mathbf{G}_i^I(t)$ representing spatial local patterns
 - 3: **Stage 1: Image Token Extraction**
 - 4: Use the backbone to extract feature maps $\mathbf{H}_i^I(t)$
 - 5: 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
 - 6: **Stage 2: Cross-Modality Attention**
 - 7: Feed image tokens $\{\mathbf{F}_{i,j}^I(t)\}_{j=1}^{N_q}$ into TSGM and obtain text tokens $\mathbf{F}_i^T(t)$
 - 8: CMAM weightedly sums $\{\mathbf{F}_{i,j}^I(t)\}_{j=1}^{N_q}$ according to their similarity with $\mathbf{F}_i^T(t)$
 - 9: **Return:** Return weighted sum $\mathbf{G}_i^I(t)$, representing the cross-modal features
-

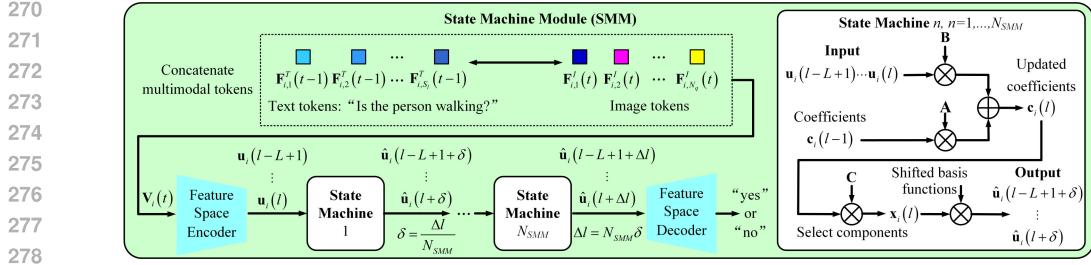
252 3.3 STAGE 2 FOR IDENTIFYING LOCAL PATTERNS
253

254 This stage further highlights local patterns by generating a sentence conditioned on image tokens
 255 and summing them based on their similarities to the generated sentence. Using the image tokens
 256 $\mathbf{F}_{i,1}^I(t), \dots, \mathbf{F}_{i,N_q}^I(t)$ from Stage 1, TSGM generates a sentence for image region i , as is shown in
 257 Fig. 2(a). TSGM utilizes SMM for inter-frame caption augmentation and a frozen Q-Former Li et al.
 258 (2023) for image-grounded text generation. SMM determines whether previous events still reside in
 259 current frame while Q-Former captions current frame. The outputs from SMM and Q-Former are
 260 combined to form the augmented sentence $\mathbf{S}_i^I(t)$. Even with incomplete observations at t , $\mathbf{S}_i^I(t)$ can
 261 recognize previously occurring events from current frame as long as the events still reside.

262 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
 263 in one sentence, is provided to CMAM. CMAM uses the first element in $\mathbf{F}_i^T(t)$ as query and the
 264 image tokens as keys and values for attention operations, as is illustrated in Fig. 2(c) and Eq. (1):
 265

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

266 $\mathbf{F}_i^T(t)$ is obtained by the text transformer in Q-Former Li et al. (2023) with first element $\mathbf{F}_i^T(t)[0]$
 267 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 $\mathbf{F}_i^I(t) = \{\mathbf{F}_{i,1}^I(t), \dots, \mathbf{F}_{i,N_q}^I(t)\}$ 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 $\mathbf{F}_i^T(t-1) = \{\mathbf{F}_{i,1}^T(t-1), \dots, \mathbf{F}_{i,S_l}^T(t-1)\}$. 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 $\mathbf{V}_i(t) = [\mathbf{F}_{i,1}^T(t-1); \dots; \mathbf{F}_{i,S_l}^T(t-1); \mathbf{F}_{i,1}^I(t); \dots; \mathbf{F}_{i,N_q}^I(t)]^\top \in \mathbb{R}^{H_d \times (S_l + N_q)}$ 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) $[g_o(1), \dots, g_o(L)]$, $o \in [1, O]$, 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 $\mathbf{U}_i(t) = [\mathbf{u}_i(l-L+1); \dots; \mathbf{u}_i(l)]^\top \in \mathbb{R}^{O \times L}$, where $\mathbf{u}_i(l') = [u_{i,1}(l'), \dots, u_{i,O}(l')]^\top$ 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:

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

In SMM, a state vector $\mathbf{c}_i(l) = [c_{i,1}(l); \dots; c_{i,O}(l)]^\top$ 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").

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

324 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
 325 Gu et al. (2020). As O grows, more diversified basis functions can represent more complex depen-
 326 dencies. Assume that $\mathbf{c}_i(l)$ encodes $\mathbf{u}_i(l-L+1), \dots, \mathbf{u}_i(l)$ based on which $\mathbf{u}_i(l+1)$ is predicted.
 327 $\mathbf{u}_i(l+1)$ denotes "yes" or "no". $\mathbf{c}_i(l+1)$ encodes $\mathbf{u}_i(l-L+2), \dots, \mathbf{u}_i(l+1)$. Eq. (3) will be derived
 328 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
 329 learnable, highlights important components, Eq. (3) is transformed to
 330

$$\mathbf{x}_i(l) = \mathbf{C}\mathbf{A}^{L-1}\mathbf{B} \sum_{o=1}^O u_{i,o}(l-L+1) + \dots + \mathbf{C}\mathbf{B} \sum_{o=1}^O u_{i,o}(l) \quad (4)$$

331 Finally, the elements of $\mathbf{x}_i(l)$ are multiplied with shifted basis functions $[g_o(1+\delta), \dots, g_o(L+\delta)]$, $o \in$
 332 $[1, O]$, $\Delta l = 1$, $\delta = \Delta l/N_{SMM}$, producing shifted weighted basis functions:
 333

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

334 The Feature Space Decoder projects $\hat{\mathbf{u}}_i(l + \delta) = [\hat{u}_{i,1}(l + \delta), \dots, \hat{u}_{i,O}(l + \delta)]$ onto a prediction
 335 ("yes" or "no"), as is shown by Fig. 3. Cross entropy loss is employed. In each batch, B_s images
 336 correspond to B_s declarative sentences which are converted into B_s questions. The tokens of each
 337 image are concatenated with those of each corresponding question before feeding into SMM.
 338

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

344 where ground truth $y_{i,j}$ takes 1 when the Qwen-Chat model Bai et al. (2023) receives question j
 345 together with image i and returns "yes", else $y_{i,j}$ takes 0. $\text{Sim}(P(i, j), \text{Emb}("yes"))$ is the cosine
 346 similarity between the embedding $P(i, j)$ of SMM's output and the embedding of "yes".
 347

351 3.5 TEMPORAL MOTION ESTIMATION AND SPATIO-TEMPORAL ANOMALY DETECTION

352 To enhance the spatial local patterns obtained from Stage 2, this paper proposes to encode frames
 353 into H.265 (HEVC) videos using FFmpeg Zeng et al. (2016). As is illustrated in Fig. 2(a), motion
 354 vectors from encoded videos are extracted, each motion vector is associated with a 8×8 macroblock.
 355 The orientation of each motion vector is computed as $\text{atan2}(y, x)$ and quantized into $D_m = 8$ equi-
 356 spaced bins, x and y are the horizontal and vertical components. The average magnitudes of motion
 357 vectors in these bins produce a D_m -dimensional histogram $\mathbf{M}_i^I(t)$ representing region i .
 358

359 To detect anomalies with anomalous local patterns or irregular dynamics, the Reconstruction Mod-
 360 ule (RM) with 7 fully-connected layers is trained on normal spatial and temporal data. As is shown
 361 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
 362 maps $H_d + D_m$ input channels to D_h output channels while the last layer maps D_h input channels
 363 to $H_d + D_m$ output channels. The 5 hidden layers have D_h input channels and D_h output chan-
 364 nels. The reconstructions of spatial and temporal features are conducted together, facilitating the
 365 reconstruction of each one to depend on the other. Reconstruction error determines anomaly scores.
 366

367 4 EXPERIMENTS AND RESULTS

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

371 4.1 EXPERIMENTAL SETUP

372 **Datasets** Experiments are conducted on seven datasets. The training sets of ShanghaiTech, Avenue
 373 and UCSD Ped2 contain only normal events and anomalies reside in test data. (1) **ShanghaiTech**
 374 dataset Liu et al. (2018) includes 330 training videos and 107 test videos. Among the two versions
 375 of ShanghaiTech dataset Liu et al. (2018) and Zhong et al. (2019); Li et al. (2022a); Zanella et al.
 376 (2023), the latter includes abnormal behaviors in both training set and test set. As our approach
 377 is unsupervised, we use the first version. (2) **CUHK Avenue** dataset Lu et al. (2013) involves 16

378
 379 Table 1: Performance (AUC, %) on the benchmarks. ST, Ave, UB, Ped2 and NWPU represent
 380 ShanghaiTech, CUHK Avenue, Ubnormal, UCSD Ped2 and NWPU Campus, respectively. Macro-
 381 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

405 training videos and 21 test videos. (3) **Ubnormal** dataset Acsintoae et al. (2022) is divided into a
 406 training set with 268 videos, a validation set with 64 videos, and a test set with 211 videos. (4)
 407 **NWPU Campus** dataset Cao et al. (2023) comprises 43 scenes, 28 classes of anomalies and 16
 408 hours of video footage. (5) **UCSD Ped2** dataset Li et al. (2014) contains 16 normal training videos
 409 and 12 test videos. (6) **UCF Crime** dataset Sultani et al. (2018) includes 1610 training videos in
 410 which 800 contain only normal behaviors. The test set includes 290 videos in which 140 include
 411 anomalies. (7) **XD Violence** Wu et al. (2020) includes 4754 videos where 2349 are non-violent and
 412 2405 are violent. There are 3954 training videos and 800 test videos where 500 are violent.

