Enhancing Trustworthiness in VAD with Rule-Based VLM-LLM Explanations

Mohamed Ibn Khedher, Faouzi Adjed, Joseph Kattan

IRT SystemX, 2 Boulevard Thomas Gobert 91120 Palaiseau, France
{mohamed.ibn-khedher, faouzi.adjed, joseph.kattan}@irt-systemx.fr

Abstract

Video Anomaly Detection is a critical task for identifying unusual events in video streams, with applications ranging from public safety surveillance to industrial monitoring. Traditional VAD methods, often based on reconstruction or prediction errors, excel at detecting deviations but typically lack semantic understanding, failing to explain *why* an event is anomalous. The recent advent of Vision-Language Models and Large Language Models has introduced a new paradigm, enabling systems to interpret and reason about video content in natural language. However, existing VLM/LLM-based approaches often focus either on rich, open-ended description or on structured, rule-based reasoning, but rarely both. In this paper, we address this gap by proposing a novel hybrid framework that synergizes the strengths of descriptive and deductive models. Our approach first leverages a powerful VLM to generate detailed, contextual scene descriptions. These descriptions are then fed into a rule-driven LLM, which uses a pre-induced set of contextual rules to make a final anomaly judgment and provide a human-readable explanation grounded in the specific rule that was violated. We validate our approach on the large-scale UCF-Crime dataset and conduct an analysis of key hyperparameters, including the VLM's input prompt and the number of frames used for description. Our results demonstrate the effectiveness of the proposed architecture and offer insights into building more interpretable, reliable, and context-aware VAD systems.

Introduction

Video Anomaly Detection (VAD) is a pivotal task in computer vision, aimed at automatically identifying events that deviate from a learned norm. Its applications are widespread and critical, spanning public surveillance, traffic monitoring, industrial process control, and autonomous systems. Historically, VAD methods have predominantly relied on low-level visual features, employing techniques like reconstruction-based models (Liu et al. 2018), future-frame prediction (Luo et al. 2021), distance-based and probability-based (Wu et al. 2024).

Recent advances in Large Language Models (LLMs) (Brown et al. 2020) and their multimodal counterparts, Vision-Language Models (VLMs) (Radford et al. 2021), have unlocked unprecedented capabilities in semantic understanding, reasoning, and natural language generation. These models, pre-trained on web-scale data, possess a remarkable ability to interpret complex visual scenes. This has paved the way for a new generation of VAD systems that move beyond simple detection to provide rich, interpretable insights.

Early explorations into VLM/LLM-based VAD have followed distinct directions. Some approaches leverage VLMs to generate detailed textual descriptions of video content, allowing an LLM to assess abnormality based on its general world knowledge (Zanella et al. 2024). Others focus on creating structured reasoning frameworks, where an LLM uses a predefined or induced set of rules to check for violations (Yang et al. 2024). While powerful, these paradigms present a trade-off: descriptive methods may lack the rigor needed for specific, high-stakes contexts, whereas rule-based systems might be constrained by the expressiveness of their rules and the VLM's ability to map visual information to them.

In this paper, we argue that the future of reliable and trustworthy VAD lies in the synthesis of these approaches. We make two primary contributions:

- We propose a novel hybrid VLM/LLM framework for VAD that aims to provide rule-grounded explanations. Our system first uses a VLM to generate rich scene descriptions and then employs an LLM to perform deductive reasoning on these descriptions against a contextual rule set. This allows the system not only to detect an anomaly but also to explain it by citing the specific rule that was violated, thereby enhancing trustworthiness.
- We empirically demonstrate that the trustworthiness of VLM-based VAD is significantly dependent on prompt engineering and visual input granularity. We show that specific, anomaly-focused prompts are essential for reliable performance, and that there is an optimal trade-off in the number of frames to maximize contextual understanding without introducing noise.

The remainder of this paper is structured as follows. We first review related works in the next section, followed by the section describing the proposed approach where a detail of our proposed hybrid framework is exposed. Then, the detail of implementation is given in the followed section.

