2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Rule Augmentation and Perception Smoothing for Training-free Video Anomaly Detection with LLMs

Dongliang Zhao[1[0000-0003-3832-7773]], Bo Sun[*1,2[0000-0003-1168-1051]],
Jun He[*1,2[0000-0002-3017-2108]], Li Yuan[2],
Mingyang Yue[1[0000-0003-2266-3481]], and Zhichao Wu[1]

[1] School of Artificial Intelligence
Beijing Normal University, Beijing 100875, China
[2] College of Education for the Future
Beijing Normal University, Zhuhai 519087, Guangdong, China
[*] Corresponding authors
{tosunbo, hejun}@bnu.edu.cn

Abstract. Video Anomaly Detection (VAD) is widely applied in the field of public safety. Recently, training-free video anomaly detection based on large language models (LLMs) has achieved remarkable progress. However, while pretrained LLMs in previous methods contain rich general-domain knowledge, they often lack a nuanced understanding of domain-specific knowledge, leading to reduced performance in specific scenarios, such as campus environments. Furthermore, these methods often overlook the temporal consistency and motion continuity between anomalous video frames when utilizing LLMs for score judgment. To address these challenges, we propose a method for video anomaly detection using rule augmentation and perception smoothing. Specifically, the rule augmentation strategy can automatically generate anomaly detection rules based on the management standards of various scenarios. Perception smoothing employs an adaptive temporal smoothing strategy to enhance the robustness of score judgment based on LLMs. Extensive experiments demonstrate that the proposed method not only outperforms state-of-the-art, training-free methods on general datasets such as UCF-Crime and XD-Violence, but also achieves significant improvements on the specific scenario dataset ShanghaiTech.

Keywords: Video Anomaly Detection, Large Language Models, Training-free, Perception Smoothing, Rule Augmentation.

1 Introduction

Video Anomaly Detection (VAD) is a critical area within the domains of artificial intelligence and computer vision, aiming to accurately identify abnormal events from large-scale video streams [1-4]. Anomalies in videos are contextually defined based on real-world scenarios and settings, typically classified into two categories. The first category involves normal behaviors or events occurring in restricted locations within a specific scene, such as cycling or rollerblading on a sidewalk [26]. The second category

includes abnormal behaviors or events that can occur at any location within a scene, such as fires or explosions [27]. Due to its significant application potential in public safety and video content analysis, video anomaly detection has been extensively studied in recent years.

Most existing video anomaly detection methods rely on training to ensure accuracy, which significantly limits their generalization capability. VAD models trained on specific datasets often perform poorly when applied to videos captured in different environmental conditions (e.g., daylight versus nighttime settings). Another closely related challenge in VAD is data collection, particularly in certain application domains such as video surveillance, where privacy concerns may severely hinder effective data acquisition. Therefore, the aforementioned training-based video anomaly detection method faces significant challenges in terms of generalization ability and data collection. To address these issues, Luca et al. [13] leveraged the powerful prior knowledge capabilities of LLMs, combined with vision-language models, to propose the first training-free video anomaly detection method.

However, several limitations persist with this approach. First, the implicit knowledge embedded in pre-trained LLMs tends to focus on general-domain knowledge, lacking a nuanced understanding of specific domain knowledge. In practical applications, specific domain scenarios usually follow fixed security management protocols, which implicitly define the rules and requirements for anomaly detection. In other words, there is a misalignment between the anomalies as understood by LLMs and the anomaly definitions required in specific scenarios. For instance, GPT-4 typically classifies 'skateboarding' as a normal activity, but in certain high-safety-demanding contexts, such as sidewalks on a campus, it should be considered an anomalous activity. Nevertheless, fine-tuning the LLM for each specific application to incorporate such domain knowledge is costly. Therefore, flexible prompting strategies are needed to guide the LLM in adapting to various VAD tasks. Secondly, when scoring anomalies based on descriptive inputs, LLMs typically evaluates the description of each frame in isolation. In fact, from a theoretical standpoint, video content typically manifests stable patterns over time, wherein successive frames exhibit temporal coherence and continuity of motion. For instance, in frame-level descriptions generated by a video captioning model, consecutive abnormal frames might individually be labeled as normal when assessed separately, even though they collectively represent a consistent anomaly.

