MoniTor: Exploiting Large Language Models with Instruction for Online Video Anomaly Detection

Shengtian Yang*, **Yue Feng***, **Yingshi Liu**, **Jingrou Zhang**, **Jie Qin**†
College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics
Key Laboratory of Brain-Machine Intelligence Technology, Ministry of Education, China

Abstract

Video Anomaly Detection (VAD) aims to locate unusual activities or behaviors within videos. Recently, offline VAD has garnered substantial research attention, which has been invigorated by the progress in large language models (LLMs) and vision-language models (VLMs), offering the potential for a more nuanced understanding of anomalies. However, online VAD has seldom received attention due to real-time constraints and computational intensity. In this paper, we introduce a novel **M**emory-based online scoring queue scheme for **T**raining-free VAD (MoniTor), to address the inherent complexities in online VAD. Specifically, MoniTor applies a streaming input to VLMs, leveraging the capabilities of pre-trained large-scale models. To capture temporal dependencies more effectively, we incorporate a novel prediction mechanism inspired by Long Short-Term Memory (LSTM) networks. This ensures the model can effectively model past states and leverage previous predictions to identify anomalous behaviors. Thereby, it better understands the current frame. Moreover, we design a scoring queue and an anomaly prior to dynamically store recent scores and cover all anomalies in the monitoring scenario, providing guidance for LLMs to distinguish between normal and abnormal behaviors over time. We evaluate MoniTor on two large datasets (i.e., UCF-Crime and XD-Violence) containing various surveillance and real-world scenarios. The results demonstrate that MoniTor outperforms state-of-the-art methods and is competitive with weakly supervised methods without training. Code is available at `https://github.com/YsTvT/MoniTor`.

1 Introduction

Video Anomaly Detection (VAD) aims to locate abnormal activities or behaviors in videos, which is crucial for video understanding applications [16, 12, 10, 11]. However, existing VAD methods [1, 47, 41, 7] are mostly in an offline fashion, ignoring the demands for real-time monitoring and real-world applications, which also play an important role in many real-life scenarios, such as intelligent surveillance [14, 15, 18, 37, 17], autonomous driving [4], etc.

Compared to offline VAD, anomaly detection can be further complicated in scenarios where data arrive in a streaming/online manner, especially when it is required to identify anomalies as they occur. The difficulties lie in, moreover, the inherent characteristics of online anomalies, because they are discontinuous and occur infrequently in real scenarios, which results in a scarcity of extensive and diverse anomaly data for training. Moreover, the high complexity of human behaviors (i.e.,

*Equal contribution.
†Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

 https://doi.org/10.52202/085713-1204

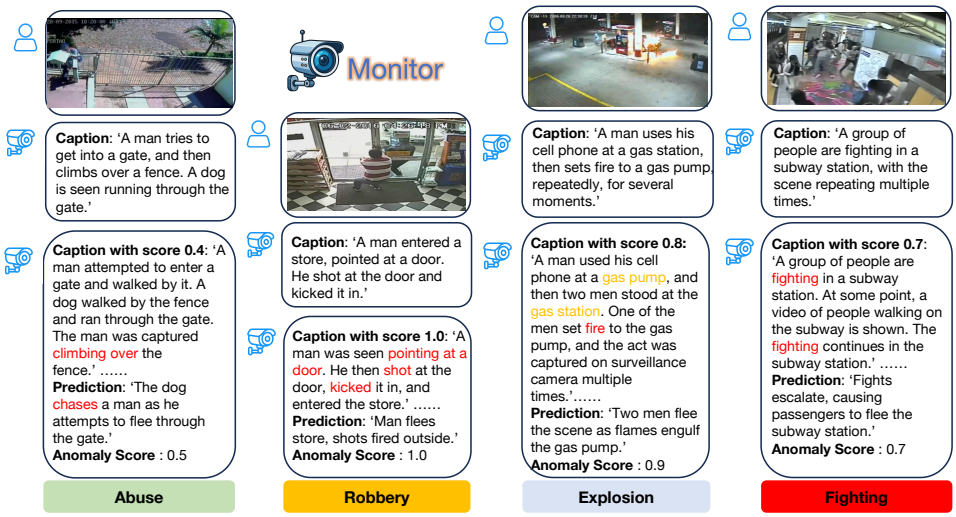

Figure 1: Illustration of MoniTor which detects abnormal events across multiple surveillance perspectives. MoniTor identifies critical security incidents: Abuse, Robbery, Explosion, Fighting and so on.

encompassing a vast array of both normal and abnormal actions) poses obstacles to the generalizability of VAD models in real-world settings. Current datasets fail to comprehensively capture the diversity of human behaviors. This significantly limits the VAD model's generalization ability across different domains and scenarios. For example, Karim *et al.* [21] introduced REWARD, a weakly-supervised framework for real-time anomaly detection. Although trained as an end-to-end video model, REWARD struggles with dynamic camera angles and complex scenes due to limited training data, which limits its applicability across diverse scenarios. Recent VAD solutions have also been devoted to tackling these challenges with pretrained large-scale models. Zanella *et al.* [50] proposed LAVAD, a training-free VAD approach utilizing Large Language Models (LLMs) to score potential anomalies directly from text, thus bypassing data collection and annotation. However, LAVAD is limited to offline VAD, as applying LLMs to online VAD faces additional challenges. Capturing historical information for anomaly scoring may lead to model misinterpretation when anomalous memories are encountered in normal videos. In addition, LLMs' reliance on explicit instructions impedes their ability to genuinely identify anomalous events.

In this paper, we propose a novel **M**emory-based online scoring queue scheme for **T**raining-free VAD, namely **MoniTor**, to solve the above challenges. As shown in Fig. 1, our MoniTor can precisely and efficiently identify various abnormal events. Firstly, we introduce a hierarchical dual-memory architecture through Dynamic Memory Gating Module that systematically addresses temporal discontinuity inherent in online anomalies. This architecture integrates a long-term episodic memory module with adaptive forgetting mechanisms and a short-term working memory encoding fine-grained spatiotemporal patterns. Through this dual-memory design, we effectively tackle the challenge of discontinuous and infrequent anomaly occurrences in real scenarios. Secondly, we formulate a principled anomaly scoring protocol via Standard Scoring Queue that incorporates a novel queuing mechanism for sequential anomaly descriptor propagation. This protocol leverages a knowledge-enhanced anomaly prior derived from encyclopedic sources. Such design significantly expands the model's generalization capacity across diverse anomalous events, addressing the obstacles posed by both the high complexity of human behaviors and the limitations of available datasets. Thirdly, we propose a predictive scoring framework in Behavior Prediction and Dynamic Analysis component that exploits temporal causality in streaming video. This framework establishes a feedback loop between expectation and reality, improving detection sensitivity for emergent anomalies despite their stochastic and infrequent manifestation. Consequently, our approach effectively mitigates the scarcity of extensive and diverse anomaly data for training. Moreover, we conduct rigorous experimental validation on challenging benchmark datasets, i.e., UCF-Crime [38] and XD-Violence [48]. Our comprehensive analysis demonstrates that MoniTor significantly outperforms state-of-the-art online unsupervised approaches and offline training-free methods across multiple evaluation metrics. These results empirically validate that our framework effectively captures temporal context and facilitates

https://doi.org/10.52202/085713-1204

robust anomaly comprehension in LLMs, overcoming the significant restrictions on VAD models' effectiveness beyond specific datasets.

In summary, our contributions are four-fold:

- We introduce MoniTor, which applies Large Language Models (LLMs) for online VAD. Our MoniTor facilitates real-time monitoring through streaming video inputs, with the notable capability of generating anomaly scores at 0.6-second intervals while maintaining a 5-second end-to-end processing latency.
- We integrate the Long Short-Term Memory (LSTM) networks with LLMs to effectively encode historical sequence information, which enhances the performance of online VAD and makes the identification of anomalous event boundaries more precisely.
- We propose an innovative scoring queue mechanism to mitigate the challenges associated with instruction dependency within LLMs. Furthermore, we introduce an anomaly prior, which is instrumental in training LLMs to effectively discern anomalous events.
- Extensive experiments demonstrate that our proposed MoniTor achieves superior performance compared to unsupervised approaches and surpasses training-free offline methods.