Approach	Principle	Supervision	Key Strength vs. Limitation
Traditional Approaches			
Traditional VAD	Learn a model of normal behavior via reconstruction, prediction, or classification. Anomalies are detected as deviations from this learned norm.	Unsupervised Weakly-Supervised Supervised	**Strength:** Effective at detecting statistical deviations. **Limitation:** Lacks semantic understanding; cannot explain *why* an event is anomalous.
VLM/LLM-based Approaches			
AnomalyRuler (Yang et al. 2024)	Uses an LLM to induce context-specific rules from normal samples, then performs deductive reasoning to check for rule violations in new videos.	Few-Shot	**Strength:** Explicit interpretability, rule-grounded reasoning. **Limitation:** Performance depends on the quality of induced rules and the VLM's perception accuracy.
LAVAD (Zanella et al. 2024)	A training-free pipeline that uses a VLM for frame captioning and an LLM to score the "abnormality" of these captions based on its general knowledge.	Training-Free	**Strength:** Zero-shot capability; requires no training data or fine-tuning. **Limitation:** Relies on the LLM's generic world knowledge, which may not align with domain-specific anomalies.
Holmes-VAU (Zhang et al. 2025)	A training-based framework that fine-tunes a multimodal LLM to generate detailed, multi-granular descriptions for long-term video understanding.	Training-Based	**Strength:** Produces detailed, and context-aware descriptions of events. **Limitation:** Requires significant labeled data and computational resources for fine-tuning.
Our proposed approach	Synergizes rich VLM-generated scene descriptions with LLM-based deductive reasoning against a pre-defined rule set.	Few-Shot	**Strength:** Combines the descriptive of generative models with the interpretability of rule-based reasoning. **Limitation:** Performance is contingent on the quality of both the VLM description and the comprehensiveness of the rule set.

Table 1: Summary and Comparison of Key Paradigms in Video Anomaly Detection.

Finally, we present and discuss the obtained results and conclude with future perspectives.

Related Work

Historically, VAD has been approached through various data supervision paradigms (Table 1). The most prevalent is the Unsupervised VAD, where models learn a representation of "normality" from unlabeled videos, often by training on reconstruction or prediction tasks (Kiran, Thomas, and Parakkal 2018). The core idea is that events deviating from this learned norm will produce high reconstruction or prediction errors. A second major paradigm is the Weakly-Supervised VAD, which utilizes video-level labels (Tian et al. 2021). These methods, often based on Multi-Instance Learning (MIL), learn to identify anomalous segments within videos known to contain anomalies, without needing precise temporal annotations. The Supervised VAD is the most used paradigm for abnormal detection by providing the model with precise annotations of each anomaly in the sample (Wu et al. 2024).

While effective at identifying deviations, a shared limitation of these traditional methods is their semantic opacity. Some efforts have attempted to provide explanations by linking anomaly scores to specific visual features like motion speed, or by training models on datasets with explicit action labels. However, these explanations remain constrained by predefined categories and lack the flexible, common-sense reasoning of a human observer (Şengönül et al. 2023).

To overcome the semantic limitations of traditional methods, recent research has turned to the powerful reasoning capabilities of LLMs and VLMs models (Wu et al. 2024). These models, pre-trained on vast amounts of text and im-