413 **Evaluation Metrics** Following previous literature Markovitz et al. (2020), Area under Curve (AUC,
 414 %) is adopted for evaluation. Differently, the accuracy on XD-Violence dataset is measured using
 415 precision-recall curve and the corresponding Average Precision (AP, %) Panariello et al. (2022).

416 **Implementation Details** To capture more contexts, bounding boxes are expanded by 50% on both
 417 sides horizontally and vertically. The benefits of box expansion will be shown in Table 4 of Appendix
 418 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 $\mathbf{F}_i^I(t) \in$
 419 $\mathbb{R}^{N_q \times H_d}$ which satisfy $S_d = 257$, $V_d = 1408$, $N_q = 32$, $H_d = 768$. Each of the N_q image tokens
 420 has embedding size H_d . Following BLIP-2 Li et al. (2023), the backbone has "ViT-L/14" structure
 421 in Radford et al. (2021). The text tokenizer in Fig. 2(a) will be detailed in Appendix C. For sentences
 422 with fewer than S_l tokens, $\mathbf{F}_i^T(t)$ is padded with zeros. RM has $D_h = 512$ in intermediate layers.

423 SMM, with $N_{SMM} = 3$ state machines, is trained on the COCO-Caption dataset Lin et al. (2014).
 424 Table 3 shows the influences of N_{SMM} . The Feature Space Encoder (H_d input channels, $O = 64$
 425 output channels) and Feature Space Decoder (O input channels, H_d output channels) are learnable
 426 fully-connected layers, the weights in $\mathbf{C} \in \mathbb{R}^{O \times O}$ are also learnable. All weights are initialized with
 427 distribution $N(0, 0.02)$. Training spans 20 epoches with initial learning rate 5×10^{-5} and decay
 428 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
 429 Adam optimizer with learning rate 10^{-3} for 10 epoches, using MSE loss. Implementations are based
 430 on Pytorch Pytorch (2018) and a NVIDIA A100 GPU. RM is trained on benchmark videos without
 431 anomalies. The influences of RM's number of layers will be shown in Appendix F. The evaluations
 432 on operational efficiency will be detailed in Appendix H.

432
433 Table 2: Performance on UCF-Crime (micro-AUC, %) and XD-Violence (AP, %). UCF and XD
434 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

442
443 Table 3: Ablations of components using Micro-AUC. TME and TSG are short for Temporal Motion
444 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

458 4.2 COMPARISONS WITH BASELINES

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

465 4.3 ABLATION STUDIES

466 **Ablation on Stage 1 and Stage 2** In Setting 1 of Table 3, the reconstruction error of backbone
467 features $\mathbf{H}_i^I(t)$ is used to detect anomalies. Setting 2 and Setting 3 show the utilization of Stage
468 1 and both stages for reconstruction, respectively. The comparison shows that Stages 1 and 2 both
469 play crucial roles in selecting text-informative local patterns, as will be illustrated in Fig. 4.

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

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

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

480 **Ablation on the SMM in Temporal Sentence Generation in Stage 2** The comparison between
481 Setting 6 and Setting 7 shows that if TSGM only uses Q-Former Li et al. (2023) for image-grounded
482 text generation without SMM to incorporate previous captions, performance drops. As a result, the
483 mixture of image tokens and text tokens from different moments contributes to more informative
484 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.

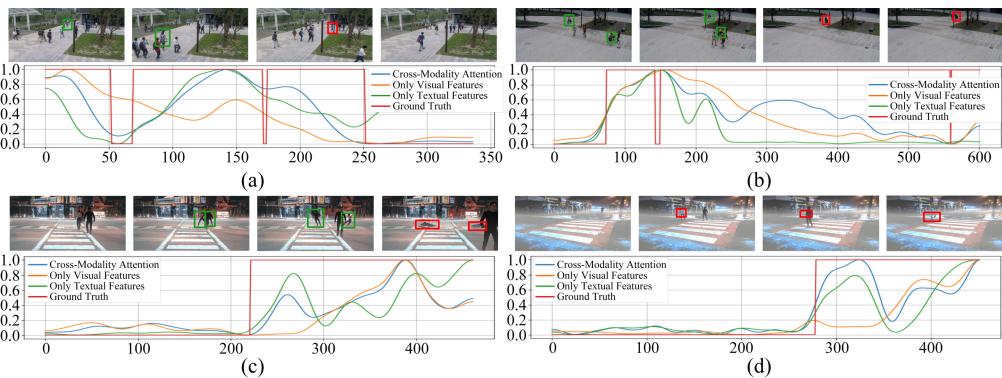


Figure 5: Anomaly scores obtained using image features, text features and combined ones. Cross-modality attention detects anomalies even under occlusion and low resolution. Green and red boxes show the anomalies detected with single modality and cross-modality attention, respectively.

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 $F_i^T(t)$ while the third rows display those for $G_i^T(t)$. 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.

540 REFERENCES
541

- 542 Davide Abati, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Latent space autoregression
543 for novelty detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and*
544 *Pattern Recognition*, pp. 481–490. IEEE, 2019.
- 545 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Ale-
546 man, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical
547 report. *arXiv preprint arXiv:2303.08774*, 2023.
- 548 Andra Acsintoa, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea,
549 Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Ubnormal: New benchmark
550 for supervised open-set video anomaly detection. *arXiv preprint arXiv:2111.08644*, 2021.
551
- 552 Andra Acsintoa, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea,
553 Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Ubnormal: New benchmark
554 for supervised open-set video anomaly detection. In *Proceedings of the IEEE/CVF Conference*
555 *on Computer Vision and Pattern Recognition*, pp. 20143–20153, 2022.
- 556 Yoav Arad and Michael Werman. Beyond the benchmark: Detecting diverse anomalies in videos.
557 *arXiv preprint arXiv:2310.01904*, 2023.
- 558 George B Arfken, Hans J Weber, and Frank E Harris. *Mathematical methods for physicists: a*
559 *comprehensive guide*. Academic press, 2011.
560
- 561 Mujtaba Asad, Jie Yang, Enmei Tu, Liming Chen, and Xiangjian He. Anomaly3d: Video anomaly
562 detection based on 3d-normality clusters. *Journal of Visual Communication and Image Represen-*
563 *tation*, 75:103047, 2021.
- 564 Marcella Astrid, Muhammad Zaigham Zaheer, and Seung-Ik Lee. Synthetic temporal anomaly
565 guided end-to-end video anomaly detection. In *Proceedings of the IEEE/CVF International Con-*
566 *ference on Computer Vision*, pp. 207–214, 2021.
567
- 568 Marcella Astrid, Muhammad Zaigham Zaheer, and Seung-Ik Lee. Constricting normal latent space
569 for anomaly detection with normal-only training data. *arXiv preprint arXiv:2403.16270*, 2024.
- 570 Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang
571 Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, local-
572 ization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.
- 573 Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video
574 understanding? In *ICML*, volume 2, pp. 4, 2021.
575
- 576 Ruichu Cai, Hao Zhang, Wen Liu, Shenghua Gao, and Zhifeng Hao. Appearance-motion memory
577 consistency network for video anomaly detection. In *Proceedings of the AAAI conference on*
578 *artificial intelligence*, volume 35, pp. 938–946, 2021.
- 579 Congqi Cao, Yue Lu, Peng Wang, and Yanning Zhang. A new comprehensive benchmark for semi-
580 supervised video anomaly detection and anticipation. In *Proceedings of the IEEE/CVF Confer-*
581 *ence on Computer Vision and Pattern Recognition*, pp. 20392–20401, 2023.
582
- 583 Raghavendra Chalapathy, A. K. Menon, and S. Chawla. Robust, deep and inductive anomaly de-
584 tection. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge*
585 *Discovery in Databases*, pp. 36–51. Springer, 2017.
- 586 Shuning Chang, Yanchao Li, Shengmei Shen, Jiashi Feng, and Zhiying Zhou. Contrastive attention
587 for video anomaly detection. *IEEE Transactions on Multimedia*, 24:4067–4076, 2021.
- 588 Yunpeng Chang, Zhigang Tu, Wei Xie, and Junsong Yuan. Clustering driven deep autoencoder for
589 video anomaly detection. In *European Conference on Computer Vision*, pp. 329–345. Springer,
590 2020.
591
- 592 Junxi Chen, Liang Li, Li Su, Zheng-jun Zha, and Qingming Huang. Prompt-enhanced multiple in-
593 stance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF*
Conference on Computer Vision and Pattern Recognition, pp. 18319–18329, 2024.