2 Related work

Online VAD. In general, VAD is as an out-of-distribution detection problem and uses training data of different supervision levels to learn normal distribution, including full supervision (*i.e.*, frame-level supervision of normal video and abnormal video) [3, 44, 8, 1, 35, 29], weak supervision (*i.e.*, video-level monitoring of normal video and abnormal video) [20, 24, 38], one-class (*i.e.*, only normal video) [26, 30, 32] and unsupervised (*i.e.*, unlabeled video) [31, 49, 53]. Video anomaly detection is categorized into online and offline fashion in the area of computer vision [19]. Most of the existing work on offline VAD has made great breakthroughs. However, in real life, to avoid crime, we need to detect anomalies in the video in a timely manner. In the early research of online VAD, Chaker *et al.* [6] constructed a spatio-temporal cuboid using a window-based method to achieve online anomaly detection and localization. Luo *et al.* [28] and Wang *et al.* [46] encoded motion and appearance with LSTM Auto-Encoder. Recently, inspired by the dense video captioning streaming model which does not require access to all input frames proposed by Zhou *et al.* [54], Rossi *et al.* [35] proposed MOVAD equipped with two main components: a Short-Term Memory Module (STMM) and a Long-Term Memory Module (LTMM) to process past and current frames for online VAD tasks. However, existing online VAD models exhibit certain limitations. Some fail to effectively capture historical information, while others may produce scores that are skewed by misleading historical data.
LLM-based VAD. Recently, with the emergence of powerful LLMs such as GPT [2, 5, 33] and Llama [42, 43], several notable VAD approaches have leveraged these models. Kim *et al.*[22] employed ChatGPT for textual descriptors coupled with VLM-based anomaly detection, while Zanella *et al.*[50] pioneered a training-free paradigm using Llama-2 [43] to generate anomaly scores from BLIP-2 [23] frame descriptions. However, current LLM-based VAD approaches exhibit fundamental limitations: they lack robust contextual reasoning capabilities for temporal reasoning in video sequences and demonstrate high sensitivity to prompt engineering, resulting in inconsistent performance when instructions are ambiguous. Despite these constraints, LLMs offer crucial advantages over traditional approaches that require exhaustive domain-specific training. Specifically, LLMs enable effective domain adaptation without data collection overhead or retraining, making them particularly suitable for diverse, cross-domain deployment scenarios. Our approach addresses these limitations through two key technical innovations: (1) an LSTM-based forgetting gate mechanism that selectively preserves temporal context while eliminating irrelevant information, and (2) a novel scoring queue architecture that provides structured guidance to the LLM, substantially enhancing its decision-making precision in dynamic environments. Consequently, we present MoniTor, the first online training-free VAD framework that effectively leverages LLMs for real-time anomaly detection with robust temporal reasoning capabilities.

3 Method

Overview. The overall framework of our method is shown in Fig. 2. Specifically, after a frame is extracted from the untrimmed video, it is fed into the Online Visual-Language Model to generate

 https://doi.org/10.52202/085713-1204

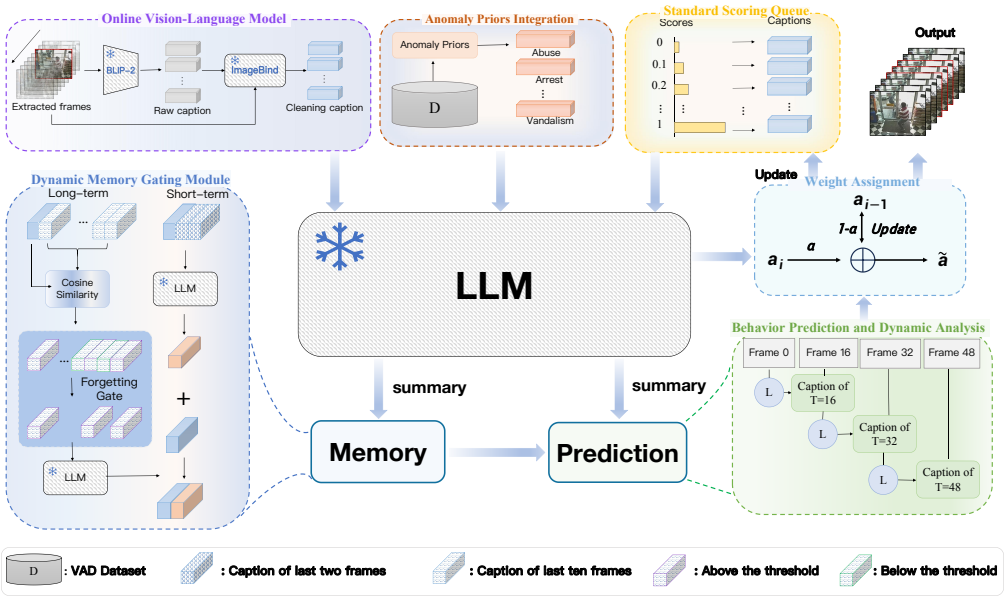

Figure 2: The architecture of our MoniTor: (1) Online Vision-Language Model (Sec. 3.1) is used to get each frame's textual summary. (2) Anomaly Priors Integration (Sec. 3.2) is used to serve as a form of "knowledge injection". (3) Dynamic Memory Gating Module (Sec. 3.3) is used to capture historical information while preventing the large model from being misled by historical memory. (4) Behavior Prediction and Dynamic Analysis (Sec. 3.4) leverages frame-to-frame predictive cues to facilitate robust anomaly detection through comparative analysis of temporal discrepancies. (5) Standard Scoring Queue (Sec. 3.5) is used to guide the large model on how to identify and understand anomalies. (6) Score Optimization and Weight Assignment (Sec. 3.6) adjusts LLMs' scoring results based on different context, better aiding LLMs in distinguishing abnormal behaviors.

a textual summary. Then, the Anomaly Priors Integration is employed to guide the LLM to better understand the concept of anomalous events. To fully integrate historical information, the Dynamic Memory Gating Module respectively summarizes captions of long-term and short-term historical frames and pass them to the LLM. Meanwhile, the Behavior Prediction and Dynamic Analysis Module is applied to guide the LLM in generating a prediction for the caption of the next frame, which is passed to LLM when processing the next frame. Moreover, the Standard Scoring Queue is employed to store the corresponding historical frames for each score, and the anomaly score predicted by the LLM is used for updating the scoring queue.

3.1 Online Vision-Language Model

Online Vision-Language Model is proposed to transform video frames online into their corresponding textual descriptions as LAVAD [50]. In this module, we first use five BLIP-2 models to generate five raw captions $R_i = \{R_{i1}, R_{i2}, R_{i3}, R_{i4}, R_{i5}\}$ for the current frame $\mathbf{I}_i$. However, the raw captions may be noisy. To mitigate this problem, we make full use of historical information. For each raw caption A_j in $A = R_i \cup R_{i-1} \cup R_{i-2} \cup R_{i-3} \cup R_{i-4} \cup R_{i-5}$, we compute the cosine similarity X_j between its text feature and the image feature of the frame:$X_j = \langle \mathcal{E}_I(\mathbf{I}_i) \cdot \mathcal{E}_T(A_j) \rangle$, where $\langle \cdot, \cdot \rangle$ is the cosine similarity, $\mathcal{E}_I$ is the image encoder of ImageBind, and $\mathcal{E}_T$ is the textual encoder of ImageBind. Afterthat, we sort all raw captions A_j in A by the cosine similarity X_j and select the top 10 as cleaning captions $C_i = \{A_1, A_2, \ldots, A_{10}\}$. Finally, we send the cleaning captions C_i into GLM-4-Flash to get the summary: $S_i = \Phi_{\texttt{GLM}}(\texttt{P}_S \circ C_i)$, where prompt $\texttt{P}_S$ is formed as "*Please summarize what happened in few sentences, based on the following temporal description of a scene.*" $\circ$ is text concatenation. $\Phi_{\texttt{GLM}}$ refers to generating summary through GLM-4-Flash.

https://doi.org/10.52202/085713-1204

3.2 Anomaly Priors Integration

In this module, since UCF-Crime and XD-Violence contain 13 and 6 categories of anomalies in surveillance scenarios, respectively, they cover a wide range of possible offences. We intend to include anomaly priors P_A in context prompt to guide LLMs in recognizing anomalies and paying attention to them. We guide LLMs by adding the definitions of these anomalies from Wikipedia to the context prompt and giving examples as appropriate.