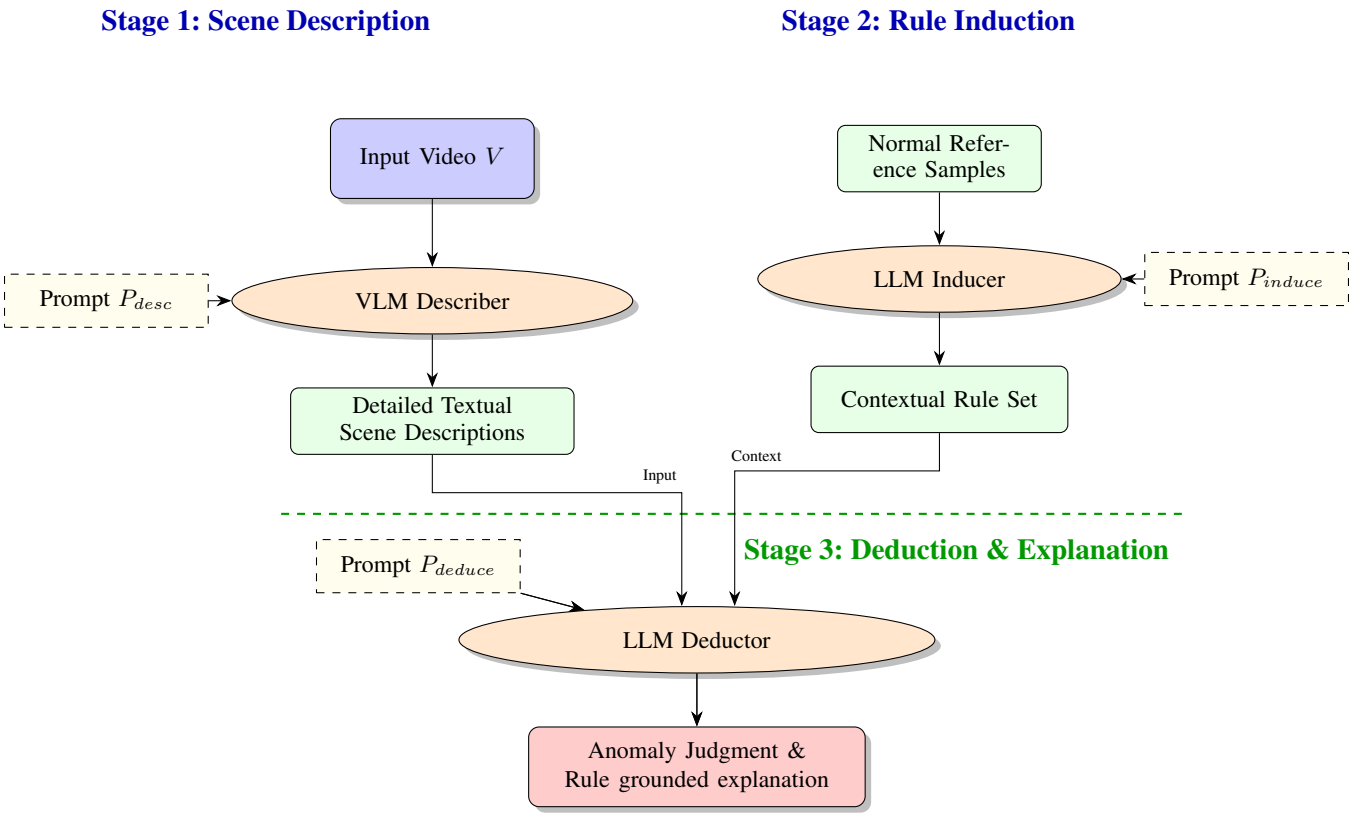


Figure 1: High-level architecture of our proposed hybrid VLM-LLM framework for video anomaly explanation. Stage 1 focuses on scene description generation using a VLM. Stage 2 (pre-computed) involves inducing a contextual rule set using an LLM. Stage 3 performs rule-driven deduction on the descriptions using LLM to provide an anomaly judgment and a grounded explanation.

age data, possess remarkable capabilities in understanding complex scenes (Tang et al. 2025), reasoning about events, and generating human-like textual descriptions. The fundamental added value is the ability of interpreting high-level concepts rather than analyzing pixel patterns. The general approach involves using a VLM to generate a textual description of the visual scene, which an LLM then assesses for abnormality.

Based on these recent approaches, Yang et al. in 2024 proposed a novel rule-based reasoning framework for VAD that explicitly leverages LLMs, named *AnomalyRuler*. Recognizing that direct LLM application may fail due to the generality of their pre-trained knowledge, *AnomalyRuler* introduces a two-stage process: induction and deduction to adapt LLM reasoning to specific VAD scenarios using only few-shot normal reference samples. Zanella et al. 2024 introduced *LAVAD* (LAnguage-based VAD), creating a truly training-free VAD paradigm by exploiting pre-trained VLMs and LLMs without any task-specific training or data collection. Zhang et al. proposed *Holmes-VAD* and *Holmes-VAU* (Zhang et al. 2024, 2025), a framework designed for long-term video anomaly understanding across various granularities (clip, event, video levels). They addressed the challenge of efficiently processing long videos and understanding anomalies with diverse temporal characteristics.

This emerging paradigm promises to transform VAD from a signal detection task into a genuine understanding and explanation task. Building on this momentum, we propose a novel hybrid framework that combines rich VLM-based scene description with rule-driven LLM reasoning. Our proposed approach, which forms a primary contribution of this paper, is detailed in the following section.

Proposed Approach

This section details our novel hybrid framework that leverages the complementary strengths of recent VLM and LLM-based approaches to provide advanced, interpretable explanations for video anomalies.

Our proposed hybrid framework operates in a three-stage process, as depicted in Figure 1:

1. **Stage 1: VLM-based scene description.** This stage employs a powerful VLM to generate rich, contextual textual narratives of the input video content. The VLM is developed in (Zhang et al. 2025).
2. **Stage 2: LLM-based contextual rule induction.** Inspired by AnomalyRuler's principles (Yang et al. 2024), this stage (which can be performed offline or once per context) uses an LLM to induce a set of behavioral rules

from few-shot normal reference samples of the target domain.

3. **Stage 3: LLM-based rule-driven deduction and explanation.** The textual descriptions from Stage 1 are then processed by another LLM, which assesses these descriptions against the rule set from Stage 2 to determine and explain anomalies.