- 594 Yingxian Chen, Zhengze Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu.
 595 Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly
 596 detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 387–
 597 395, 2023.
- 598 Zhi Chen, Jiang Duan, Li Kang, and Guoping Qiu. Supervised anomaly detection via conditional
 599 generative adversarial network and ensemble active learning. *IEEE Transactions on Pattern Anal-*
 600 *ysis and Machine Intelligence*, 45(6):7781–7798, 2022.
- 602 MyeongAh Cho, Taeoh Kim, Woo Jin Kim, Suhwan Cho, and Sangyoun Lee. Unsupervised video
 603 anomaly detection via normalizing flows with implicit latent features. *Pattern Recognition*, 129:
 604 108703, 2022.
- 605 MyeongAh Cho, Minjung Kim, Sangwon Hwang, Chaewon Park, Kyungjae Lee, and Sangyoun
 606 Lee. Look around for anomalies: Weakly-supervised anomaly detection via context-motion re-
 607 lational learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*
 608 *Recognition*, pp. 12137–12146, 2023.
- 609 Bruno Degardin and Hugo Proen  a. Iterative weak/self-supervised classification framework for
 610 abnormal events detection. *Pattern Recognition Letters*, 145:50–57, 2021.
- 612 Andong Deng, Taojiannan Yang, and Chen Chen. A large-scale study of spatiotemporal representa-
 613 tion learning with a new benchmark on action recognition. pp. 20519–20531, 2023.
- 615 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep
 616 bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- 617 Chhavi Dhiman and Dinesh Kumar Vishwakarma. View-invariant deep architecture for human ac-
 618 tion recognition using two-stream motion and shape temporal dynamics. *IEEE Transactions on*
 619 *Image Processing*, 29:3835–3844, 2020.
- 621 Mingyu Ding, Zhenfang Chen, Tao Du, Ping Luo, Josh Tenenbaum, and Chuang Gan. Dynamic
 622 visual reasoning by learning differentiable physics models from video and language. In *Advances*
 623 *in Neural Information Processing Systems*, volume 34, 2021.
- 625 Keval Doshi and Yasin Yilmaz. Continual learning for anomaly detection in surveillance videos.
 626 In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Work-*
 627 *shops*, pp. 254–255, 2020.
- 628 Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for
 629 open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF*
 630 *Conference on Computer Vision and Pattern Recognition*, pp. 14084–14093, 2022.
- 632 Yidan Fan, Yongxin Yu, Wenhuan Lu, and Yahong Han. Weakly-supervised video anomaly de-
 633 tection with snippet anomalous attention. *IEEE Transactions on Circuits and Systems for Video*
 634 *Technology*, 2024.
- 635 Zhiwen Fang, Joey Tianyi Zhou, Yang Xiao, Yanan Li, and Feng Yang. Multi-encoder towards
 636 effective anomaly detection in videos. *IEEE Transactions on Multimedia*, 23:4106–4116, 2020.
- 638 Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training frame-
 639 work for video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer*
 640 *vision and pattern recognition*, pp. 14009–14018, 2021.
- 641 Alessandro Flaborea, Luca Collorone, Guido Maria D’Amely Di Melendugno, Stefano D’Arrigo,
 642 Bardh Prenkaj, and Fabio Galasso. Multimodal motion conditioned diffusion model for skeleton-
 643 based video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on*
 644 *Computer Vision*, pp. 10318–10329, 2023.
- 645 Jie Gao, Bineng Zhong, and Yan Chen. Robust tracking via learning model update with unsupervised
 646 anomaly detection philosophy. *IEEE Transactions on Circuits and Systems for Video Technology*,
 647 2022.

- 648 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
 649 learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12742–12752. IEEE, 2021a.
 650
 651
- 652 Mariana Iuliana Georgescu, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and
 653 Mubarak Shah. A background-agnostic framework with adversarial training for abnormal event
 654 detection in video. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):4505–
 655 4523, 2021b.
 656
- 657 Dong Gong, L. Liu, V. Le, B. Saha, M.R. Mansour, S. Venkatesh, and A.V. Den Hengel. Memorizing
 658 normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly
 659 detection. In *Proceedings of IEEE International Conference on Computer Vision*, pp. 1705–1714.
 660 IEEE, 2019.
- 661 Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. Hippo: Recurrent memory
 662 with optimal polynomial projections. *Advances in neural information processing systems*, 33:
 663 1474–1487, 2020.
- 664
 665 Chongye Guo, Hongbo Wang, Yingjie Xia, and Guorui Feng. Learning appearance-motion synergy
 666 via memory-guided event prediction for video anomaly detection. *IEEE Transactions on Circuits
 667 and Systems for Video Technology*, 2023.
- 668 Nitin Hardeniya, Jacob Perkins, Deepti Chopra, Nisheeth Joshi, and Iti Mathur. *Natural language
 669 processing: python and NLTK*. Packt Publishing Ltd, 2016.
- 670
 671 Or Hirschorn, Shai Avidan, and Shai Avidan. Normalizing flows for human pose anomaly detec-
 672 tion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13545–
 673 13554, 2023.
- 674 Xin Huang, Yutao Hu, Xiaoyan Luo, Jungong Han, Baochang Zhang, and Xianbin Cao. Boosting
 675 variational inference with margin learning for few-shot scene-adaptive anomaly detection. *IEEE
 676 Transactions on Circuits and Systems for Video Technology*, 2022.
- 677
 678 Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. Object-
 679 centric auto-encoders and dummy anomalies for abnormal event detection in video. In *Proceed-
 680 ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7842–7851,
 681 2019.
- 682 Hyekang Kevin Joo, Khoa Vo, Kashu Yamazaki, and Ngan Le. Clip-tsa: Clip-assisted temporal self-
 683 attention for weakly-supervised video anomaly detection. In *2023 IEEE International Conference
 684 on Image Processing (ICIP)*, pp. 3230–3234. IEEE, 2023.
- 685
 686 Siwon Kim, Kukjin Choi, Hyun-Soo Choi, Byunghan Lee, and Sungroh Yoon. Towards a rigorous
 687 evaluation of time-series anomaly detection. In *Proceedings of the AAAI Conference on Artificial
 688 Intelligence*, volume 36, pp. 7194–7201. AAAI Press, 2022.
- 689 Bo Li, Sam Leroux, and Pieter Simoens. Decoupled appearance and motion learning for effi-
 690 cient anomaly detection in surveillance video. *Computer Vision and Image Understanding*, 210:
 691 103249, 2021a.
- 692
 693 Guoqiu Li, Guanxiong Cai, Xingyu Zeng, and Rui Zhao. Scale-aware spatio-temporal relation
 694 learning for video anomaly detection. In *European Conference on Computer Vision*, pp. 333–
 695 350. Springer, 2022a.
- 696 Jing Li, Qingwang Huang, Yingjun Du, Xiantong Zhen, Shengyong Chen, and Ling Shao. Varia-
 697 tional abnormal behavior detection with motion consistency. *IEEE Transactions on Image Pro-
 698 cessing*, 31:275–286, 2021b.
- 699
 700 Junnan Li, Caiming Xiong, and Steven CH Hoi. Learning from noisy data with robust representation
 701 learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.
 9485–9494, 2021c.