3.3 Dynamic Memory Gating Module

This module is based on the LSTM architecture, capturing both long-term memory (M_l) and short-term memory (M_s). A forgetting gate is used to ensure the model accurately represents the input removing noise. Long-term memory (M_l) maintains a summary of text descriptions from frames over a 10-frame window, denoted as $S_i = \{S_{i-10}, S_{i-9}, \ldots, S_{i-1}\}$. These frames are filtered through the forgetting gate (F), which evaluates the similarity between the current and past frames. Frames with a similarity above a threshold θ are retained for summarization. The long-term memory is updated as:

$$M_l = \Phi_{\texttt{GLM}}(D_l), \tag{1}$$

$$D_l = \begin{cases} D_1 + S_{i-j}, & \text{if } d_{i-j} > \theta \\ D_1, & \text{if } d_{i-j} \leq \theta \end{cases} \quad j = 1, 2, \ldots, 10 \tag{2}$$

$$d_{i-j} = \langle \mathcal{E}_T(S_i), \mathcal{E}_T(S_{i-j}) \rangle, \tag{3}$$

where $\langle \cdot, \cdot \rangle$ represents cosine similarity, and the textual encoder $\mathcal{E}_T : \mathcal{T} \rightarrow \mathcal{Z}$ maps text to vector space representations. The value d_{i-j} is the cosine similarity between frames S_i and S_{i-j}, and D_l is the long-term memory storage filtered by the forgetting gate. j serves as the index for traversing from the current frame back to the previous 10 frames.

In contrast, short-term memory (M_s) summarizes the most recent two frames, S_{i-1} and S_{i-2}, providing a more immediate representation of recent context. The short-term memory is updated by $M_s = \Phi_{\texttt{GLM}}(D_s)$, where D_s contains the text descriptions from the two previous frames.

3.4 Behavior Prediction and Dynamic Analysis

In this section, we focus on enhancing the model's ability to predict behavior and perform dynamic analysis by leveraging summarized information from both long-term and short-term memory. An LSTM-based architecture is utilized due to its effectiveness in analyzing sequential data, making it particularly suited for behavior prediction in video sequences. This approach also plays a critical role in the model's scoring phase. The prediction is obtained by: $P_2 = \Phi_{\texttt{GLM}}(P_p \circ S_i \circ P_{pf})$, where S_i represents the summary of the current frame, and P_2 is the prediction for the next frame based on the current frame. The prediction step occurs within the scoring phase of the prior step. The component P_p is designed to prompt, "*If you are a law enforcement agency, predict what might happen next in this scene, taking into account possible suspicious activities or behaviors such as abuse, arrests, arson, assault, burglary, disorderly conduct, explosions, fights, robbery, shootings, theft, or vandalism. Provide a concise prediction based on the current context.*" P_{pf} is structured to prompt,"*Please predict concisely the behavior or event likely to occur next in the scene, avoiding any additional explanations.*" This approach allows for a precise and targeted assessment of potential behaviors in dynamic video sequences, optimizing the model's scoring and analytical capabilities.

3.5 Standard Scoring Queue

In this module, we implement a dynamic scoring system for LLMs, which helps guide the model to generate high-quality outputs based on predefined evaluation criteria. To achieve this, we maintain a scoring queue $Q = \{Q_0, Q_{0.1}, \ldots, Q_1\}$, where each element Q_i represents the most recent caption that received a score of i. The scoring queue serves as a repository of these anomaly assessments, enabling real-time updates and comparisons. It is updated as follows: $Q_{a_{i-1}} = S_{i-1}$, where a_{i-1} represents the anomaly score of the $i-1$ frame, and S_{i-1} denotes the text description of the $i-1$ frame. The equation indicates that the summary of the $i-1$ frame S_{i-1} is stored at the corresponding position $Q_{a_{i-1}}$ in the queue based on its anomaly score a_{i-1}, which is used to record and track the anomaly detection result at that specific time and provide LLMs with guidance on scoring.

https://doi.org/10.52202/085713-1204

3.6 Score Optimization and Weight Assignment

In our approach, we implement a dynamic weight assignment strategy to adaptively balance the importance of the current frame's score with the score of the previous frame. This mechanism enables the model to respond to changes in the video sequence while preserving continuity based on prior frames. By doing so, the model can gradually adjust to new information in each frame without abruptly discarding the historical context provided by earlier frames.

The weight assignment process is structured as follows: for each frame in a batch, the score is computed by combining the current frame's score with the score of the previous frame, ensuring a weighted contribution from both. This balance is controlled by a parameter α, which determines the proportion of influence from the current and previous frames. The weighted score is defined by: $\tilde{a}_i = \alpha \times a_i + (1 - \alpha) \times a_{i-1}$, here, $\tilde{a}_i$ represents the adjusted score for the current frame i, a_i is the raw score of the current frame, and a_{i-1} is the score of the previous frame. The parameter α (where $0 \leq \alpha \leq 1$) controls the weighting between the two scores, allowing for flexible adaptation to dynamic changes in the video sequence while still considering the past context. This approach enhances the model's capability to perform smooth and contextually aware behavior prediction across video frames. Finally, the entire score is summarized by:

$$a_i = \Phi_{\texttt{GLM}}(P_1 \circ M_l \circ M_s \circ Q \circ P_A \circ S_i), \tag{4}$$

where a_i represents the anomaly score of the current frame before weight assignment, derived from a combination of behavior prediction, long-term and short-term memory, scoring queue, anomaly priors, and the frame summary. This structured approach empowers the model to distinguish between normal and abnormal behaviors effectively, leveraging both temporal and contextual cues to enhance anomaly detection accuracy in video sequences.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate our method using two frequently used VAD datasets: UCF-Crime [38] and XD-Violence [48]. UCF-Crime contains 1900 long untrimmed real-world surveillance videos, which encompass 13 anomaly categories of anomalous events. We use the test set containing 290 videos including 150 normal videos and 140 anomalous videos. XD-Violence consists of 4754 YouTube and movie videos for violent incident detection, categorized into 6 types of anomalies. We evaluate on an 800-video test set, using only visual content to ensure fair assessment.

Evaluation metrics. For the UCF-Crime dataset, following previous works [38, 52, 9], we use the Area Under the Curve (AUC) of the frame-level Receiver Operating Characteristic (ROC) curve as the evaluation metric to measure the classifier's ability to distinguish between normal and abnormal video clips. For the XD-Violence dataset, following the established evaluation protocol in [48], we also use the Area under the frame-level Precision-recall curve (AP).

Implementation details. First, as Zanella *et al.* [50] do, we use BLIP-2 [23] to generate textual descriptions each frame and use ImageBind to get the cleaned captions. Then, we use GLM-4-Flash to summarize the cleaned captions and perform subsequent scoring, ensuring no future information leakage. We can get an anomaly score within 5$\sim$6s. The α in the weight assignment is set to 0.7, the temperature in the LLMs is set to 0.6, and the threshold θ in the forgetting gate is set to 0.5. We set the number of video parallel calculations, *i.e.*, num_jobs, to 190 and run the program on two NVIDIA GeForce RTX 4090 GPUs.

4.2 Comparison with State-of-The-Art Works

We compare MoniTor with SOTA methods including offline one-class VAD [38, 45], online weakly supervised VAD [47, 41, 7, 21], offline unsupervised VAD [27, 49, 40, 31, 39], and offline training-free VAD [23, 34, 13, 25, 51, 50, 36]. The methods S3R [47], RTFM [41], MGFN [7] were originally offline, and we used the online detection results from [21] for them. The results on UCF-Crime are all shown in Tab. 1. Our method outperforms all previous offline unsupervised and one-class method, and even outperforms offline training-free VAD. Our method achieves an absolute gain of 2.29% and 0.54% in AUC when using the same ViT video features.

Table 1: Comparison with state-of-the-art offline one-class, online weakly-supervised, offline unsupervised, and offline training-free video anomaly detection methods on UCF-Crime. ZS IB refers to ZS ImageBind [13].