To address the challenges of LLMs lacking domain-specific knowledge and failing to account for the temporal consistency and motion continuity between anomalous video frames, we propose a rule augmentation and perception-smoothing video anomaly detection method. We design a set of adaptive templates capable of automatically generating anomaly detection rules based on the management protocols specific to each scenario, thereby producing a well-defined set of anomaly rules. Additionally, an adaptive temporal smoothing strategy is employed to enhance the robustness of anomaly judgments made by LLMs.

The contributions of this work are summarized as follows:

1) We designed a series of prompt templates that automatically generate anomaly detection rules based on the management specifications of different scenarios. These

templates enhance the ability of large language models to understand domain-specific knowledge, thereby enabling more effective anomaly detection.

2) When performing anomaly detection on each frame, the large language model does not consider the temporal consistency and motion continuity between frames. To address this limitation, we introduce an adaptive time smoothing strategy, which enhances the robustness of anomaly detection.
3) We propose a rule augmentation and perception-smoothing video anomaly detection method. The proposed method not only outperforms state-of-the-art training-free methods on the UCF-Crime and XD-Violence crime datasets, but also achieves significant improvements on the scene-specific ShanghaiTech dataset.

2 Related Works

VAD aims to identify anomalous frames within untrimmed long videos. Recently, deep learning methods have dominated the field of VAD and can be broadly categorized into weakly supervised [5-10], unsupervised [11-12, 15-18], one-class [19-22], and fully supervised approaches [23]. Unsupervised methods train exclusively on normal videos to learn normal patterns and are typically designed as reconstruction-based, prediction-based, or hybrid approaches. Some methods have also explored a fully unsupervised [18] setting, including both normal and anomalous videos without labels in the training set. Weakly supervised methods leverage normal and anomalous videos annotated at the video level, utilizing multi-instance learning loss during training [6]. These methods are popular due to their reduced annotation time and relatively high effectiveness. One-class methods rely solely on normal data during training. Most of these approaches learn normal patterns through self-supervised pretext tasks, operating under the assumption that the model performs poorly on anomalous data. Fully supervised methods, which require precise frame-level annotations, are less common due to the high cost of annotation. However, these training-based anomaly detection methods face significant challenges related to generalization and data collection, which severely limit their applicability.

Recently, with the advent of LLMs, Luca et al. [13] leveraged LLMs and vision-language models (VLM) to address temporal anomaly detection in videos, introducing the first VAD method that requires neither training nor data collection. This method employs a captioning model to extract captions from video frames and designs prompts for LLMs to provide anomaly scores. However, the implicit knowledge acquired during LLMs pretraining primarily focuses on general-domain knowledge, lacking a deep understanding of domain-specific information. As a result, the method lacks flexibility and accuracy for diverse real-world scenarios. Moreover, it fails to consider the temporal consistency and motion continuity between anomalous video frames. To address these issues, we propose a rule augmentation and perception-smoothing video anomaly detection method. This approach designs a set of adaptive templates to automatically generate anomaly detection rules tailored to the management protocols of different scenarios. Additionally, it incorporates an adaptive temporal smoothing strategy to enhance the robustness of anomaly judgments.

3 Method

As shown in Fig. 1, the overall framework of our method is presented. The visual description module utilizes a VLM that has not been fine-tuned on specific datasets. The rule generation module generates anomaly detection rules using LLMs, based on safety management protocols and designed prompts. The LLMs then evaluate the generated descriptions to assign anomaly scores. Finally, to enhance the robustness of these evaluations, an adaptive moving average strategy is applied. The following sections provide a detailed explanation of each module and the strategies employed.