- 702 Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image
 703 pre-training with frozen image encoders and large language models. In *International conference*
 704 *on machine learning*, pp. 19730–19742. PMLR, 2023.
- 705 Nanjun Li, Faliang Chang, and Chunsheng Liu. Spatial-temporal cascade autoencoder for video
 706 anomaly detection in crowded scenes. *IEEE Transactions on Multimedia*, 23:203–215, 2020a.
- 708 Nanjun Li, Faliang Chang, and Chunsheng Liu. A self-trained spatial graph convolutional network
 709 for unsupervised human-related anomalous event detection in complex scenes. *IEEE Transactions*
 710 *on Cognitive and Developmental Systems*, 2022b.
- 711 Shuo Li, Fang Liu, and Licheng Jiao. Self-training multi-sequence learning with transformer for
 712 weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial*
 713 *Intelligence*, volume 24. AAAI Press, 2022c.
- 715 Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in
 716 crowded scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):18–
 717 32, 2014.
- 719 Xuelong Li, Mulin Chen, and Qi Wang. Quantifying and detecting collective motion in crowd
 720 scenes. *IEEE Transactions on Image Processing*, 29:5571–5583, 2020b.
- 721 Swee Kiat Lim, Yi Loo, Ngoc-Trung Tran, Ngai-Man Cheung, Gemma Roig, and Yuval Elovici.
 722 Doping: Generative data augmentation for unsupervised anomaly detection with gan. In *2018*
 723 *IEEE international conference on data mining (ICDM)*, pp. 1122–1127. IEEE, 2018.
- 725 Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr
 726 Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European*
 727 *conference on computer vision*, pp. 740–755. Springer, 2014.
- 728 Xiangru Lin, Yuyang Chen, Guanbin Li, and Yizhou Yu. A causal inference look at unsupervised
 729 video anomaly detection. In *Proceedings of the Thirty-sixth AAAI Conference on Artificial*
 730 *Intelligence*. AAAI Press, 2022.
- 732 Boyang Liu, Pang-Ning Tan, and Jiayu Zhou. Unsupervised anomaly detection by robust density
 733 estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 2022a.
- 734 Wen Liu, W. Luo, D. Lian, and S. Gao. Future frame prediction for anomaly detection—a new
 735 baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*,
 736 pp. 6536–6545. IEEE, 2018.
- 738 Wenrui Liu, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Diversity-measurable
 739 anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*
 740 *Recognition*, pp. 12147–12156, 2023a.
- 741 Yang Liu, Dingkang Yang, Yan Wang, Jing Liu, Jun Liu, Azzedine Boukerche, Peng Sun, and Liang
 742 Song. Generalized video anomaly event detection: Systematic taxonomy and comparison of deep
 743 models. *ACM Computing Surveys*, 2023b.
- 745 Yuejiang Liu, Riccardo Cadei, Jonas Schweizer, Sherwin Bahmani, and Alexandre Alahi. Towards
 746 robust and adaptive motion forecasting: A causal representation perspective. In *Proceedings of*
 747 *the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17081–17092. IEEE,
 748 2022b.
- 749 Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. A hybrid video anomaly
 750 detection framework via memory-augmented flow reconstruction and flow-guided frame predi-
 751 cation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13588–
 752 13597, 2021.
- 754 Zuhao Liu, Xiao-Ming Wu, Dian Zheng, Kun-Yu Lin, and Wei-Shi Zheng. Generating anomalies for
 755 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.

- 756 Cewu Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the*
 757 *IEEE International Conference on Computer Vision*, pp. 2720–2727. IEEE, 2013.
 758
- 759 Yiwei Lu, Frank Yu, Mahesh Kumar Krishna Reddy, and Yang Wang. Few-shot scene-adaptive
 760 anomaly detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK,*
 761 *August 23–28, 2020, Proceedings, Part V 16*, pp. 125–141. Springer, 2020.
- 762 Yue Lu, Congqi Cao, Yifan Zhang, and Yanning Zhang. Learnable locality-sensitive hashing for
 763 video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 33
 764 (2):963–976, 2022.
- 765 Weixin Luo, W. Liu, and S. Gao. A revisit of sparse coding based anomaly detection in stacked rnn
 766 framework. In *Proceedings of IEEE International Conference on Computer Vision*, pp. 341–349.
 767 IEEE, 2017.
- 768 Weixin Luo, Wen Liu, Dongze Lian, Jinhui Tang, Lixin Duan, Xi Peng, and Shenghua Gao. Video
 769 anomaly detection with sparse coding inspired deep neural networks. *IEEE transactions on pat-*
 770 *tern analysis and machine intelligence*, 43(3):1070–1084, 2019.
- 771 Weixin Luo, Wen Liu, and Shenghua Gao. Normal graph: Spatial temporal graph convolutional
 772 networks based prediction network for skeleton based video anomaly detection. *Neurocomputing*,
 773 444:332–337, 2021a.
- 774 Weixin Luo, Wen Liu, Dongze Lian, and Shenghua Gao. Future frame prediction network for video
 775 anomaly detection. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):
 776 7505–7520, 2021b.
- 777 Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Learning normal dynamics in
 778 videos with meta prototype network. *arXiv preprint arXiv:2104.06689*, 2021.
- 779 Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. Unbiased multiple in-
 780 stance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF*
 781 *Conference on Computer Vision and Pattern Recognition*, pp. 8022–8031, 2023.
- 782 Neelu Madan, Arya Farkhondeh, Kamal Nasrollahi, Sergio Escalera, and Thomas B Moeslund.
 783 Temporal cues from socially unacceptable trajectories for anomaly detection. In *Proceedings of*
 784 *the IEEE/CVF International Conference on Computer Vision*, pp. 2150–2158, 2021.
- 785 Neelu Madan, Nicolae-Cătălin Ristea, Radu Tudor Ionescu, Kamal Nasrollahi, Fahad Shahbaz
 786 Khan, Thomas B Moeslund, and Mubarak Shah. Self-supervised masked convolutional trans-
 787 former block for anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelli-*
 788 *gence*, 2023.
- 789 Amir Markovitz, Gilad Sharir, Itamar Friedman, Lihi Zelnik-Manor, and Shai Avidan. Graph em-
 790 bedded pose clustering for anomaly detection. In *Proceedings of the IEEE/CVF Conference on*
 791 *Computer Vision and Pattern Recognition*, pp. 10539–10547. IEEE, 2020.
- 792 Jakub Micorek, Horst Possegger, Dominik Narnhofer, Horst Bischof, and Mateusz Kozinski. Mulde:
 793 Multiscale log-density estimation via denoising score matching for video anomaly detection.
 794 In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.
 795 18868–18877, 2024.
- 796 Romero Morais, V. Le, T. Tran, B. Saha, M. Mansour, and S. Venkatesh. Learning regularity in
 797 skeleton trajectories for anomaly detection in videos. In *Proceedings of the IEEE Conference on*
 798 *Computer Vision and Pattern Recognition*, pp. 11996–12004. IEEE, 2019.
- 799 Trong Nguyen Nguyen and J. Meunier. Anomaly detection in video sequence with appearance mo-
 800 tion correspondence. In *Proceedings of the IEEE International Conference on Computer Vision*,
 801 pp. 1273–1283. IEEE, 2019.
- 802 Aniello Panariello, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Consistency-based self-
 803 supervised learning for temporal anomaly localization. In *European Conference on Computer*
 804 *Vision*, pp. 338–349. Springer, 2022.

- 810 Guansong Pang, Cheng Yan, Chunhua Shen, Anton van den Hengel, and Xiao Bai. Self-trained
 811 deep ordinal regression for end-to-end video anomaly detection. In *Proceedings of the IEEE/CVF*
 812 *Conference on Computer Vision and Pattern Recognition*, pp. 12173–12182. IEEE, 2020.
- 813
- 814 Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly
 815 detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog-*
 816 *nition*, pp. 14372–14381. IEEE, 2020.
- 817 Lorenzo Perini, Vincent Vercruyssen, and Jesse Davis. Transferring the contamination factor be-
 818 tween anomaly detection domains by shape similarity. In *Proceedings of the AAAI Conference on*
 819 *Artificial Intelligence*, volume 36, pp. 4128–4136. AAAI Press, 2022.
- 820 Yujiang Pu, Xiaoyu Wu, and Shengjin Wang. Learning prompt-enhanced context features for
 821 weakly-supervised video anomaly detection. *arXiv preprint arXiv:2306.14451*, 2023.
- 822
- 823 Didik Purwanto, Yie-Tarng Chen, and Wen-Hsien Fang. Dance with self-attention: A new look
 824 of conditional random fields on anomaly detection in videos. In *Proceedings of the IEEE/CVF*
 825 *International Conference on Computer Vision*, pp. 173–183, 2021.
- 826 Automatic Differentiation In Pytorch. Pytorch, 2018.
- 827
- 828 Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language under-
 829 standing by generative pre-training. 2018.
- 830
- 831 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal,
 832 Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual
 833 models from natural language supervision. In *International conference on machine learning*, pp.
 8748–8763. PMLR, 2021.
- 834
- 835 Bharathkumar Ramachandra, Michael J Jones, and Ranga Raju Vatsavai. A survey of single-scene
 836 video anomaly detection. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):
 837 2293–2312, 2020.
- 838
- 839 Tal Reiss and Yedid Hoshen. Attribute-based representations for accurate and interpretable video
 840 anomaly detection. *arXiv preprint arXiv:2212.00789*, 2022.
- 841
- 842 Nicolae-Catalin Ristea, Neelu Madan, Radu Tudor Ionescu, Kamal Nasrollahi, Fahad Shahbaz
 843 Khan, Thomas B Moeslund, and Mubarak Shah. Self-supervised predictive convolutional at-
 844 tentive block for anomaly detection. *arXiv preprint arXiv:2111.09099*, 2021.
- 845
- 846 Mohammad Sabokrou, M. Khalooei, M. Fathy, and E. Adeli. Adversarially learned one-class classi-
 847 fier for novelty detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern*
 848 *Recognition*, pp. 3379–3388. IEEE, 2018.
- 849
- 850 Hitesh Sapkota and Qi Yu. Bayesian nonparametric submodular video partition for robust anomaly
 851 detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog-*
 852 *nition*, pp. 3212–3221, 2022.
- 853
- 854 Fumiaki Sato, Ryo Hachiuma, and Taiki Sekii. Prompt-guided zero-shot anomaly action recogni-
 855 tion using pretrained deep skeleton features. In *Proceedings of the IEEE/CVF Conference on*
 856 *Computer Vision and Pattern Recognition*, pp. 6471–6480, 2023.
- 857
- 858 Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh,
 859 and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based local-
 860 ization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626,
 861 2017.
- 862
- 863 Chenrui Shi, Che Sun, Yuwei Wu, and Yunde Jia. Video anomaly detection via sequentially learning
 864 multiple pretext tasks. In *Proceedings of the IEEE/CVF International Conference on Computer*
 865 *Vision*, pp. 10330–10340, 2023a.
- 866
- 867 Haoyue Shi, Le Wang, Sanping Zhou, Gang Hua, and Wei Tang. Abnormal ratios guided multi-phase
 868 self-training for weakly-supervised video anomaly detection. *IEEE Transactions on Multimedia*,
 869 2023b.