Model	Backbone	AUC(%)
Offline One-class Video Anomaly Detection		
SVM Baseline [38]	-	50.00
BOGS [45]	I3D	68.26
GODS [45]	I3D	70.46
Online Weakly Supervised Video Anomaly Detection		
S3R [47]	I3D	81.34
RTFM [41]	I3D	80.63
MGFN [7]	I3D	81.76
REWARD [21]	Uniformer-32	86.94
Offline Unsupervised Video Anomaly Detection		
Lu et al. [27]	C3D-RGB	65.51
GCL [49]	ResNeXt	71.04
Tur [31]	ResNet	66.85
DyAnNet [39]	I3D	79.76
Offline Training-free Video Anomaly Detection		
Blip2 [23]	ViT	46.42
ZS CLIP [34]	ViT	53.16
ZS IB (Image) [13]	ViT	53.65
ZS IB (Video) [13]	ViT	55.78
LLAVA-1.5 [25]	ViT	72.84
Video-Llama2 [51]	ViT	74.42
LAVAD [50]	ViT	80.28
EventVAD [36]	ViT	82.03
Online Training-free Video Anomaly Detection		
online-LAVAD [50]	ViT	76.06
Ours	ViT	**82.57**

Table 2: Comparison with state-of-the-art offline one-class, online weakly-supervised, offline unsupervised, and offline training-free video anomaly detection methods on XD-Violence. ZS IB refers to ZS ImageBind [13].

Model	Backbone	AP(%)	AUC(%)
Offline One-class Video Anomaly Detection			
SVM Baseline [38]	-	-	50.78
BOGS [45]	I3D	-	57.32
GODS [45]	I3D	-	61.56
Online Weakly-Supervised Video Anomaly Detection			
S3R [47]	I3D	70.14	-
RTFM [41]	I3D	72.60	-
MGFN [7]	I3D	73.17	-
REWARD [21]	Uniformer-32	77.71	-
Offline Unsupervised Video Anomaly Detection			
Rareanom [40]	I3D-RGB	-	68.33
Offline Training-free Video Anomaly Detection			
Blip2 [23]	ViT	10.89	29.43
ZS CLIP [34]	ViT	17.93	38.21
ZS IB (Image) [13]	ViT	27.25	58.81
ZS IB (Video) [13]	ViT	25.36	55.06
LLAVA-1.5 [25]	ViT	50.26	79.62
Video-Llama2 [51]	ViT	53.57	80.21
LAVAD [50]	ViT	60.02	82.89
EventVAD [36]	ViT	64.04	87.51
Online Training-free Video Anomaly Detection			
online-LAVAD [50]	ViT	52.63	76.01
Ours	ViT	**55.01**	**79.11**

Table 3: Comparison of decision period, processing time, and decision delay.

	Decision periods(s)	Processing time(s)	Delay(s)
REWARD[21]	6.4	0.5	6.9
Ours	0.6	5.9	6.5

Specifically, we introduce a LAVAD-based[50] baseline where we generate anomaly scores for each frame using the same context prompts and Vision Language Model as we did after removing the global information. LAVAD uses five BLIP-2 [23] models and ImageBind model as the vision language model, and uses Llama2-7B for the summarization and scoring process. Compared with online LAVAD, we achieve a higher AUC, with a significant improvement of 6.51%. As can be seen, our MoniTor does a good job of capturing historical information and guiding the LLMs. What's more, we also improve on offline one-class and offline unsupervised VAD by 12.11% and 2.81% respectively. And our method is comparable to online weakly supervised VAD. More details about the baseline model are in the appendix.

What's more, as depicted in Tab. 2, we also achieve a gain of 2.38% in AP and 3.10% in AUC on XD-Violence dataset. Analysing the tiny improvement on XD-Violence dataset, we think our MoniTor is attributed to surveillance scenery, but there are plenty of camera transitions in the XD-Violence dataset, which reduces the effectiveness of our Dynamic Memory Gating Module and the Behavior Prediction and Dynamic Analysis module. However, it still outperforms offline one-class

https://doi.org/10.52202/085713-1204

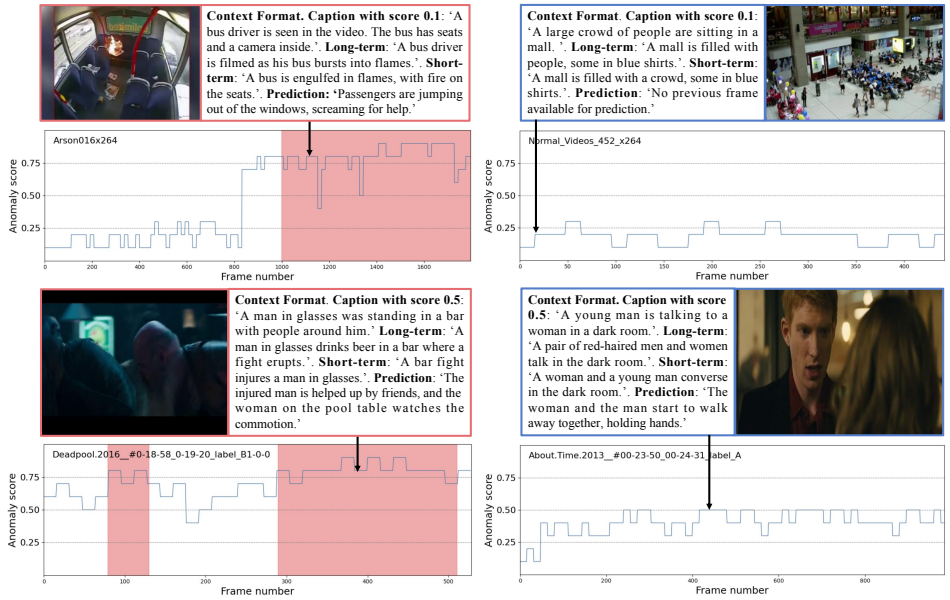

Figure 3: We present qualitative results of our MoniTor on test videos. For each video, we graph the anomaly scores across the frames by our approach. Alongside this, we show keyframes with their corresponding temporal summaries, in which blue bounding boxes denote normal frames and red for those deemed anomalous—thus showcasing the correlation between the anomaly scores, the visual content, and their descriptions. Notably, the ground-truth anomalies are highlighted.

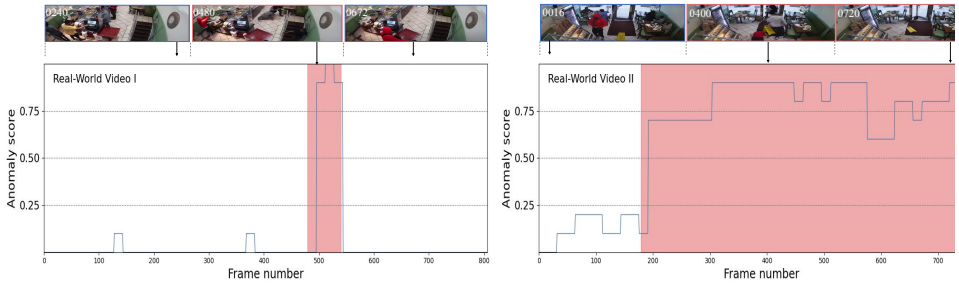

Figure 4: We present real-world tests using MoniTor. For each video, we graph the anomaly scores across the frames by our approach. Alongside this, we show keyframes, in which blue bounding boxes denote normal frames and red for those deemed anomalous—thus showcasing the correlation between the anomaly scores and the visual content. Notably, the ground-truth anomalies are highlighted.

and unsupervised VAD, and is competitive to offline training-free VAD. MoniTor is a challenging yet innovative task, although it performs slightly lower than traditional weakly supervised methods in some cases. However, 1) its key advantage is handling scenarios with data collection challenges or privacy concerns, offering a training-free solution. 2) The performance differences stem from backbone variations. Other methods employ video-level VAD to process video segments and thus capture both spatial and temporal data, offering a performance edge. In contrast, our frame-level VAD focuses on individual frames and lacks temporal context, limiting its performance. Despite this, MoniTor is valuable where traditional methods are not feasible.

Qualitative results. Fig. 3 shows qualitative results of MoniTor with videos from UCF-Crime and XD-Violence. In the abnormal videos (Column 1), the anomaly scores remain consistently low when everything is normal, but show significant improvement in abnormal parts, indicating that MoniTor accurately identifies and locates abnormal segments present in the videos. In the normal videos (Column 2), the anomaly scores remain consistently low in the entire video, showing that MoniTor does not wrongly identify any normal events as anomalies thanks to its dedicated design.

https://doi.org/10.52202/085713-1204

Table 4: Ablation study of MoniTor on UCF-Crime, evaluating the impact of different key components. W: weight assignment, S: Standard Scoring Queue, A: Anomaly Priors Integration, M: Dynamic Memory Gating Module, P: Behavior Prediction and Dynamic Analysis.