3.1 Rule Generation

In the domain of natural language processing, researchers have effectively implemented rule-based methodologies for specific tasks, yielding substantial outcomes [24]. In real visual world applications, each specific scenario, such as hospitals, supermarkets, different classrooms within schools, laboratories, and playgrounds, operates under distinct safety management protocols. These protocols implicitly define rules governing prohibited anomalous activities within these environments. Due to the variability of scenarios, the types of anomalous events are virtually limitless. Moreover, the implicit knowledge acquired during the pretraining of LLMs primarily focuses on general contexts and lacks a deep understanding of domain-specific knowledge. Therefore, it is necessary to design a framework that automatically generates anomaly rules based on safety management protocols, effectively guiding LLMs in anomaly detection and judgment.

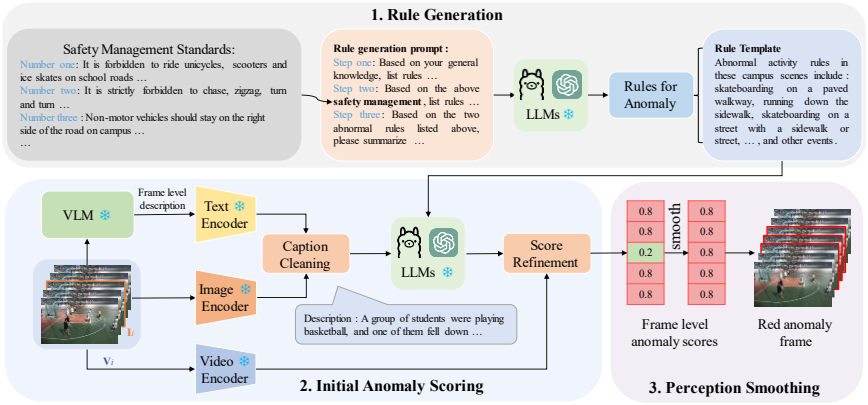


Fig. 1. Method overview. This study introduces two strategies: the rule augmentation strategy and the perception smoothing strategy. First, based on safety management specifications and a designed prompt template, the LLMs is employed to generate anomaly detection rules tailored to the specific scenario. Subsequently, the VLM is used to obtain descriptions for each video frame, with the Text Encoder and Image Encoder models of ImageBind employed to generate high-quality captions. The initial anomaly score is computed by the LLMs model, which uses the previously generated captions and anomaly rules. In the next step, the refined anomaly score is derived by utilizing the Video Encoder and Text Encoder models of ImageBind. Finally, an adaptive sliding window smoothing strategy is applied to the obtained anomaly scores, enhancing the robustness of the LLMs-based assessments.

Specifically, we first exploit the implicit knowledge embedded within the LLMs to extract anomaly management rules, denoted as Nnorm_1. The LLMs utilized in this work is GPT-4. A prompt, denoted as $\mathbf{P}_\text{l}$, is provided to facilitate this process: "*Based on your general knowledge, list rules regarding prohibited human anomalous activities on campus sidewalks. Do not explain the activities after listing them; simply provide the prohibited activities directly*." The output from the LLMs is denoted as $\mathbf{R}_\text{l}$:

$$\mathbf{R}_1 = \{LLMs(Nnorm_1,\ \mathbf{P}_l)\}, \tag{1}$$

Next, based on the ShanghaiTech dataset, we refer to its safety management protocols for school sidewalks. A prompt $\mathbf{P}_\text{r}$ is provided: "*Based on the above management regulations, list rules regarding prohibited human anomalous activities on campus sidewalks. Do not explain the activities after listing them; simply provide the prohibited activities directly*." This protocol and the prompt $\mathbf{P}_\text{r}$ are input into GPT-4 to generate Nnorm_2, with the output denoted as:

$$\mathbf{R}_2 = \{LLMs(Nnorm_2,\ \mathbf{P}_r,\ and\ Norm)\}, \tag{2}$$

Finally, the two sets of rules, $\mathbf{R}_1$ and $\mathbf{R}_2$, along with a new prompt $\mathbf{P}$, are input into GPT-4 to consolidate them into the required anomaly rule format. The prompt $\mathbf{P}$ is: "*Based on the anomalous activities you previously listed using your general knowledge and the activities listed based on the 'Management Regulations,' provide a simplified list of prohibited human activities on campus sidewalks. Avoid adding any explanations or numbering. Simply separate the activity names with commas. For example: 'Anomalous activities on campus sidewalks include: ***, ***.' Please provide your answer in this format*."

$$\mathbf{R} = \{LLMs(Nnorm,\ Nnorm_1,\ Nnorm_2,\ and\ \mathbf{P})\}, \tag{3}$$

3.2 Initial Anomaly Scoring

As shown in Fig. 1, after the rule generation module exports a set of robust rules, the initial Anomaly Scoring module initializes the anomaly score for each frame of data. Next, we will provide a detailed explanation of the specific process of this module.