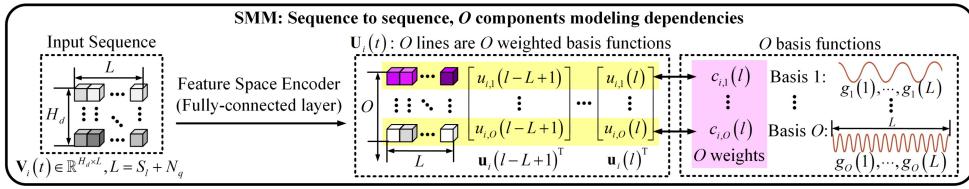
- 864 Ashish Singh, Michael J Jones, and Erik G Learned-Miller. Eval: Explainable video anomaly local-
 865 ization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog-
 866 nition*, pp. 18717–18726, 2023.
- 867 Alexandros Stergiou, Brent De Weerd, and Nikos Deligiannis. Holistic representation learning for
 868 multitask trajectory anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on
 869 Applications of Computer Vision*, pp. 6729–6739, 2024.
- 870 Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance
 871 videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*,
 872 pp. 6479–6488, 2018.
- 873 Che Sun, Yunde Jia, Yao Hu, and Yuwei Wu. Scene-aware context reasoning for unsupervised
 874 abnormal event detection in videos. In *Proceedings of the 28th ACM International Conference on
 875 Multimedia*, pp. 184–192, 2020.
- 876 Shengyang Sun and Xiaojin Gong. Hierarchical semantic contrast for scene-aware video anomaly
 877 detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog-
 878 nition*, pp. 22846–22856, 2023.
- 879 Weijun Tan, Qi Yao, and Jingfeng Liu. Overlooked video classification in weakly supervised video
 880 anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Com-
 881 puter Vision*, pp. 202–210, 2024.
- 882 Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo
 883 Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude
 884 learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp.
 885 4975–4986, 2021.
- 886 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez,
 887 Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural informa-
 888 tion processing systems*, 30, 2017.
- 889 Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-
 890 freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF
 891 conference on computer vision and pattern recognition*, pp. 7464–7475, 2023.
- 892 Guodong Wang, Yunhong Wang, Jie Qin, Dongming Zhang, Xiuguo Bao, and Di Huang.
 893 Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. *arXiv preprint
 894 arXiv:2207.10172*, 2022a.
- 895 Siqi Wang, Yijie Zeng, Guang Yu, Zhen Cheng, Xinwang Liu, Sihang Zhou, En Zhu, Marius Kloft,
 896 Jianping Yin, and Qing Liao. E3 outlier: a self-supervised framework for unsupervised deep
 897 outlier detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):2952–
 898 2969, 2022b.
- 899 Xuanzhao Wang, Zhengping Che, Bo Jiang, Ning Xiao, Ke Yang, Jian Tang, Jieping Ye, Jingyu
 900 Wang, and Qi Qi. Robust unsupervised video anomaly detection by multipath frame prediction.
 901 *IEEE transactions on neural networks and learning systems*, 33(6):2301–2312, 2021.
- 902 Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. Self-
 903 supervised sparse representation for video anomaly detection. In *European Conference on Com-
 904 puter Vision*, pp. 729–745. Springer, 2022.
- 905 Peihao Wu, Wenqian Wang, Faliang Chang, Chunsheng Liu, and Bin Wang. Dss-net: Dynamic
 906 self-supervised network for video anomaly detection. *IEEE Transactions on Multimedia*, 2023.
- 907 Peng Wu and Jing Liu. Learning causal temporal relation and feature discrimination for anomaly
 908 detection. *IEEE Transactions on Image Processing*, 30:3513–3527, 2021.
- 909 Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not
 910 only look, but also listen: Learning multimodal violence detection under weak supervision. In
 911 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020,
 912 Proceedings, Part XXX 16*, pp. 322–339. Springer, 2020.

- 918 Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, and Yanning Zhang.
 919 Open-vocabulary video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Com-*
 920 *puter Vision and Pattern Recognition*, pp. 18297–18307, 2024a.
 921
- 922 Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning
 923 Zhang. Vadclip: Adapting vision-language models for weakly supervised video anomaly detec-
 924 tion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6074–6082,
 925 2024b.
- 926 Cheng Yan, Shiyu Zhang, Yang Liu, Guansong Pang, and Wenjun Wang. Feature prediction diffu-
 927 sion model for video anomaly detection. In *Proceedings of the IEEE/CVF International Confer-*
 928 *ence on Computer Vision*, pp. 5527–5537, 2023.
- 929 Yuchen Yang, Kwonjoon Lee, Behzad Dariush, Yinzhi Cao, and Shao-Yuan Lo. Follow the
 930 rules: Reasoning for video anomaly detection with large language models. *arXiv preprint*
 931 *arXiv:2407.10299*, 2024a.
- 932 Zhen Yang, Yuanfang Guo, Junfu Wang, Di Huang, Xiuguo Bao, and Yunhong Wang. Towards
 933 video anomaly detection in the real world: A binarization embedded weakly-supervised network.
 934 *IEEE Transactions on Circuits and Systems for Video Technology*, 2023a.
- 935 Zhiwei Yang, Peng Wu, Jing Liu, and Xiaotao Liu. Dynamic local aggregation network with adap-
 936 tive clusterer for anomaly detection. In *European Conference on Computer Vision*, pp. 404–421.
 937 Springer, 2022.
- 938 Zhiwei Yang, Jing Liu, Zhaoyang Wu, Peng Wu, and Xiaotao Liu. Video event restoration based
 939 on keyframes for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on*
 940 *Computer Vision and Pattern Recognition*, pp. 14592–14601, 2023b.
- 941 Zhiwei Yang, Jing Liu, and Peng Wu. Text prompt with normality guidance for weakly supervised
 942 video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and*
 943 *Pattern Recognition*, pp. 18899–18908, 2024b.
- 944 Fei Yu, Mo Zhang, Hexin Dong, Sheng Hu, Bin Dong, and Li Zhang. Dast: Unsupervised domain
 945 adaptation in semantic segmentation based on discriminator attention and self-training. In *Pro-*
 946 , volume 35, pp. 10754–10762, 2021.
- 947 Guang Yu, Siqi Wang, Zhiping Cai, Xinwang Liu, Chuanfu Xu, and Chengkun Wu. Deep anomaly
 948 discovery from unlabeled videos via normality advantage and self-paced refinement. In *Pro-*
 949 , pp. 13987–13998,
 950 2022a.
- 951 Jiahuo Yu, Jinyu Liu, Ying Cheng, Rui Feng, and Yuejie Zhang. Modality-aware contrastive in-
 952 stance learning with self-distillation for weakly-supervised audio-visual violence detection. In *Pro-*
 953 , pp. 6278–6287, 2022b.
- 954 Shoubin Yu, Zhongyin Zhao, Haoshu Fang, Andong Deng, Haisheng Su, Dongliang Wang, Weihao
 955 Gan, Cewu Lu, and Wei Wu. Regularity learning via explicit distribution modeling for skeletal
 956 video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.
- 957 M Zaigham Zaheer, Arif Mahmood, M Haris Khan, Mattia Segu, Fisher Yu, and Seung-Ik Lee.
 958 Generative cooperative learning for unsupervised video anomaly detection. In *Proceedings of the*
 959 *IEEE/CVF conference on computer vision and pattern recognition*, pp. 14744–14754, 2022a.
- 960 Muhammad Zaigham Zaheer, Jin-ha Lee, Marcella Astrid, and Seung-Ik Lee. Old is gold: Re-
 961 defining the adversarially learned one-class classifier training paradigm. In *Proceedings of the*
 962 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14183–14193. IEEE,
 963 2020a.
- 964 Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. Claws: Cluster-
 965 ing assisted weakly supervised learning with normalcy suppression for anomalous event detection.
 966 In *European Conference on Computer Vision*, pp. 358–376. Springer, 2020b.