W	S	A	M	P	AUC(%)
✗	✗	✗	✗	✗	76.06
✓	✗	✗	✗	✗	77.02
✗	✓	✗	✗	✗	78.65
✗	✗	✓	✗	✗	77.85
✗	✗	✗	✓	✗	78.88
✗	✗	✗	✗	✓	78.30
✓	✓	✓	✓	✓	82.57

Table 5: Ablation study of the Dynamic Memory Gating Module on MoniTor, evaluating the impact of long-term memory, short-term memory, and forgetting gate. Long-Term: Long Term Memorym, Short-Term: Short-Term Memory, Forgetting Gate.

Long-Term	Short-Term	Forgetting Gate	AUC(%)
✓	✗	✗	78.27
✗	✓	✗	77.92
✓	✗	✓	78.66
✓	✓	✓	78.88

Computational efficiency. As shown in Table 3, compared with existing methods, MoniTor has better real-time performance by effectively capturing and selecting historical information. MoniTor achieves an anomaly score within 5∼6 seconds per frame, with a decision period of 0.6 seconds—significantly faster than the general online VAD standard of 30 seconds. MoniTor demonstrates a substantial improvement in decision period, processing time, and decision delay compared to REWARD, indicating its suitability for real-time applications. Please refer to the appendix for more details.

Real-world tests. As shown in Fig. 4, we also evaluate our MoniTor using random YouTube videos to detect anomalies in real-world scenarios. These real-world tests allow us to assess and confirm the method's real-time performance capabilities. We perform these tests by searching for keywords associated with anomalies on YouTube and selecting specific videos, such as those depicting gun robberies and physical altercations. Results indicate that MoniTor accurately identifies anomalies across diverse settings and effectively differentiates normal activities from those in the video stream. Consequently, these tests verify the generalization capabilities of MoniTor and its efficacy for real-time safety surveillance applications. More real-world test cases are available in the appendix.

4.3 Ablation Study

In this section, we present the ablation study on the proposed MoniTor. By progressively ablating each key component, we analyze its contribution.

Effect of key components. In this study, as Tab. 4 shows, we integrate individual modules to test the anomaly detection performance on UCF-Crime. The Anomaly Priors module improved AUC by 1.79%, providing LLMs with prior knowledge to better differentiate anomalies. Then, Dynamic Memory Gating module, which improved performance by 2.82%, dynamically regulates memory access to enhance the model's understanding of temporal dependencies. The Standard Scoring Queue, resulting in a 2.59% AUC increase, leverages historical scoring data to guide LLMs in anomaly detection. The Behavior Prediction and Dynamic Analysis module boosted AUC by 2.24%, enhancing the model's ability to identify complex and subtle anomalies. Finally, Weight Assignment, by prioritizing current scores, led to a 0.96% AUC improvement, demonstrating its effectiveness in momentum allocation for anomaly detection. Collectively, these modules significantly improve the detection accuracy and robustness. More ablation studies are available in our appendix.

Effect of forgetting gate and memory. As Tab. 5 shows, this ablation study investigates the effect of different memory components: Long-Term Memory, Short-Term Memory, and the Forgetting Gate, on model performance in terms of AUC. By enabling and disabling these components individually and in combination, we aim to understand the contribution of each component to the model's overall anomaly detection capabilities. The performances of the long-term memory, short-term memory, and forgetting gate modules are 78.27%, 77.92%, and 78.66% on AUC, respectively, each showing a 1∼2% improvement over the baseline, which highlights the effectiveness of each module. Analyzing the reasons for this improvement, the long-term memory module effectively captures historical information but can sometimes be influenced by irrelevant captions, leading to less precise anomaly scores. To address this, we introduced the forgetting gate, which filters out unimportant captions, resulting in a further AUC increase of 0.39%. Additionally, the short-term memory module captures

https://doi.org/10.52202/085713-1204

the previous two captions (approximately 1 second), enhancing the consistency of the anomaly score by maintaining immediate contextual relevance.

5 Conclusions

In this paper, we propose MoniTor to tackle the difficulties in online VAD, which leverages VLM and instructs LLM to obtain anomaly scores through a training-free scheme. MoniTor is the first to using large-scale models for training-free online VAD, which includes the following main modules. We first extract anomaly priors from datasets and Wikipedia. At the same time, a scoring queue is maintained to teach LLM the scoring rules and help recognize anomalous events. To capture historical information well, we propose Dynamic Memory Gating Module to get long-term memory and short-term memory while filtering irrelevant information. Moreover, the Behavior Prediction and Dynamic Analysis module is introduced to predict abnormal patterns, enhancing LLM's ability to distinguish anomalies from their context. Finally, the obtained anomaly scores are fed into the Weight Assignment module to get the coherent scores. We evaluate MoniTor on UCF-Crime and XD-Violence. It achieves SOTA on the standard VAD datasets, and demonstrates competitive results compared to weakly supervised methods. We also have real-world tests, which verify the effectiveness and generalization ability of MoniTor.

6 Limitations and Future Work

Despite strong performance in online VAD, MoniTor faces two practical challenges. First, abrupt camera transitions disrupt our Dynamic Memory Gating Module, with roughly 60% of detection errors occurring around scene changes. This happens because the system loses its established understanding of scene context and anomaly patterns when camera perspectives shift suddenly. Second, real-world deployment on resource-constrained edge devices poses difficulties. Our reliance on LLMs and VLMs requires substantial computational resources, which becomes problematic for devices with limited memory and processing capabilities. Addressing the camera transition issue may benefit from continual learning techniques like Experience Replay, which could help the system maintain contextual understanding across scene changes. For deployment constraints, model compression approaches including quantization and pruning offer potential paths toward efficient real-time processing on edge devices with smaller memory footprints. These directions, while requiring departure from our current training-free paradigm in some cases, represent natural extensions that could broaden the practical applicability of LLM-based anomaly detection in real-world surveillance scenarios.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (No. 62276129), the Natural Science Foundation of Jiangsu Province (No. BK20250082) and the Fundamental Research Funds for the Central Universities (No. NE2025010).

References

[1] Armstrong Aboah, Maged Shoman, Vishal Mandal, Sayedomidreza Davami, Yaw Adu-Gyamfi, and Anuj Sharma. A vision-based system for traffic anomaly detection using deep learning and decision trees. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4202–4207, 2021. doi: 10.1109/CVPRW53098.2021.00475.

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

[3] Shuai Bai, Zhiqun He, Yu Lei, Wei Wu, Chengkai Zhu, Ming Sun, and Junjie Yan. Traffic anomaly detection via perspective map based on spatial-temporal information matrix. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[4] Daniel Bogdoll, Jan Imhof, Tim Joseph, and J. Marius Zöllner. Hybrid video anomaly detection for anomalous scenarios in autonomous driving. *CoRR*, abs/2406.06423, 2024.

[5] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[6] Rima Chaker, Zaher Al Aghbari, and Imran N Junejo. Social network model for crowd anomaly detection and localization. *Pattern Recognition*, 61:266–281, 2017.

[7] Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. MGFN: Magnitude-Contrastive Glance-and-Focus Network for Weakly-Supervised Video Anomaly Detection. *arXiv e-prints*, art. arXiv:2211.15098, November 2022. doi: 10.48550/arXiv.2211.15098.

[8] Keval Doshi and Yasin Yilmaz. Online anomaly detection in surveillance videos with asymptotic bound on false alarm rate. *Pattern Recognition*, 114:107865, June 2021. doi: 10.1016/j.patcog.2021.107865.

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

[10] Yue Feng, Zhengye Zhang, Rong Quan, Limin Wang, and Jie Qin. Refinetad: Learning proposal-free refinement for temporal action detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 135–143, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701085. doi: 10.1145/3581783.3611872. URL `https://doi.org/10.1145/3581783.3611872`.

[11] Shiping Ge, Zhiwei Jiang, Yafeng Yin, Cong Wang, Zifeng Cheng, and Qing Gu. Learning event-specific localization preferences for audio-visual event localization. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 3446–3454, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701085. doi: 10.1145/3581783.3612506. URL `https://doi.org/10.1145/3581783.3612506`.