Following the setup of LAVAD, given a test video $\mathbf{V} = [\text{I}_1, \ldots, \text{I}_\text{M}]$, descriptions for each frame I are first generated using the VLM model BLIP2 [29]. Subsequently, a visual-language encoder is used to assign the most semantically relevant caption to each frame.

$$\hat{C}_i = \arg\max_{C \in \mathbf{C}} \langle \varepsilon_I(\mathbf{I}_i) \cdot \varepsilon_T(C) \rangle, \tag{4}$$

Where, $\langle \cdot, \cdot \rangle$ is the cosine similarity, ε_I and ε_T are image encoder and text encoders of the VLM , and $\mathbf{C} = [\text{C}_1, \ldots, \text{C}_\text{M}]$ is a sequence of captions by the BLIP2 model.

Due to the lack of temporal information in the obtained frame-level descriptions, we leverage an LLMs to summarize temporary summaries. Specifically, we define a time window of $\mathbf{T}$ seconds centered around frame $\mathbf{I}_\text{i}$. Within this window, we uniformly

sample **N** frames to form a video segment $\mathbf{V}_i$. Subsequently, we query the LLMs using and the prompt $\mathbf{P}_S$ to generate a temporal summary $\mathbf{S}_i$ centered on frame $\mathbf{I}_i$.

$$S_i = \Phi_{\mathbf{LLMs}}(\mathrm{P}_s \circ \hat{\mathbf{C}}_i) \tag{5}$$

The prompt $\mathbf{P}_S$ is defined as: "Please summarize what happens in the following scene in a few sentences, focusing on the temporal description without including any unnecessary details or descriptions."

The anomaly score estimation is treated as a classification task, where the Φ_{LLMs} is required to select a single score from a list of 11 uniformly sampled values within the interval [0,1], with 0 indicating normality and 1 indicating an anomaly. This approach results in a textual description Si that is semantically and temporally richer than $\mathbf{C}_i$. The rule **R** we obtained in sec. 3.1 is then fed into the LLMs along with prompt to evaluate the anomaly score. The resulting scores are:

$$a_i = \Phi_{\mathbf{LLMs}}(\mathrm{P}_C \circ \mathrm{R} \circ \mathrm{P}_F \circ S_i) \tag{6}$$

Where, $\mathbf{P}_C$ is defined as: "If you are a law enforcement agency, please score the semantics of the described scene according to the abnormal activity rules. It ranges from 0 to 1, where 0 represents the standard scenario and 1 represents the scenario that conforms to the abnormal activity rules. The description semantics conform to the abnormal activity rules, and the score is 1 or close to 1." while $\mathbf{P}_F$ provides information about the desired output format.

The scores are further refined by aggregating them from semantically similar frames using visual information. Specifically, we encode the video segment $\mathbf{V}_i$ centered on frame $\mathbf{I}_i$ using ε_V , and encode all temporal summaries using ε_T . We define $\mathbf{K}_i$ as the index set of the K-closest temporal summaries $\{S_1, \ldots, S_M\}$ that are most similar to $\mathbf{V}_i$, where the similarity between $\mathbf{V}_i$ and a summary $\mathbf{S}_j$ is measured by cosine similarity, i.e. $\langle \varepsilon_V(\mathbf{V}_i), \varepsilon_T(S_j) \rangle$. We obtain the refined anomaly score:

$$a_i = \sum_{k \in K_i} a_k \cdot \frac{e^{\langle \varepsilon_V(\mathbf{V}_i), \varepsilon_T(S_k) \rangle}}{\sum_{k \in K_i} e^{\langle \varepsilon_V(\mathbf{V}_i), \varepsilon_T(S_k) \rangle}} \tag{7}$$

Where, $a = [a_1, \cdots, a_M]$ denotes the anomaly score of the video, and $\langle \cdot, \cdot \rangle$ represents the cosine similarity.