- 972 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.
- 973
- 974
- 975
- 976 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.
- 977
- 978
- 979 Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing 980 large language models for training-free video anomaly detection. In *Proceedings of the IEEE/CVF* 981 *Conference on Computer Vision and Pattern Recognition*, pp. 18527–18536, 2024.
- 982
- 983 Hao Zeng, Zhiyong Zhang, and Lulin Shi. Research and implementation of video codec based on 984 ffmpeg. In *2016 international conference on network and information systems for computers* (ICNISC), pp. 184–188. IEEE, 2016.
- 985
- 986 Xianlin Zeng, Yalong Jiang, Wenrui Ding, Hongguang Li, Yafeng Hao, and Zifeng Qiu. A hierar- 987 chical spatio-temporal graph convolutional neural network for anomaly detection in videos. *IEEE* 988 *Transactions on Circuits and Systems for Video Technology*, 2021.
- 989
- 990 Chen Zhang, Guorong Li, Yuankai Qi, Shuhui Wang, Laiyun Qing, Qingming Huang, and Ming- 991 Hsuan Yang. Exploiting completeness and uncertainty of pseudo labels for weakly supervised 992 video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and* 993 *Pattern Recognition*, pp. 16271–16280, 2023a.
- 994
- 995 Menghao Zhang, Jingyu Wang, Qi Qi, Haifeng Sun, Zirui Zhuang, Pengfei Ren, Ruilong Ma, and 996 Jianxin Liao. Multi-scale video anomaly detection by multi-grained spatio-temporal represen- 997 tation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern* 998 *Recognition*, pp. 17385–17394, 2024.
- 999
- 1000 Sijia Zhang, Maoguo Gong, Yu Xie, A Kai Qin, Hao Li, Yuan Gao, and Yew-Soon Ong. Influence- 1001 aware attention networks for anomaly detection in surveillance videos. *IEEE Transactions on* 1002 *Circuits and Systems for Video Technology*, 32(8):5427–5437, 2022a.
- 1003
- 1004 Xinfeng Zhang, Jinpeng Fang, Baoqing Yang, Shuhan Chen, and Bin Li. Hybrid attention and 1005 motion constraint for anomaly detection in crowded scenes. *IEEE Transactions on Circuits and* 1006 *Systems for Video Technology*, 2022b.
- 1007
- 1008 Yu Zhang, Xiushan Nie, Rundong He, Meng Chen, and Yilong Yin. Normality learning in multi- 1009 space for video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Tech-* 1010 *nology*, 31(9):3694–3706, 2020.
- 1011
- 1012 Yu-Xuan Zhang, Hua Meng, Xue-Mei Cao, Zhengchun Zhou, Mei Yang, and Avik Ranjan Ad- 1013 hikary. Interpreting vulnerabilities of multi-instance learning to adversarial perturbations. *Pattern* 1014 *Recognition*, pp. 109725, 2023b.
- 1015
- 1016 Jia-Xing Zhong, Nannan Li, Weijie Kong, Thomas H Li Shan Liu, and Ge Li. Graph convolutional 1017 label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings* 1018 *of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1237–1246. IEEE, 1019 2019.
- 1020
- 1021 Yuanhong Zhong, Xia Chen, Yongting Hu, Panliang Tang, and Fan Ren. Bidirectional spatio- 1022 temporal feature learning with multiscale evaluation for video anomaly detection. *IEEE Transac-* 1023 *tions on Circuits and Systems for Video Technology*, 32(12):8285–8296, 2022.
- 1024
- 1025 Hang Zhou, Junqing Yu, and Wei Yang. Dual memory units with uncertainty regulation for weakly 1026 supervised video anomaly detection. *arXiv preprint arXiv:2302.05160*, 2023a.
- Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic 1027 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* 1028 *Conference on Computer Vision*, pp. 395–412. Springer, 2022.

1026 A DETAILS ABOUT SMM

1028 We model the sequence $\mathbf{V}_i(t) = [\mathbf{F}_{i,1}^T(t-1); \dots; \mathbf{F}_{i,S_l}^T(t-1); \mathbf{F}_{i,1}^I(t); \dots; \mathbf{F}_{i,N_q}^I(t)]^\top \in$
 1029 $\mathbb{R}^{H_d \times (S_l + N_q)}$, $L = S_l + N_q$ of object i consisting of both S_l columns denoting the sentence from
 1030 $t-1$ 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
 1031 the Feature Space Encoder which is a fully-connected layer in Fig. 3. The l' -th column in $\mathbf{U}_i(t)$ is
 1032 $[u_{i,1}(l'), \dots, u_{i,O}(l')]^\top$, $l' \in [l-L, l]$. l varies along the column dimension of $\mathbf{V}_i(t)$ and $\mathbf{U}_i(t)$,
 1033 $(l-L, l]$ is the window of columns which are encoded by $\mathbf{c}_i(l)$ together. For simplicity, we ignore
 1034 index t in the following parts which conduct analysis along the column dimension of input tensor at
 1035 any moment t .



1044 Figure 6: Basis functions in state machines for modeling sequences.

1045 Fig. 6 shows the basis functions in state machines for modeling sequences. The O rows in
 1046 $\mathbf{U}_i(t)$ are the weighted version of the O length- L Legendre polynomials Arfken et al. (2011)
 1047 $[g_o(1), \dots, g_o(L)]$, $o \in [1, O]$. Specifically, the o -th row can be represented as:

$$1048 [u_{i,o}(l-L+1), \dots, u_{i,o}(l)] = c_{i,o}(l)[g_o(1), \dots, g_o(L)], o \in [1, O] \quad (7)$$

1049 In SMM, a state vector $\mathbf{c}_i(l) \in \mathbb{R}^O$ with O weights $c_{i,1}(l), \dots, c_{i,O}(l)$ are the weights of polynomials
 1050 and encode the dependencies among the columns in $\mathbf{V}_i(t)$.

1051 As can be seen from Fig. 6, $\mathbf{c}_i(l_1)$ encodes the dependencies among $l_1 - L + 1, \dots, l_1$ columns,
 1052 $\mathbf{c}_i(l_2)$ encodes the dependencies among $l_2 - L + 1, \dots, l_2$ columns. According to Gu et al. (2020),
 1053 the dynamics of a 1-dimensional sequence $f_i(l)$ across a period can be represented by $\mathbf{c}_i(l) \in \mathbb{R}^O$,
 1054 satisfying $[f_i(l-L+1), \dots, f_i(l)] = \sum_{o=1}^O c_{i,o}(l)[g_o(1), \dots, g_o(L)]$, $f_i(l') \in \mathbb{R}$, $l' \in [l-L+1, l]$.
 1055 The transitions from $\mathbf{c}_i(l_1)$ to $\mathbf{c}_i(l_2)$ facilitates the prediction in the sequence:

$$1056 \frac{d}{dt} \mathbf{c}_i(l) = \mathbf{A}_{HiPPO} \mathbf{c}_i(l) + \mathbf{B}_{HiPPO} f_i(l) \quad (8)$$

1057 By combining Eq. (7) with Eq. (8), we can obtain:

$$1058 \frac{d}{dt} \mathbf{c}_i(l) = \mathbf{A}_{HiPPO} \mathbf{c}_i(l) + \mathbf{B}_{HiPPO} \sum_{o=1}^O u_{i,o}(l) \quad (9)$$