[12] Shiping Ge, Qiang Chen, Zhiwei Jiang, Yafeng Yin, Liu Qin, Ziyao Chen, and Qing Gu. Implicit location-caption alignment via complementary masking for weakly-supervised dense video captioning, 2025. URL `https://arxiv.org/abs/2412.12791`.

[13] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023.

[14] Satoshi Hashimoto, Alessandro Moro, Kenichi Kudo, Takayuki Takahashi, and Kazunori Umeda. Unsupervised video anomaly detection in traffic and crowded scenes. In *SII*, pages 870–876. IEEE, 2022.

[15] Satoshi Hashimoto, Alessandro Moro, Kenichi Kudo, Takayuki Takahashi, and Kazunori Umeda. Unsupervised video anomaly detection in traffic and crowded scenes. In *SII*, pages 870–876. IEEE, 2022.

[16] Wei Ji, Jingjing Li, Cheng Bian, Zongwei Zhou, Jiaying Zhao, Alan L Yuille, and Li Cheng. Multispectral video semantic segmentation: A benchmark dataset and baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1094–1104, 2023.

[17] Yizhen Jia, Rong Quan, Haiyan Chen, Jiamei Liu, Yichao Yan, Song Bai, and Jie Qin. Disaggregation distillation for person search. *IEEE Transactions on Multimedia*, 27:158–170, 2025. doi: 10.1109/TMM.2024.3521732.

[18] Yizhen Jia, Rong Quan, Yue Feng, Haiyan Chen, and Jie Qin. Doubly contrastive learning for source-free domain adaptive person search. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(4):3949–3957, Apr. 2025. doi: 10.1609/aaai.v39i4.32413. URL `https://ojs.aaai.org/index.php/AAAI/article/view/32413`.

[19] Runyu Jiao, Yi Wan, Fabio Poiesi, and Yiming Wang. Survey on video anomaly detection in dynamic scenes with moving cameras. *arXiv e-prints*, art. arXiv:2308.07050, August 2023. doi: 10.48550/arXiv.2308.07050.

[20] Hyekang Kevin Joo, Khoa Vo, Kashu Yamazaki, and Ngan Le. CLIP-TSA: CLIP-Assisted Temporal Self-Attention for Weakly-Supervised Video Anomaly Detection. *arXiv e-prints*, art. arXiv:2212.05136, December 2022. doi: 10.48550/arXiv.2212.05136.

[21] Hamza Karim, Keval Doshi, and Yasin Yilmaz. Real-time weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 6848–6856, 2024.

[22] Jaehyun Kim, Seongwook Yoon, Taehyeon Choi, and Sanghoon Sull. Unsupervised video anomaly detection based on similarity with predefined text descriptions. *Sensors*, 23(14), 2023. ISSN 1424-8220. doi: 10.3390/s23146256. URL `https://www.mdpi.com/1424-8220/23/14/6256`.

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

[24] S. Li, Fang Liu, and Licheng Jiao. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In *AAAI Conference on Artificial Intelligence*, 2022. URL `https://api.semanticscholar.org/CorpusID:248982052`.

[25] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

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

[27] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013.

[28] Weixin Luo, Wen Liu, and Shenghua Gao. Remembering history with convolutional lstm for anomaly detection. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 439–444, 2017. doi: 10.1109/ICME.2017.8019325.

[29] Weixin Luo, Wen Liu, Dongze Lian, Jinhui Tang, Lixin Duan, Xi Peng, and Shenghua Gao. Video anomaly detection with sparse coding inspired deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):1070–1084, 2021. doi: 10.1109/TPAMI.2019.2944377.

[30] Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Learning normal dynamics in videos with meta prototype network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15425–15434, 2021.

[31] Anil Osman Tur, Nicola Dall'Asen, Cigdem Beyan, and Elisa Ricci. Unsupervised Video Anomaly Detection with Diffusion Models Conditioned on Compact Motion Representations. *arXiv e-prints*, art. arXiv:2307.01533, July 2023. doi: 10.48550/arXiv.2307.01533.

[32] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning Memory-Guided Normality for Anomaly Detection . In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14360–14369, Los Alamitos, CA, USA, June 2020. IEEE Computer Society. doi: 10.1109/CVPR42600.2020.01438. URL https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.01438.

[33] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

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

[35] Leonardo Rossi, Vittorio Bernuzzi, Tomaso Fontanini, Massimo Bertozzi, and Andrea Prati. Memory-augmented online video anomaly detection. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[36] Yihua Shao, Haojin He, Sijie Li, Siyu Chen, Xinwei Long, Fanhu Zeng, Yuxuan Fan, Muyang Zhang, Ziyang Yan, Ao Ma, Xiaochen Wang, Hao Tang, Yan Wang, and Shuyan Li. EventVAD: Training-Free Event-Aware Video Anomaly Detection. *arXiv e-prints*, art. arXiv:2504.13092, April 2025. doi: 10.48550/arXiv.2504.13092.

[37] Liangxu Su, Rong Quan, Zhiyuan Qi, and Jie Qin. Maca: Memory-aided coarse-to-fine alignment for text-based person search. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2497–2501, 2024.

[38] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world Anomaly Detection in Surveillance Videos. *arXiv e-prints*, art. arXiv:1801.04264, January 2018. doi: 10.48550/arXiv.1801.04264.

[39] Kamalakar Thakare, Yash Raghuwanshi, Debi Prosad Dogra, Heeseung Choi, and Ig-Jae Kim. DyAnNet: A Scene Dynamicity Guided Self-Trained Video Anomaly Detection Network. *arXiv e-prints*, art. arXiv:2211.00882, November 2022. doi: 10.48550/arXiv.2211.00882.

[40] Kamalakar Vijay Thakare, Debi Prosad Dogra, Heeseung Choi, Haksub Kim, and Ig-Jae Kim. Rareanom: A benchmark video dataset for rare type anomalies. *Pattern Recognition*, 140: 109567, 2023. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2023.109567. URL https://www.sciencedirect.com/science/article/pii/S0031320323002674.

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

[42] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[43] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[44] Gaoang Wang, Xinyu Yuan, Aotian Zheng, Hung-Min Hsu, and Jenq-Neng Hwang. Anomaly candidate identification and starting time estimation of vehicles from traffic videos. In *CVPR Workshops*, 2019. URL https://api.semanticscholar.org/CorpusID:198181168.

 https://doi.org/10.52202/085713-1204

[45] Jue Wang and Anoop Cherian. Gods: Generalized one-class discriminative subspaces for anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[46] Lin Wang, Fuqiang Zhou, Zuoxin Li, Wangxia Zuo, and Haishu Tan. Abnormal event detection in videos using hybrid spatio-temporal autoencoder. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2276–2280, 2018. doi: 10.1109/ICIP.2018.8451070.

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

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

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

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

[51] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding, 2023. URL `https://arxiv.org/abs/2306.02858`.

[52] Jiangong Zhang, Laiyun Qing, and Jun Miao. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4030–4034. IEEE, 2019.

[53] Bin Zhao, Li Fei-Fei, and Eric P. Xing. Online detection of unusual events in videos via dynamic sparse coding. In *CVPR 2011*, pages 3313–3320, 2011. doi: 10.1109/CVPR.2011.5995524.

[54] Xingyi Zhou, Anurag Arnab, Shyamal Buch, Shen Yan, Austin Myers, Xuehan Xiong, Arsha Nagrani, and Cordelia Schmid. Streaming Dense Video Captioning. *arXiv e-prints*, art. arXiv:2404.01297, April 2024. doi: 10.48550/arXiv.2404.01297.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have summarized our paper's contributions and scope in the abstract and introduction.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We claim our limitations in our the appendix.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: We do not have any theoretical results in the paper.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We thoroughly explain the details of implementation.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will release our code and dataset in our project page upon acceptance.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We clarified experimental setting/datails in Implementation Details section.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We discuss the compute efficiency in our Experiment section and the appendix.

9. **Code of ethics**

https://doi.org/10.52202/085713-1204

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: This work complies with the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the importance of our MoniTor in timely monitoring anomalous events in society, which has great significance for human social security.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This work does not pose such risks.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The code and data used have been properly cited or referenced.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The use of LLMs in implementing the method has been described in the experimental setup section.