3.3 Perception Smoothing

In the process of score evaluation by LLMs, they primarily rely on the current descriptive information to assign scores, often overlooking the temporal consistency and action continuity of anomalous video frames. This can, to some extent, limit the comprehensiveness and accuracy of the evaluation. To enhance classification robustness, existing studies have explored and implemented random smoothing strategies [25], which have been shown to be quite effective. Therefore, to further improve the robustness of

anomaly score judgment, we have introduced an innovative approach, utilizing an adaptive moving average strategy for the obtained anomaly scores.

Based on the predefined ratio, we calculate the size of the initial window.

$$w = \max(\lceil len(a) \cdot r \rceil, 3) \tag{8}$$

Where , r is the ratio of window size to data length. Then the smoothed score can be expressed as:

$$x_i = \frac{1}{w} \sum_{j=i-w/2}^{i+w/2} a_j \tag{9}$$

Where, the form takes xi as the center and takes ⌊w/2⌋ points before and after to average. For data on array boundaries, the window size may be adjusted to accommodate the range of data.

4 Experiments

4.1 Datasets

In our study, we utilized three widely used VAD datasets: ShanghaiTech [26], UCF-Crime [27], and XD-Violence [28]. The ShanghaiTech dataset comprises 437 videos collected from multiple surveillance cameras within a university campus. It captures 130 anomalous events across 13 different scenes, covering 17 anomaly classes. Following the configuration of Zhong et al., the dataset is divided into 238 training videos and 199 testing videos. In our study, we used only the 199 testing videos. The UCF-Crime dataset is a large-scale dataset of real-world surveillance videos, consisting of 1900 untrimmed videos that represent 13 real-world anomalies with significant implications for public safety. The training set includes 800 normal videos and 810 anomalous videos, while the test set comprises 150 normal and 140 anomalous videos. The XD-Violence dataset is a large-scale dataset designed for violence detection, containing 4,754 untrimmed videos with audio signals and weak labels. Of these, 3,954 videos were used for training, and 800 videos were designated for testing. The dataset has a total duration of 217 hours, covering various scenes and six anomaly categories.

4.2 Experimental Details

To enhance computational efficiency, we follow the configuration of the LAVAD method, sampling each video every 16 frames. BLIP-2 [29] is used as the captioning module ΦC, and Llama-2-13b-chat [30] serves as the LLMs module Llama-2-13b-chat. For multi-modal encoding, we utilize the pre-trained multi-modal encoder from ImageBind [31]. Specifically, the temporal window is set to **T**=10 seconds, consistent with the pretraining of the ImageBind video encoder. We use **K**=10 for the textual representation S of the video. Additionally, GPT-4 is utilized as the LLMs for rule summarization. The adaptive window ration for the moving average strategy is set to 0.15.

4.3 Main Results

We conducted experiments using the UCF-Crime and XD-Violence datasets and compared the algorithm with the latest untrained video anomaly detection methods. Since the events in these datasets are highly recognizable and large language models possess rich implicit knowledge of dangerous events in general scenarios, as shown in Fig. 2, the image accurately represents that the red area falls within the anomalous rule. However, in the predicted description, the large model does not provide a very precise description but includes the word "thrown" If the model were to rely solely on its implicit knowledge, the anomaly score would be 0.8, whereas with the rule added, the anomaly score drops to 0.4. This is because the inclusion of the rule restricts the model's anomaly detection ability. Therefore, for general datasets, we did not introduce additional rules.