1059 The matrices \mathbf{A}_{HiPPO} and \mathbf{B}_{HiPPO} are defined in Gu et al. (2020) with $o, h \in [1, O]$:

$$1060 \mathbf{A}_{HiPPO}(o, h) = \begin{cases} -\frac{(2o+1)^{0.5}(2h+1)^{0.5}}{L} & \text{if } o > h, \\ 0 & \text{if } o < h, \\ -\frac{o+1}{L} & \text{if } o = h. \end{cases} \quad (10)$$

$$1061 \mathbf{B}_{HiPPO}(o) = -\frac{(2o+1)^{0.5}}{L} \quad (11)$$

1062 To discretize Eq. (9), we obtain

$$\begin{aligned}
& \lim_{\Delta \rightarrow 0} \frac{\mathbf{c}_i(l + \Delta) - \mathbf{c}_i(l)}{\Delta} = \lim_{\Delta \rightarrow 0} \left(\frac{\mathbf{A}_{HiPPO} \mathbf{c}_i(l) + \mathbf{B}_{HiPPO} \sum_{o=1}^O u_{i,o}(l)}{2} + \right. \\
& \quad \left. \frac{\mathbf{A}_{HiPPO} \mathbf{c}_i(l + \Delta) + \mathbf{B}_{HiPPO} \sum_{o=1}^O u_{i,o}(l + \Delta)}{2} \right), \Delta = 1
\end{aligned} \tag{12}$$

which can be transformed to

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

which simplifies to

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

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

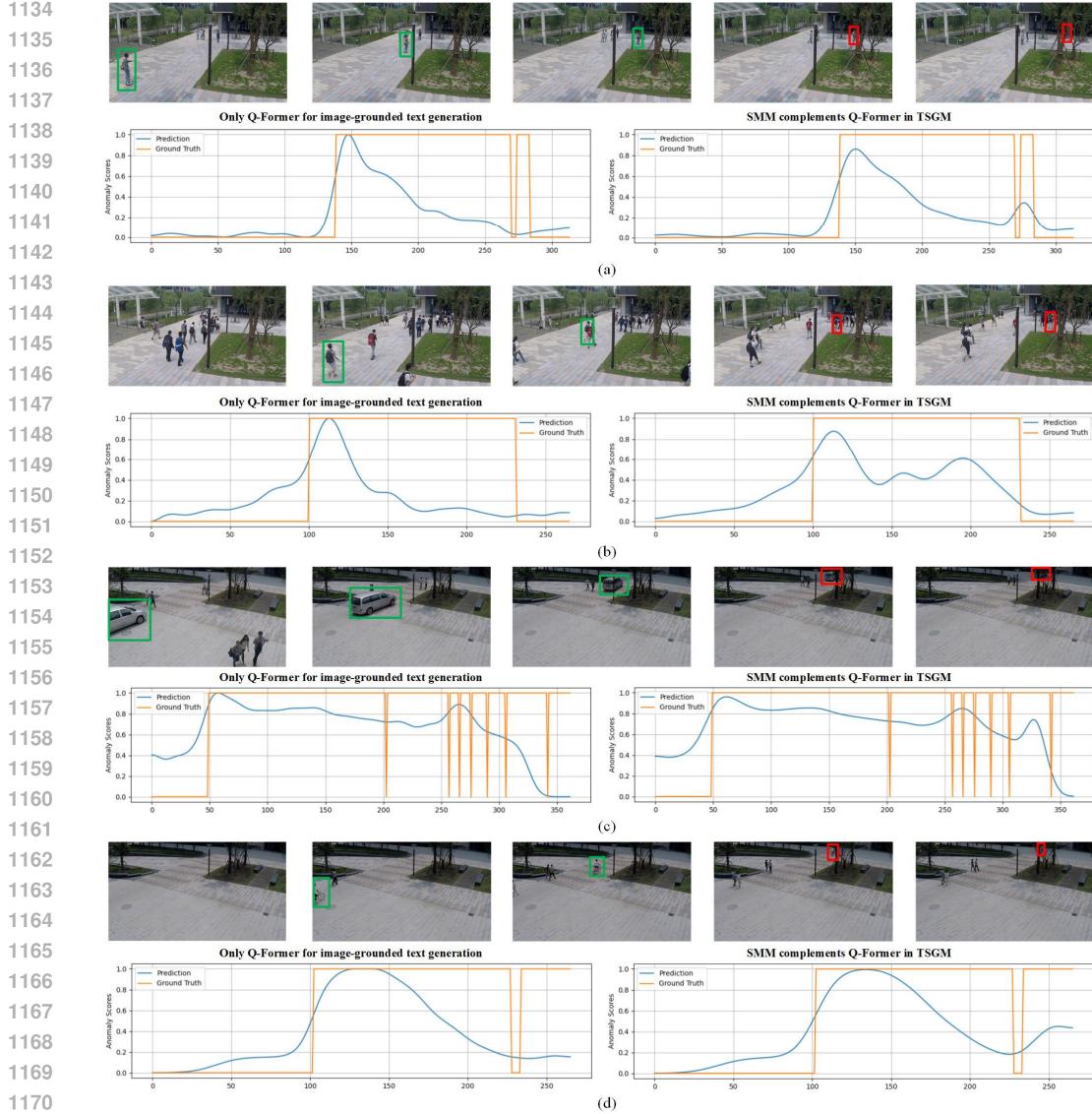
In Fig. 7, the green bounding boxes indicate the anomalies that can be detected by directly applying the Q-Former, as described in Li et al. (2023), for image-grounded text generation in TSGM. The red bounding boxes show the cases where only with the combination of SMM and Q-Former in TSGM can the anomalies 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 $\mathbf{E}_j^I(t) = [\mathbf{E}_{j,1}^I(t), \mathbf{E}_{j,2}^I(t), \dots, \mathbf{E}_{j,S_l}^I(t)] \in \mathbb{R}^{S_l \times H_d}$ in a sentence with maximum length S_l attend to each other. $S_l = N_q$ and $\mathbf{E}_{j,1}^I(t) \in \mathbb{R}^{H_d}, \mathbf{E}_{j,2}^I(t) \in \mathbb{R}^{H_d}, \dots, \mathbf{E}_{j,S_l}^I(t) \in \mathbb{R}^{H_d}$.



1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187

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 $\mathbf{S}_{i,1}(t), \mathbf{S}_{i,2}(t), \dots, \mathbf{S}_{i,k}(t)$ followed by $S_l - k$ zero placeholders, producing embeddings $\mathbf{E}_{i,1}^T(t), \mathbf{E}_{i,2}^T(t), \dots, \mathbf{E}_{i,S_l}^T(t)$. The visual embeddings are concatenated with text tokens, producing

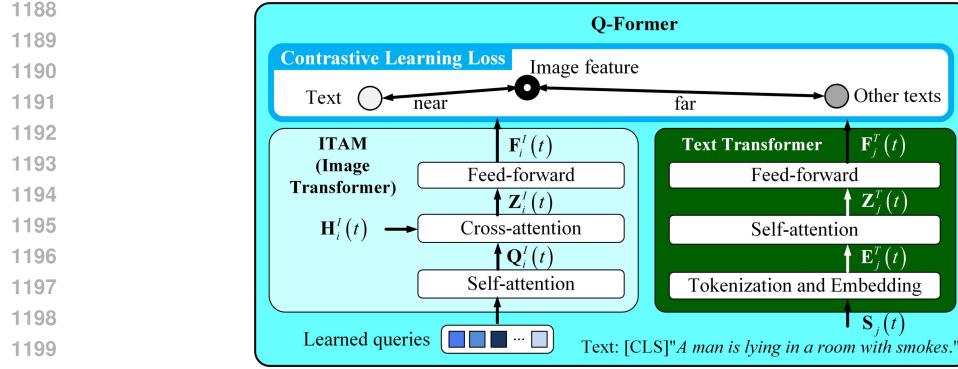


Figure 8: ITAM is the image transformer in Q-Former Li et al. (2023), it identifies the local features through aligning the visual features from image transformer with textual features from text transformer.

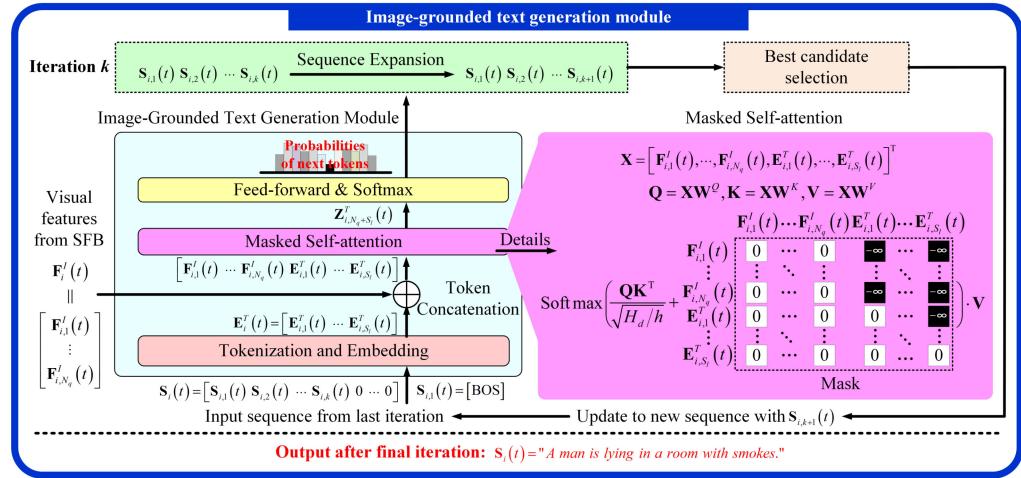


Figure 9: Structure of the module for image-grounded text generation, the module is part of the Q-Former Li et al. (2023).