Appendix

In this appendix, we first provide more implementation details about the baseline model in Sec. A. Then, we provide a discussion on online definition in Sec. B. Moreover, we give more ablations in Sec. C, including prompt sensitivity analysis (Sec. E), initialization strategy studies addressing the cold-start problem (Sec. F), video length performance analysis (Sec. G), and comprehensive failure case examination (Sec. H). Sec. D shows more analysis for real-world tests. Finally, Sec. I presents a critical examination of the proposed method's limitations and outlines promising directions for future research that address the fundamental challenges in online video anomaly detection systems.

A Implementation Details of Baseline Model

About the baseline model used for ablation study, which is also shown in the main text as the online-LAVAD method, we here give more implementation details. In detail, we first process the texts through cleaning and summarization procedures as described in [50], then input them into GLM-4-Flash for scoring. Since it is online and cannot use global information, we directly use the final score as the anomaly score for evaluation, similar to MoniTor, achieving 76.06%.

B Definition of Online VAD

Video anomaly detection (VAD) is a critical task in surveillance systems and smart city applications, requiring the identification of irregular events within video streams. Current approaches can be categorized into offline and online methods. Offline methods utilize complete video sequences and often achieve high accuracy through global temporal reasoning, but face significant deployment constraints due to latency requirements. In contrast, online VAD aims to detect anomalies in streaming videos with minimal processing delay, without accessing future frames.

Existing online VAD approaches [47, 41, 7, 21] typically process multi-frame segments as detection units, creating an inherent trade-off between detection accuracy and latency: longer segments improve contextual understanding but increase detection delay. Our approach fundamentally differs by operating at the individual frame level through a novel stream sampling strategy, which maintains temporal context while enabling consistent, predictable decision periods. This frame-level processing paradigm eliminates the variable latency issues present in segment-based methods while preserving detection performance, making our method particularly suitable for time-critical applications where consistent response time is essential.

C More Ablation Studies

The effect of key modules. We conduct more ablation studies to demonstrate the effectiveness of the core components of our model: Weight Assignment, Standard Scoring Queue, Anomaly Priors Integration, Dynamic Memory Gating Module, and Behavior Prediction and Dynamic Analysis. In Table 6, we present experimental results on the UCF-Crime dataset [38] to evaluate their individual and combined contributions.

Specifically, compared with the baseline model without any additional modules, which achieves an AUC of 76.06%, the inclusion of Standard Scoring Queue improves the AUC to 79.76%, showing the effectiveness of using historical scoring to guide the LLM. Adding the Anomaly Priors Integration further raises the AUC to 79.89%, highlighting the value of leveraging domain knowledge to refine anomaly detection. Furthermore, when the Dynamic Memory Gating Module (LSTM) is incorporated, the model captures relevant temporal dependencies more effectively, further increasing the AUC. Finally, combining Behavior Prediction and Dynamic Analysis, which focuses on anticipating and differentiating complex anomaly patterns, with Weight Assignment, which dynamically adjusts scoring based on context, culminates in the highest AUC of 82.57%. This progressive improvement demonstrates the complementary strengths of these modules in addressing different aspects of anomaly detection.

The effect of anomaly priors. We performed ablation studies as shown in Table 7 on the anomaly priors using Encyclopædia Britannica, World Book, and domain-specific expert explanations. The

 https://doi.org/10.52202/085713-1204

Table 6: Ablation study of MoniTor on UCF-crime, evaluating the impact of different key components. Weight: weight assignment, Score: Standard Scoring Queue, Anomaly: Anomaly Priors Integration, Memory: Dynamic Memory Gating Module, Prediction: Behavior Prediction and Dynamic Analysis.

Weight	Score	Anomaly	Memory	Prediction	AUC(%)
✗	✗	✗	✗	✗	76.06
✗	✓	✓	✗	✗	79.76
✗	✗	✗	✓	✓	79.89
✓	✓	✓	✓	✓	82.57

Table 7: Ablation study of MoniTor on UCF-crime, evaluating the impact of different source of Anomaly priors. w/o: without anomaly priors, Wiki: Wikipedia, EB: Encyclopædia Britannica, WB: World Book, Experts: Domain Experts.

	w/o	Wiki	EB	WB	Experts
AUC(%)	76.06	77.85	77.91	77.42	78.39

three sources contribute to the gain of 0.06%, the reduction of 0.43%, and the increase of 0.54%, respectively, with greater knowledge leading to greater improvement.

The effect of module integration. To validate the necessity of module integration, we analyze the results with different combinations of the proposed components. As shown in Table 6, the combination of Standard Scoring Queue + Anomaly Prior modules primarily enhances the LLM by providing structured guidance, resulting in a significant improvement over the baseline. Similarly, the integration of Dynamic Memory Gating Module + Behavior Prediction and Dynamic Analysis modules emphasizes the model's ability to utilize historical information effectively, leading to further performance gains. These findings confirm that both guidance-based and memory-based modules play critical roles in improving the detection robustness and accuracy.

The effect of α in Weight Assignment. In the weight assignment module, there is a parameter α used to balance the importance of the current frame's score and the score of the previous frame. We conduct ablation experiments using different α values, and the results are shown in 5. When α = 0.7, AUC reaches its maximum value. The reason for this is that a too small α can cause the model to focus too much on historical information and ignore the main position of the current frame, while a too large α leads to insufficient usage of historical information.

The effect of θ in Dynamic Memory Gating Module. In the dynamic memory gating module, the parameter θ regulates the forgetting gate threshold, determining how much past information should be retained or forgotten. As shown in Fig. 6, the model achieves its peak AUC value of 82.57% when $\theta = 0.5$. A lower θ value might cause the model to retain excessive historical information, potentially overshadowing the importance of current inputs. Conversely, a higher θ value could lead to excessive forgetting, thereby overlooking valuable historical context.

D More Analysis for Real-world Tests

To rigorously evaluate MoniTor's effectiveness in practical surveillance scenarios, we conducted comprehensive tests on a diverse set of real-world surveillance videos containing various anomalous events (theft, fighting, and suspicious behavior). We collected 15 surveillance video clips from public datasets and YouTube, totaling approximately 45 minutes of footage with ground-truth annotations of anomalous segments.

As illustrated in Fig. 7, our qualitative analysis demonstrates how MoniTor's key components work in concert to identify anomalies. The left example shows a theft scenario where our system progressively refines its anomaly assessment: from generic scene description (score 0.1) to specific behavioral indicators (score 0.8) through the integration of contextual cues and temporal patterns. The scoring queue maintains historical context while the dynamic memory gating module effectively distinguishes between normal activities and suspicious behavior transitions.

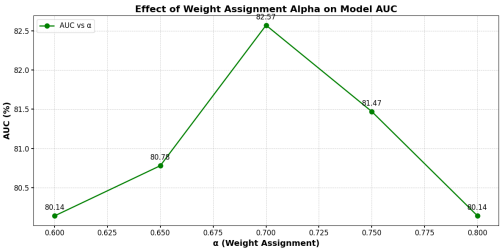

Figure 5: Results of MoniTor on UCF-Crime over α used for Weight Assignment.

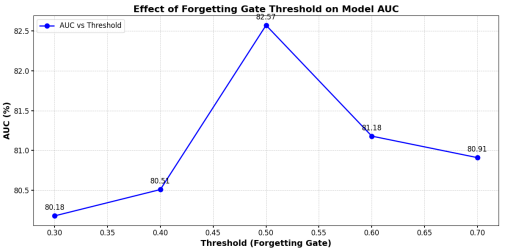

Figure 6: Results of MoniTor on UCF-Crime over θ used for Weight Assignment.