In the following, we elaborate on each of these stages in detail.

Stage 1: VLM-Based Scene Description

The initial stage of our framework is responsible for transforming raw video input into a comprehensive textual representation, crucial for subsequent reasoning.

a) **Input video processing:** The input video V is processed to extract relevant visual information. This may involve temporal segmentation into segments V_k.

b) **VLM-based description generation:** A pre-trained VLM, denoted $\mathcal{F}_{VLM\_desc}$ is prompted with P_{desc} for each segment V_k . The VLM generates a detailed textual description $D_k = \mathcal{F}_{VLM\_desc}(V_k, P_{desc})$, capturing objects, actions, interactions, and scene context.

c) **Scene descriptions as output:** The output is a sequence of textual descriptions $\{D_k\}$ that serves as the factual basis for the subsequent deductive reasoning stage.

Stage 2: LLM-Based Contextual Rule Induction

To enable context-aware and interpretable anomaly detection, our framework incorporates a rule induction stage, inspired by AnomalyRuler (Yang et al. 2024).

a) **Normal reference data:** A small set of normal reference samples (e.g., textual descriptions of typical scenes, D_{normal_ref}, or even video clips of normal behavior) from the target domain is provided.

b) **LLM-based rule generation:** An LLM, $\mathcal{F}_{LLM\_induce}$, is prompted with P_{induce} and D_{normal_ref}. The LLM's task is to summarize normal patterns and, by contrast, infer rules that characterize potential anomalies within that specific context. This results in a contextual rule set $\mathcal{R}_{context}$, differentiating rules for normal/anomalous human activities and environmental objects.

If a sufficiently robust set of rules is already available for the domain, or if the operational context is very general, this stage might be substituted with a manually curated rule set. The primary advantage of LLM-based induction is its adaptability.

Our approach mirrors real-world unsupervised VAD, where only normal data is available and anomalies are rare and diverse. While inducing rules from normal samples alone may increase false positives, it crucially allows for the detection of novel, unseen anomaly types. To ensure transparency, the rules are stored in a simple text file, enumerating behaviors under explicit categories (e.g., *Anomalous Activities*: running, fighting; *Normal Activities*: walking). This structured, human-readable format allows the rules to be easily inspected and validated.

Stage 3: LLM-Based Rule-Driven Deduction and Explanation

With the scene descriptions $\{D_k\}$ from Stage 1 and the contextual rule set $\mathcal{R}_{context}$ from Stage 2 (or provided externally), the final stage performs deductive reasoning and generates explanations.

a) **LLM-based rule verification:** For each scene description D_k, an LLM, $\mathcal{F}_{LLM\_deduce}$ (which can be the same or different from $\mathcal{F}_{LLM\_induce}$), is employed. It takes D_k, the rule set $\mathcal{R}_{context}$, and a deductive prompt P_{deduce} as input.

b) **Anomaly judgment and explanation generation:** The LLM $\mathcal{F}_{LLM\_deduce}$ is tasked to:

- Parse D_k to identify relevant entities and actions.
- Compare these against the rules in $\mathcal{R}_{context}$.
- Output a judgment (normal/anomaly).
- Generate a natural language explanation E_k for its judgment, explicitly referencing the rule(s) from $\mathcal{R}_{context}$ that were either violated (in case of an anomaly) or adhered to.

The final output for each analyzed segment is thus $\{Judgment_k, Explanation_k\}$.

Implementation

To validate our proposed hybrid framework, we conduct experiments on a large-scale, challenging benchmark for video anomaly detection. We detail the dataset and the evaluation protocol below.

UCF-Crime Dataset

We evaluate our approach on the UCF-Crime dataset (Sultani, Chen, and Shah 2018). This is a large-scale, real-world dataset specifically designed for VAD. It consists of 1900 long, untrimmed surveillance videos with a total duration of 128 hours. The dataset is particularly well-suited for our task for several reasons:

- **Realism and diversity:** It contains 13 realistic anomaly types, including *Arson, Assault, Burglary, Explosion, Fighting,* and *Shooting*, alongside a vast amount of normal background activity. This diversity challenges the model to recognize and differentiate between various complex events.
- **Categorical labels:** Each anomaly is labeled with its category. These category labels (e.g., "Arson") serve as the ground truth for evaluating the semantic correctness of our model's generated explanations.

The official experimental protocol divides the dataset into a training set (containing only normal videos) and a testing set (containing both normal and anomalous videos), which we follow in our experiments.