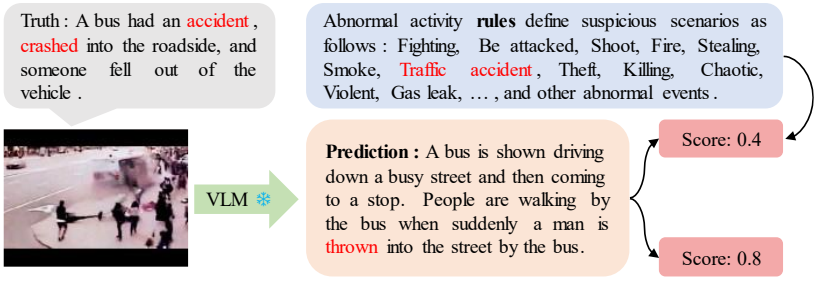


Fig. 2. Analysis of video scores after adding rules to the UCF-Crime dataset.

As shown in Table 1, our method is compared with state-of-the-art weakly supervised, unsupervised, and training-free methods on the XD-Violence dataset. It is worth noting that training-free VAD is a challenging task. Our results show that compared to the state-of-the-art train-free LAVAD method, AP improved by +2.05%, and AUC increased by nearly one point. Similarly, as shown in Table 2, our method is compared with state-of-the-art weakly supervised, unsupervised, and training-free methods on the UCF-Crime dataset. It can be observed that our method outperforms the state-of-the-art training-free methods with an improvement of +2.18%.

Table 1. Comparison with state-of-the-art training-free methods on the XD-Violence dataset.

	Method	Backbone	AP(%)	AUC(%)
Weakly-supervised	VADCLIP (AAAI'24) [5]	ViT	84.51	-
	PE-MIT (CVPR'24) [6]	I3D	88.05	-
Unsupervised	RAREANOM [15] (Pattern Recognit)	I3D	-	68.33
Training-free	ZS CLIP (ICML'21) [32]	ViT	17.83	38.21
	ZS IMAGEBIND (CVPR'23) [29]	ViT	25.36	55.06
	LLAVA-1.5 (CVPR'24) [33]	ViT	50.26	79.62
	LAVAD (CVPR'24) [13]	ViT	62.01	85.36
	TRLVAD (Ours)	ViT	**64.06**	**86.16**

As shown in Table 3, we compare our method with state-of-the-art unsupervised and training-free LAVAD video anomaly detection methods. Unlike the results in Table 1

and Table 2, Table 3 reveals that without incorporating rules and the adaptive moving average strategy, the accuracy of our method is only 53.51%. This significant drop in accuracy is attributed to the lack of in-depth understanding of domain-specific knowledge by large models in certain scenarios. In response, we further introduced rules and the moving average strategy, resulting in a substantial improvement, with the accuracy increasing by nearly 13 percentage points.

Table 2. Comparison with state-of-the-art training-free methods on the UCF-Crime dataset.

	Method	Backbone	AUC(%)
Weakly-supervised	VadCLIP (AAAI'24) [5]	ViT	88.02
	PE-MIT (CVPR'24) [6]	I3D	86.83
Unsupervised	DYANNET (WACV'23) [16]	I3D	79.76
	LANP-UVAD (ECCV'24) [11]	I3D	80.02
Training-free	ZS CLIP (ICML'21) [32]	ViT	53.16
	ZS IMAGEBIND (CVPR'23) [31]	ViT	55.78
	LLAVA-1.5 (CVPR'24) [23]	ViT	72.84
	LAVAD (CVPR'24) [13]	ViT	80.28
	TRLVAD (Ours)	ViT	**82.46**

Table 3. Comparison with state-of-the-art training-free methods on the ShanghaiTech dataset.

	Method	Backbone	AUC(%)
Unsupervised	GCL(CVPR'22) [17]	R3D	79.60
	FPDM (ICCV'23) [18]	Image	78.60
	DiffVAD (IJCAI'24) [12]	Image	81.90
Training-free	LAVAD (CVPR'24) [13]	ViT	53.51
	TRLVAD (Ours)	ViT	**66.43**