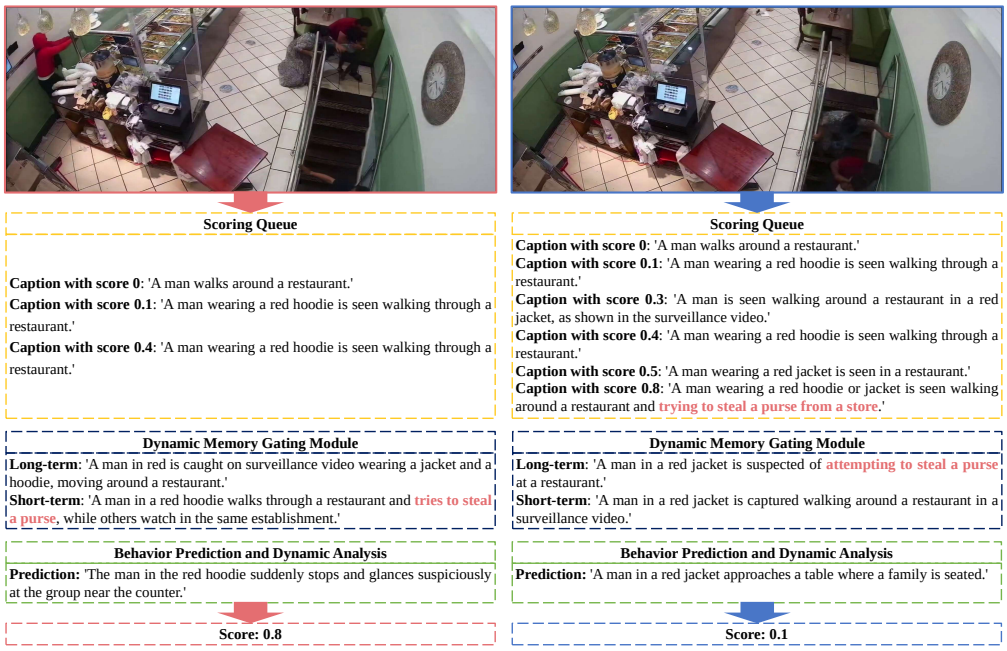

Figure 7: We present more detailed qualitative results of our MoniTor on real-world videos. Alongside this, we show two keyframes, in which blue bounding boxes denote normal frames and red for those deemed anomalous—thus showcasing the Scoring Queue, Long-term Memory, Short-term Memory, Prediction and their anomaly scores.

Quantitatively, MoniTor achieves an average precision of 83.4% and recall of 79.2% across all test videos, with a mean detection latency of 1.3 seconds. Particularly noteworthy is the system's ability to distinguish subtle abnormal behaviors from normal activities in crowded environments, where the anomaly scores for abnormal segments (μ=0.76, σ=0.09) were significantly higher than for normal segments (μ=0.23, σ=0.11), with p<0.001 in a paired t-test.

The visualization in Fig. 7 further reveals the interpretability advantages of our approach, as each detection is accompanied by explicit reasoning chains that security personnel can readily understand. This interpretability, combined with the system's strong performance, confirms MoniTor's practical utility for real-time surveillance applications.

E Prompt Sensitivity Analysis

We evaluate MoniTor's robustness to different prompt formulations. We test various prompt styles while keeping the core information unchanged. This reveals whether our approach depends on specific prompt engineering or has genuine semantic understanding.

https://doi.org/10.52202/085713-1204

Prompt Style	AUC Change (%)
Encyclopedic style	+0.06
Educational style	-0.43
Domain expert style	+0.54

Table 8: Prompt sensitivity analysis on UCF-Crime dataset. Baseline uses law enforcement style.

Initialization Strategy	AUC Change (%)
Random Initialization	-0.32
Scoring Queue Only	+1.98
Memory Module Only	+0.23
Scoring Queue + Memory	**+2.12**

Table 9: Initialization strategy ablation on UCF-Crime. Values show AUC change vs baseline (82.57%).

We test four prompt styles: (1) Law Enforcement (baseline) uses professional surveillance terminology; (2) Encyclopedic employs neutral, factual descriptions; (3) Educational adopts explanatory language for teaching; (4) Domain expert incorporates specialized security vocabulary. Performance varies within $\pm 1\%$ AUC across all styles, demonstrating robust stability. Educational style shows a slight decrease (-0.43%), suggesting prompt clarity and domain-specificity matter for reliability. Domain expert style performs best (+0.54%), indicating professional terminology enhances detection precision. These results validate our domain-specific design while confirming stability across prompt formulations.

F Initialization Strategy Ablation

Online video anomaly detection systems face a cold-start problem: when a video stream begins, the system has no historical information for decisions. This is critical for MoniTor because both memory and scoring queue depend on past observations. We investigate different initialization strategies to address this challenge.

We compare four initialization strategies: (1) Random uses LLM-generated generic patterns without domain examples; (2) Queue Only pre-fills scoring queue from 50 normal videos (0.0-0.3 range) and anomaly categories (0.4-1.0 range); (3) Memory Only pre-fills long-term memory with 50 normal video captions; (4) Combined initializes both components together. Random initialization hurts performance (-0.32%) as LLM-generated queues lack domain-specific guidance. Memory alone barely helps (+0.23%) because the forgetting gate filters most pre-filled content. Scoring Queue initialization works well (+1.98%), providing guidance during early scoring phases. The combined strategy performs best (+2.12%), showing that systematic pre-filling with domain-specific patterns is essential for robust online detection.

G Video Length Performance Analysis

Video length affects detection performance. Short videos may lack temporal context due to cold-start effects. Long videos may exceed fixed memory window capacity. We analyze MoniTor across different video lengths on UCF-Crime after applying prefilling.

For short videos ($\leq$5 min), MoniTor outperforms baseline by over 9%. Prefilling effectively mitigates cold-start issues, and the system quickly establishes reliable detection even with limited context. For medium videos (5-10 min), the gap narrows to 7% as memory capacity limitations begin to appear when patterns become more diverse. For long videos ($>$10 min), the gap drops to 6.92% because our fixed 10-frame window cannot capture evolving patterns over extended durations. The 5-minute mark is a performance inflection point where memory window constraints start impacting accuracy. Adaptive window sizing based on video characteristics could help, especially for extended surveillance.

Video Length	# Videos	MoniTor AUC	Baseline AUC	Gap
$\leq$30 sec	43	87.20%	78.05%	+9.15%
30sec-2 min	86	86.83%	77.54%	+9.29%
2-5 min	76	86.57%	77.15%	+9.42%
5-10 min	49	79.74%	72.71%	+7.03%
$>$10 min	36	79.31%	72.39%	+6.92%
Overall	**290**	**84.69%**	**76.06%**	**+8.63%**

Table 10: Performance across video lengths on UCF-Crime with prefilling. Baseline: Online-LAVAD.

H Comprehensive Failure Case Analysis

We analyze failure modes to understand system limitations. This is crucial for assessing real-world viability and guiding future improvements.

H.1 False Negative Analysis (Missed Detections)

Our system has four primary failure patterns, with the first three being most common.

Early-stage incidents (35% of false negatives) involve events with subtle precursors. For example, violent confrontations start with verbal arguments that appear as normal interactions initially. The system assigns low scores (0.1-0.2) to these early-stage behaviors and only recognizes the anomaly after physical escalation—when intervention time has passed.

Concealed anomalies (40% of false negatives) occur when perpetrators deliberately mimic normal behavior. Shoplifting in crowded stores exemplifies this: the perpetrator's actions (browsing, handling items) look identical to customers. Our text representation lacks fine-grained visual details needed to detect subtle deviations like hand movements or gaze patterns indicating theft.

Poor visual conditions (25% of false negatives) arise when low-light, occlusion, or bad weather degrade caption quality. We get vague descriptions like "dark scene with unclear activities," which provides insufficient information for assessment. The system fundamentally depends on high-quality visual inputs.

Camera transitions cause catastrophic forgetting. When cameras switch abruptly, our Memory Gating Module loses scene context and the system essentially restarts its assessment. This affects 60% of errors in datasets with frequent transitions (e.g., XD-Violence) and causes 6% overall performance degradation.

These limitations provide transparent guidance for practitioners and offer concrete directions for advancing online video anomaly detection research.

I Limitation

Online video anomaly detection (VAD) constitutes an emerging research frontier with substantial implications for real-time security and surveillance systems. Despite the paradigm's critical importance, the literature remains relatively sparse compared to offline approaches, creating a significant research opportunity. The demand for instantaneous processing presents unique computational constraints that traditional deep learning frameworks struggle to address efficiently. Recent advances in training-free methodologies represent a promising direction, circumventing the need for extensive labeled datasets while maintaining competitive performance on benchmark datasets such as UCF-Crime. However, current approaches face fundamental speed-accuracy trade-offs that limit practical deployment, particularly on resource-constrained edge devices. The integration of statistical boundary detection with efficient neural network architectures offers a promising pathway forward, potentially enabling sub-linear computational complexity while preserving detection fidelity. Future research should focus on hardware-aware algorithm design and adaptive computation frameworks that dynamically allocate resources based on scene complexity, potentially transforming how safety-critical systems perceive and respond to anomalous events in streaming video contexts.

https://doi.org/10.52202/085713-1204