To streamline our multi-class analysis, we fused the original 13 anomaly classes into 6 semantically coherent supercategories, due to their similarity. This fusion allows us to improve generalization and reduce confusion during classification, despite the importance of having multiple similar classes for more precision and more details. The resulting fused classes are shown in Table 2.

ID	Fused Class	Original Classes
1	Stealing	Stealing, Shoplifting, Burglary, Robbery
2	Assault	Shooting, Fighting, Assault, Abuse
3	Vandalism	Vandalism, Arson
4	Explosion	Explosion
5	RoadAccidents	RoadAccidents
6	Arrest	Arrest
7	Normal	Normal

Table 2: Semantic fusion of original UCF-Crime anomaly classes.

Experimental Protocol

To rigorously evaluate our approach, we systematically vary two hyperparameters that influence the generation of scene descriptions: 1) the number of input frames and 2) the content of the VLM prompt. This protocol allows us to understand their impact on the final detection and explanation quality. In this paper, we used a public VLM [1] and a public Mistral model [2].

Visual Input Granularity (Number of Frames) The number of frames provided to the VLM for generating a single description is an important hyperparameter in video processing. Using too few frames might lead to a loss of temporal context, while using too many can introduce noise or increase computational load. To investigate this trade-off, we vary the number of sampled frames, denoted as N_f, provided to the VLM. We test the following values for N_f: $N_f \in \{1, 2, 4, 6, 8, \ldots, 30\}$. This facilitates identifying the optimal visual granularity for both detection accuracy and explanation quality.

VLM Prompt Formulation The prompt given to the VLM fundamentally guides the focus and style of the generated textual description. A well-formulated prompt can steer the model toward relevant details, while a vague one may result in generic or unhelpful text. To measure this effect, we experiment with four distinct prompts, ranging from highly specific and anomaly-focused to very general and open-ended. The prompts tested are detailed in Table 3.

This experimental setup is applied across two evaluation schemes, which are detailed in the following subsections: a binary evaluation to measure the model's core ability to distinguish normal from abnormal events, and a multi-class evaluation to assess its capacity to correctly identify specific anomaly types.

Results

In this section, we present a comprehensive evaluation of our proposed hybrid framework on the UCF-Crime dataset. Our experiments are designed to assess the model's performance on both binary and multi-class anomaly detection tasks, and to analyze its sensitivity to key hyperparameters.

[1]https://huggingface.co/ppxin321/HolmesVAU-2B

[2]https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

ID	Prompt Text
P1	"Provide a summary of the video, focusing on any unusual or anomalous events."
P2	"Provide a summary of the video."
P3	"Summarize the anomaly events in the video."
P4	"Describe this video."

Table 3: VLM prompts used to guide scene description generation. The prompts are designed to test different levels of guidance, from explicit anomaly-seeking to general description.

Binary Classification Results

This section presents the results for the binary anomaly detection task. The goal here is to evaluate the model's fundamental ability to distinguish anomalous videos from normal ones, based on the descriptions generated using different prompts and frame counts. We primarily focus on the F1-score as it provides a balanced measure between precision and recall.

Performance Overview To provide a comparison, Table 4 summarizes the highest F1 score achieved for each of the four prompts, along with the corresponding number of frames (N_f) required to reach that peak.

We observe a significant disparity in performance based on the prompt's formulation. Prompt P3 ("Summarize the anomaly events in the video") clearly emerges as the top performer, achieving a maximum F1-score of 0.852 when using 14 frames. Prompt P1, which is also anomaly-focused, follows with a significant score of 0.837. In contrast, the more generic prompts P2 and P4 yield substantially lower performance, highlighting the critical importance of guiding the VLM's focus.

Prompt ID	Max F1-Score	N_f at Peak	Time (s)
P1	0.837	10	3.25
P2	0.581	16	3.47
P3	**0.852**	**14**	**3.96**
P4	0.632	30	4.55

Table 4: Summary of peak binary classification performance for each prompt. We report the maximum F1-score achieved and the number of frames (N_f) at which this peak was observed.

Performance Analysis To analyze the relationship between visual input granularity and detection performance in more detail, Figure 2 plots the F1-score as a function of the number of frames (N_f) for all four prompts.

The graph visually confirms the superiority of the anomaly-focused prompts (P1 and P3). For these prompts, a clear trend is visible: performance rapidly increases as N_f grows from 1 to around $10 - 14$ frames. This indicates that providing more temporal context is highly beneficial up to a certain point. Beyond this peak, the performance tends to

plateau or slightly decrease, suggesting that adding too many frames may introduce irrelevant information or noise that dilutes the description. Conversely, the generic prompts (P2 and P4) show only marginal improvement with additional frames, never reaching the performance levels of their more guided counterparts.