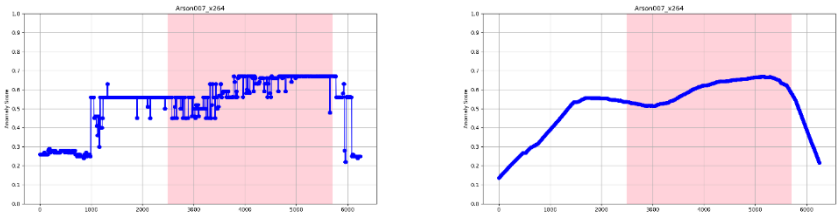
4.4 Ablation Study

As shown in Table 4, we conducted ablation experiments on the ShanghaiTech dataset to evaluate the impact of different strategies on video anomaly detection accuracy. The experiments explored the effects of introducing rules, the adaptive moving average strategy, and the simultaneous application of both strategies. The data in the Table 4 indicates that adding rules alone improved the accuracy by 6 percentage points, while applying the adaptive moving average strategy alone resulted in a 4-percentage-point increase in accuracy. Notably, when both rules and the adaptive moving average strategy were introduced simultaneously, the accuracy significantly increased to 66.43%. These results strongly validate the effectiveness of the proposed strategies in enhancing video anomaly detection accuracy.

Table 4. Results of Ablation Experiments on the ShanghaiTech Dataset.

Rule	Smoothing Strategy	AUC(%)
×	×	53.51
√	×	59.01
×	√	57.09
√	√	**66.43**

Fig. 3 shows some qualitative results on UCF-Crime. X represents the time sequence, with the subscript indicating the frame number, while the y-axis represents the anomaly score for each frame. The closer the score is to 1, the more anomalous the frame is. The pink background highlights the time period of the anomaly event in the video, and the blue line represents the anomaly score for each frame. In the red-background area, some anomalous frames have lower scores, while some non-anomalous frames have higher scores. However, by introducing the moving average smoothing strategy, this issue is significantly mitigated, resulting in final scores that better align with the temporal consistency and action continuity between video frames.



(a) The moving smoothing strategy was not employed. (b) The moving smoothing strategy was applied.

Fig. 3. Visualizations of frame-level anomaly scores for test videos from UCF-Crime.

5 Conclusions

In this study, we focus on the limitations of LLMs, specifically their lack of domain-specific knowledge and their neglect of temporal consistency and motion continuity during score assessment. We propose a video anomaly detection method that combines rule augmentation and perception smoothing. The method automatically generates anomaly detection rules using adaptive templates and introduces an adaptive temporal smoothing strategy to enhance the robustness of LLMs-based score judgments. Experimental results demonstrate that the proposed method not only outperforms state-of-the-art training-free methods on general datasets such as UCF-Crime and XD-Violence, but also achieves significant improvements on the scenario-specific dataset, ShanghaiTech. In future research, we will focus on optimizing VLMs to obtain more accurate video descriptions. This improvement aims to provide a more reliable descriptive foundation for large models during score assessments.

Acknowledgment. This work is supported by the National Natural Science Foundation of China (Grant No. 62177006, 62077009), and Guangdong Provincial Natural Science Foundation (Grant

No. 2025A1515010136, 2022A1515011541); and partially supported by the Guangdong Province Undergraduate Course Teaching and Research Office Construction Project (Grant No. jx2022303). Besides, this work is supported by the Interdisciplinary Intelligence Super Computer Center of Beijing Normal University at Zhuhai.