$\mathbf{X} = [\mathbf{F}_{i,1}^I(t), \mathbf{F}_{i,2}^I(t), \dots, \mathbf{F}_{i,N_q}^I(t), \mathbf{E}_{i,1}^T(t), \mathbf{E}_{i,2}^T(t), \dots, \mathbf{E}_{i,S_l}^T(t)]^T$ 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

$$\mathbf{Z}_i^T(t) = \text{Softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{H_d/h}} + \mathbf{M}\right)\mathbf{V} \quad (15)$$

where the values in mask \mathbf{M} are shown by black and white rectangles in Fig. 9. $h = 12$ denotes the number of heads. $\mathbf{Z}_i^T(t) = [\mathbf{Z}_{i,1}^T(t), \dots, \mathbf{Z}_{i,N_q+S_l}^T(t)]^T \in \mathbb{R}^{(N_q+S_l) \times H_d}$. Only the last token $\mathbf{Z}_{i,N_q+S_l}^T(t)$ 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 $N_{\text{vocabulary}}$ output channels, producing $N_{\text{vocabulary}} = 30,523$ probabilities indicating the likelihood of candidate tokens. $N_{\text{vocabulary}}$ 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 $\mathbf{S}_{i,1}(t), \mathbf{S}_{i,2}(t), \dots, \mathbf{S}_{i,k}(t)$ 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 training set of anomaly detection benchmark 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.

1296
 1297 Table 5: Ablations of ITAM’s structure using Micro-AUC. TME and TSG are short for Temporal
 1298 Motion Estimation and Temporal Sentence Generation, respectively. N_q is the number of image
 1299 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

1315 As is shown in Fig. 2(b), TSGM firstly converts the declarative sentence "The man is pushing a
 1316 stroller on the street." to an interrogative sentence "Is the man pushing a stroller on the street?"
 1317 The conversion is based on nltk library Hardeniya et al. (2016) and the procedures are shown in
 1318 Algorithm 2:

Algorithm 2 Algorithm for Converting Declarative Sentences to Interrogative Sentences

- 1: Input sentence: $D \leftarrow S_i^I(t-1) = \text{'The man is pushing a stroller on the street.'}$
 - 2: Tokenization: $D \rightarrow [\text{'The', 'man', 'is', 'pushing', 'a', 'stroller', 'on', 'the', 'street', '.'}]$
 - 3: Locate first verb: $D_{firstverb} = \text{'is'}$
 - 4: Divide sentence using first verb: $D \rightarrow D_1 + D_{firstverb} + D_2$, $D_1 = \text{'The man', } D_2 = \text{'pushing a stroller on the street'}$
 - 5: Change the order of parts: $Q \leftarrow D_{firstverb} + D_1 + D_2$
 - 6: **return** Q
-

H OPERATIONAL EFFICIENCY

1334 All experiments are conducted on an NVIDIA A100 GPU and an Intel(R) Xeon(R) Gold 6248R
 1335 CPU. For object detection, we have employed both YOLOv7 detector Wang et al. (2023) and another
 1336 detector based on Qwen-VL-7B Bai et al. (2023). As is shown by Table 4, both detectors achieve
 1337 similar accuracy. In terms of inference speed, the YOLO detector Wang et al. (2023) processes
 1338 each frame in 1.5 milliseconds, whereas Qwen-VL-7B Bai et al. (2023) requires 5.2 seconds per
 1339 frame. Consequently, we evaluate the operational efficiency of the proposed framework using the
 1340 YOLO detector. The inference times of all components in the proposed framework are measured and
 1341 summarized in Table 6. With all components considered, the proposed method achieves an average
 1342 frame rate of 12 FPS with an average of 5 objects per frame. The average number of 5 objects is
 1343 based on the findings of Wang et al. (2022a).

1344 In the future, we aim to explore the methods that utilize a fixed number of bounding boxes per frame
 1345 to maintain a constant inference time, even with an increased number of objects. Additionally, we
 1346 will investigate parsing multiple objects within a single bounding box to maintain a fixed number of
 1347 bounding boxes per frame.

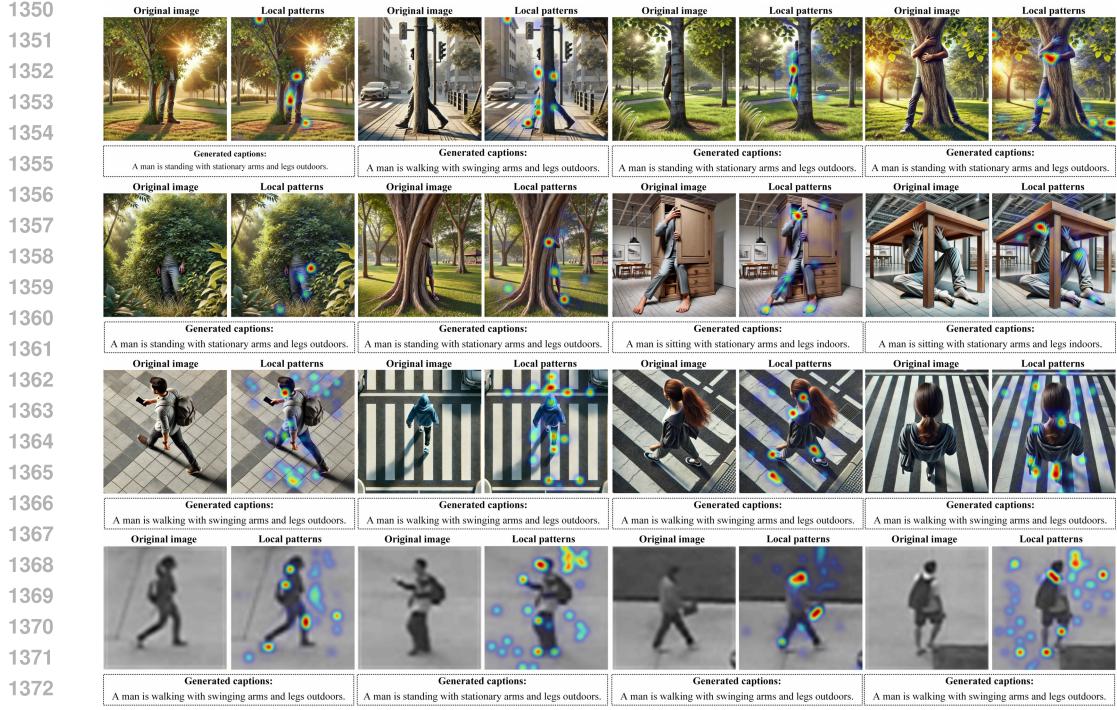


Figure 10: Visualization of the heatmaps of local patterns under occlusions, viewpoint changes and low resolutions.

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

1404 **I VISUALIZATION OF LOCAL PATTERNS UNDER OCCLUSIONS AND**
1405 **VIEWPOINT CHANGES**

1407 In real-world surveillance videos, occlusions, viewpoint variations and low-resolution conditions are
1408 common. Fig. 10 shows some examples of the local patterns identified by image-text alignment and
1409 cross-modality attention. The local patterns capture semantically meaningful features such as body
1410 joints which are consistent across the variations. The compact representations ignore redundant
1411 details and contribute to generalizable embeddings.

1413 **J FUTURE WORK**

1415 One limitation of the current framework is the reliance on object detectors. Currently, the perfor-
1416 mance of current Vision-Language Models (VLMs) is limited by their fields of view. For example,
1417 when processing an image with a large scene, a vision-language model tends to overlook many
1418 details, highlighting the necessity of object detectors that facilitate the processing of local regions
1419 independently. Table 4 shows that object detectors significantly outperform sliding windows, the
1420 poor performance of the latter may result from an incorrect strategy. As a result, we will try more
1421 efficient and effective ways to parse events in complex scenes and images with large fields of view
1422 where many objects reside. Specifically, we will explore the integration of object detectors in an
1423 end-to-end large model. In simpler scenes with fewer objects, an input image is embedded with
1424 fewer vision tokens. As scenes become more complex, more objects are involved, then an input
1425 image is encoded with an increased number of vision tokens each of which describes one or more
1426 objects. Besides, we will explore ways to improve efficiency.

1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457