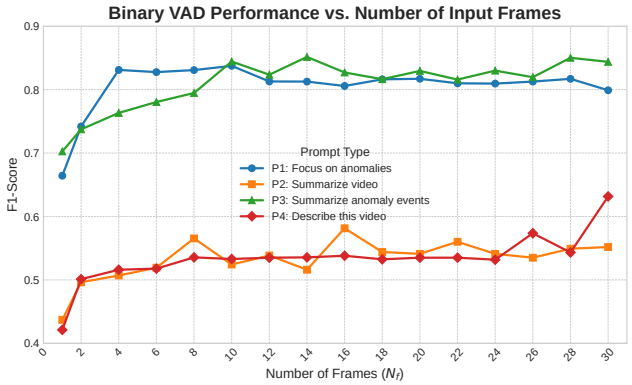


Figure 2: F1-Score for binary anomaly detection as a function of the number of input frames (N_f) for each of the four tested prompts.

Multi-Class Classification Results

While binary classification assesses the model's ability to detect anomalies, a more challenging and practical evaluation involves correctly identifying the *type* of anomaly. In this section, we evaluate the performance of our framework on a multi-class classification task. As described in our experimental setup, we group the original 13 UCF-Crime anomaly types into 6 semantically coherent "fused classes" (e.g., grouping *Stealing*, *Shoplifting*, *Burglary*, and *Robbery* into a single "Stealing" super-category) to mitigate ambiguity and focus on high-level understanding.

Performance Overview Table 5 presents the highest F1-score for each prompt in the multi-class setting. The weighted F1-score is used here as it accounts for class imbalance, providing a more robust measure of overall performance across the different anomaly types.

Similar to the binary results, a clear performance gap exists between the prompts. Prompt P3 ("Summarize the anomaly events in the video") again demonstrates superior performance, achieving the highest F1-score of 0.647 at 14 frames. Prompt P1 also performs reasonably well. The generic prompts P2 and P4 lag significantly behind, confirming that for both detection and classification, providing explicit guidance to the VLM is essential. The overall lower scores compared to the binary task are expected, as correctly classifying the anomaly type is inherently more difficult than simply detecting its presence.

Multi-Class Performance Analysis Figure **??** illustrates the trend of the weighted F1-score as a function of the number of input frames for the multi-class task.

For the top-performing prompts (P1 and P3), there is again an optimal range for N_f, generally between 6 and 14 frames, where the model receives enough context to make an informed classification without being diluted by irrelevant details. The performance of P3, in particular, shows a more pronounced peak at $N_f = 14$, suggesting that this level of visual context is crucial for distinguishing between different types of anomalies. For the generic prompts (P2 and P4), the F1-score remains low and relatively flat, indicating that without proper guidance, the VLM fails to extract the specific semantic cues necessary for accurate classification, regardless of the amount of visual information provided.

Prompt ID	Max F1-Score	N_f at Peak	Time (s)
P1	0.631	6	3.13
P2	0.496	16	3.47
P3	**0.647**	**14**	**3.96**
P4	0.568	30	4.55

Table 5: Summary of peak multi-class classification performance. We report the maximum weighted F1-score and the corresponding number of frames (N_f) for each prompt.