References

1. Ramachandra, B., Jones, M.J., Vatsavai, R.R.: A survey of single-scene video anomaly detection. IEEE transactions on pattern analysis and machine intelligence 44(5), 2293–2312 (2020)
2. Suarez, J.J.P., Naval Jr, P.C.: A survey on deep learning techniques for video anomaly detection. arXiv preprint arXiv:2009.14146 (2020)
3. Nayak, R., Pati, U.C., Das, S.K.: A comprehensive review on deep learning-based methods for video anomaly detection. Image and Vision Computing 106, 104078 (2021)
4. Duong, H.T., Le, V.T., Hoang, V.T.: Deep learning-based anomaly detection in video surveillance: A survey. Sensors 23(11), 5024 (2023)
5. Wu, P., Zhou, X., Pang, G., Zhou, L., Yan, Q., Wang, P., Zhang, Y.: Vadclip:Adapting vision-language models for weakly supervised video anomaly detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp.6074–6082 (2024)
6. Chen, J., Li, L., Su, L., Zha, Z.J., Huang, Q.: Prompt-enhanced multiple instance learning for weakly supervised video anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18319–18329 (2024)
7. Li, C., Chen, M.: Dy-mil: dynamic multiple-instance learning framework for video anomaly detection. Multimedia Systems 30(1), 11 (2024)
8. Joo, H.K., Vo, K., Yamazaki, K., Le, N.: Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 3230–3234. IEEE (2023)
9. Chen, Y., Liu, Z., Zhang, B., Fok, W., Qi, X., Wu, Y.C.: Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detec-tion. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp.387–395 (2023)
10. Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J.W., Carneiro, G.: Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4975–4986 (2021)
11. Shi, H., Wang, L., Zhou, S., Hua, G., Tang, W.: Learning anomalies with normality prior for unsupervised video anomaly detection. In: European Conference on Computer Vision. pp. 163–180. Springer (2024)
12. Zhang, M., Wang, J., Qi, Q., Ren, P., Sun, H., Zhuang, Z., Zhang, L., Liao, J.: Safeguarding sustainable cities: unsupervised video anomaly detection through diffusion-based latent pattern learning. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. pp. 7572–7580 (2024)
13. Zanella, L., Menapace, W., Mancini, M., Wang, Y., Ricci, E.: Harnessing large language models for training-free video anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18527–18536 (2024)
14. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
15. Thakare, K.V., Dogra, D.P., Choi, H., Kim, H., Kim, I.J.: Rareanom: A benchmark video dataset for rare type anomalies. Pattern Recognition 140, 109567 (2023)

16. Thakare, K.V., Raghuwanshi, Y., Dogra, D.P., Choi, H., Kim, I.J.: Dyannet: A scene dynamicity guided self-trained video anomaly detection network. In: Proceedings of the IEEE/CVF Winter conference on applications of computer vision. pp. 5541–5550 (2023)
17. Zaheer, M.Z., Mahmood, A., Khan, M.H., Segu, M., Yu, F., Lee, S.I.: Generative cooperative learning for unsupervised video anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14744–14754 (2022)
18. Yan, C., Zhang, S., Liu, Y., Pang, G., Wang, W.: Feature prediction diffusion model for video anomaly detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5527–5537 (2023)
19. Hirschorn, O., Avidan, S.: Normalizing flows for human pose anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13545–13554 (2023)
20. Shi, C., Sun, C., Wu, Y., Jia, Y.: Video anomaly detection via sequentially learning multiple pretext tasks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10330–10340 (2023)
21. Wang, J., Cherian, A.: Gods: Generalized one-class discriminative subspaces for anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8201–8211 (2019)
22. Li, G., Cai, G., Zeng, X., Zhao, R.: Scale-aware spatio-temporal relation learning for video anomaly detection. In: European Conference on Computer Vision. pp. 333–350. Springer (2022)
23. Wang, G., Yuan, X., Zheng, A., Hsu, H.M., Hwang, J.N.: Anomaly candidate identification and starting time estimation of vehicles from traffic videos. In: CVPR workshops. pp. 382–390 (2019)
24. Wang, S., Wei, Z., Choi, Y., Ren, X.: Can llms reason with rules? logic scaffolding for stress-testing and improving llms. arXiv preprint arXiv:2402.11442 (2024)
25. Cohen, J., Rosenfeld, E., Kolter, Z.: Certified adversarial robustness via randomized smoothing. In: international conference on machine learning. pp. 1310–1320. PMLR (2019)
26. Liu, W., Luo, W., Lian, D., Gao, S.: Future frame prediction for anomaly detection a new baseline. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6536–6545 (2018)
27. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6479–6488 (2018)
28. Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., Yang, Z.: Not only look, but also listen: Learning multimodal violence detection under weak supervision. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16. pp. 322–339. Springer (2020)
29. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
30. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Open and efficient foundation language models. Preprint at arXiv. https://doi. org/10.48550/arXiv 2302(3) (2023)
31. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15180–15190 (2023)

32. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
33. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26296–26306 (2024)