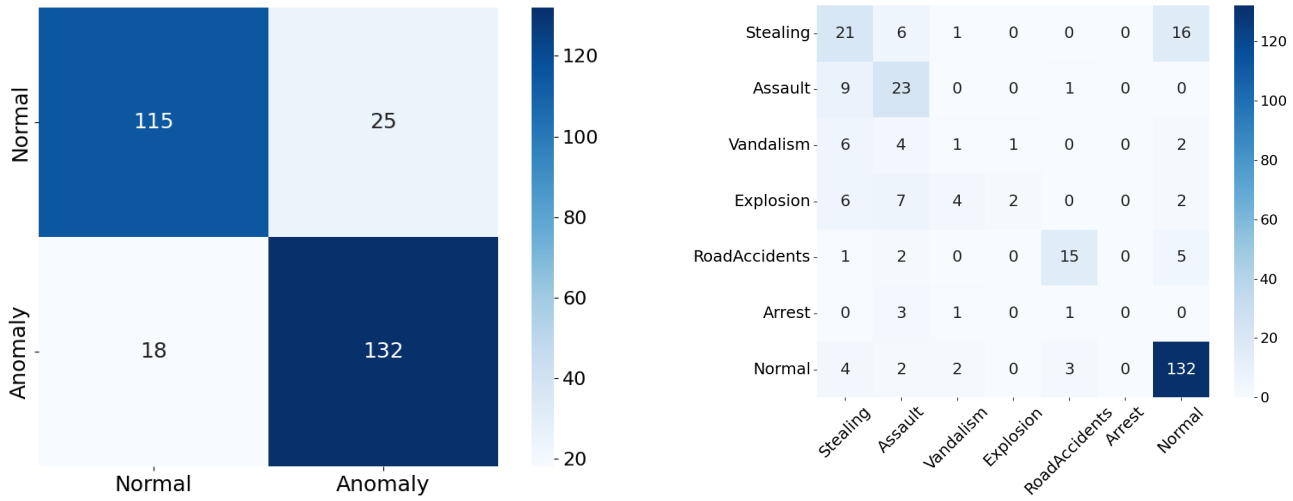
Discussion

The experimental results presented in the previous section provide several critical insights into the design and behavior of VLM/LLM-based VAD systems. Beyond the raw performance metrics, these findings have significant implications for building AI systems that are not only accurate but also reliable and trustworthy. We distill our analysis into three key observations.

Observation 1: A Trustworthy VAD Starts With a Good Prompt Our most significant finding is the important performance gap between anomaly-focused prompts (P1, P3) and generic prompts (P2, P4). This demonstrates that without explicit guidance, a powerful VLM can easily generate descriptions that are semantically correct but irrelevant to the downstream task of VAD. A generic prompt like P4 ("Describe this video.") might lead to a description of the weather or the color of a building, completely missing a subtle but critical anomalous event.

This highlights a fundamental principle for trustworthy AI: a system's behavior must be predictable and steerable. Relying on the unguided, "zero-shot" common sense of a VLM introduces a significant degree of unpredictability, making the system inherently less reliable. The success of prompts P1 and P3 shows that achieving trustworthiness in this context requires actively steering the model's focus. By instructing the VLM to look for "anomalous events," we align its "attention" with the user's intent, thereby creating a more robust and dependable system. Therefore, prompt engineering should not be seen as a mere optimization but as a core component for ensuring system reliability.

Observation 2: An Optimal Granularity for Visual Context Our analysis of the number of frames (N_f) reveals a clear trade-off. For the effective prompts, performance improves as more frames are added, but only up to an optimal point (around $10 - 14$ frames), after which it plateaus or



(a) Binary Classification.

(b) Multi-Class Classification.

Figure 3: Confusion matrices for the best-performing configuration (Prompt P3, $N_f = 14$). (a) Reconstructed from performance metrics, showing a balanced detection of Normal vs. Anomaly. (b) Illustrates plausible performance for the multi-class task, highlighting potential confusion between semantically similar categories like ‘Stealing‘ and ‘Assault‘.

even slightly degrades. This suggests that while sufficient temporal context is vital for understanding an event, an excess of visual information can lead to ”semantic dilution,” where the critical anomaly is averaged out or lost within a longer, mostly normal sequence.

This finding is important for the practical deployment of trustworthy VAD systems. A trustworthy system should be both effective and efficient (Awadid et al. 2024; Khedher, Ibn-Khedher, and Hadji 2021; Ibn-Khedher, Khedher, and Hadji 2021). Blindly assuming that ”more data is always better” can lead to a system that is not only less accurate but also computationally more expensive. Identifying this optimal visual granularity is therefore key to building a well-calibrated and efficient system. It recommends that future implementations should not treat the number of frames as an arbitrary choice but as a critical hyperparameter to be empirically validated.

Observation 3: The Gap Between Detection and Explanation Quality A comparison between the binary and multi-class results reveals a significant performance drop (e.g., peak F1-score dropping from $\sim$ 0.85 to $\sim$ 0.65). This disparity, visually captured by our confusion matrices in Figure 3, underscores a critical challenge in explainable AI: it is far easier for a model to signal that *something* is wrong than to accurately explain *what* is wrong. While the binary confusion matrix (Figure 3a) demonstrates a robust and balanced separation of Normal and Anomaly classes, the multi-class matrix (Figure 3b) highlights the system’s struggle with finer-grained distinctions. It reveals notable confusion between semantically similar categories like *Stealing* and *Assault*, proving that a deeper semantic understanding is required to move beyond simple detection.

This gap directly impacts the trustworthiness of the explanation itself. An AI system that correctly detects an anomaly but misclassifies it (e.g., flagging a medical emergency as an ”Assault”) can erode user trust and lead to incorrect real-world responses. While our framework represents a significant step towards interpretable VAD, these results highlight that achieving high-fidelity, trustworthy explanations remains a frontier. Future work should focus on advanced reasoning to achieve deeper semantic understanding.

Conclusion and Perspectives

In this paper, we addressed the critical need for more interpretable and trustworthy Video Anomaly Detection systems. Moving beyond traditional methods that lack semantic understanding, we explored the emerging landscape of VLM/LLM-based VAD. In this work, we introduced a novel hybrid framework for explainable VAD and demonstrated through a rigorous analysis that its trustworthiness is fundamentally tied to prompt design and visual input granularity.

Our experimental evaluation on the UCF-Crime dataset revealed that system reliability is critically dependent on two factors: the use of specific, anomaly-focused VLM prompts and the identification of an optimal visual input granularity.

Future work could explore more advanced reasoning techniques or prompt strategies to better capture the subtle nuances differentiating various event types. Additionally, investigating methods for model distillation and reducing the computational cost of the pipeline will be important for practical, real-world deployment. Ultimately, this work paves the way for a new generation of VAD systems that are not only effective detectors but also transparent and accountable partners in critical decision-making processes. Furthermore, we plan to quantitatively assess the quality of the generated textual explanations using standard NLP evaluation metrics.

Acknowledgments

This work has been supported by the French Government under the "France 2030" program, as part of the SystemX Technological Research Institute. The authors would also like to thank Sana TMAR, Project Manager at IRT-SystemX, for her invaluable guidance and management.

References

Awadid, A.; Amokrane-Ferka, K.; Sohier, H.; Mattioli, J.; Adjed, F.; Gonzalez, M.; and Khalfaoui, S. 2024. AI systems trustworthiness assessment: State of the art. In *Workshop on Model-based System Engineering and AI, 12th International Conference on Model-Based Software and Systems Engineering (Modelsward)*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Ibn-Khedher, H.; Khedher, M. I.; and Hadji, M. 2021. Mathematical Programming Approach for Adversarial Attack Modelling. In *ICAART (2)*, 343–350.

Khedher, M. I.; Ibn-Khedher, H.; and Hadji, M. 2021. Dynamic and Scalable Deep Neural Network Verification Algorithm. In *ICAART (2)*, 1122–1130.

Kiran, B. R.; Thomas, D. M.; and Parakkal, R. 2018. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *Journal of Imaging*, 4(2): 36.

Liu, W.; Luo, W.; Lian, D.; and Gao, S. 2018. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6536–6545.

Luo, W.; Liu, W.; Lian, D.; and Gao, S. 2021. Future frame prediction network for video anomaly detection. *IEEE transactions on pattern analysis and machine intelligence*, 44(11): 7505–7520.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Şengönül, E.; Samet, R.; Abu Al-Haija, Q.; Alqahtani, A.; Alturki, B.; and Alsulami, A. A. 2023. An analysis of artificial intelligence techniques in surveillance video anomaly detection: A comprehensive survey. *Applied Sciences*, 13(8): 4956.

Sultani, W.; Chen, C.; and Shah, M. 2018. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6479–6488.

Tang, Y.; Bi, J.; Xu, S.; Song, L.; Liang, S.; Wang, T.; Zhang, D.; An, J.; Lin, J.; Zhu, R.; et al. 2025. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*.

Tian, Y.; Pang, G.; Chen, Y.; Singh, R.; Verjans, J. W.; and Carneiro, G. 2021. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4975–4986.

Wu, P.; Pan, C.; Yan, Y.; Pang, G.; Wang, P.; and Zhang, Y. 2024. Deep learning for video anomaly detection: A review. *arXiv preprint arXiv:2409.05383*.

Yang, Y.; Lee, K.; Dariush, B.; Cao, Y.; and Lo, S.-Y. 2024. Follow the rules: Reasoning for video anomaly detection with large language models. In *European Conference on Computer Vision*, 304–322. Springer.

Zanella, L.; Menapace, W.; Mancini, M.; Wang, Y.; and Ricci, E. 2024. Harnessing large language models for training-free video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18527–18536.

Zhang, H.; Xu, X.; Wang, X.; Zuo, J.; Han, C.; Huang, X.; Gao, C.; Wang, Y.; and Sang, N. 2024. Holmes-vad: Towards unbiased and explainable video anomaly detection via multi-modal llm. *arXiv preprint arXiv:2406.12235*.

Zhang, H.; Xu, X.; Wang, X.; Zuo, J.; Huang, X.; Gao, C.; Zhang, S.; Yu, L.; and Sang, N. 2025. Holmes-vau: Towards long-term video anomaly understanding at any granularity. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 13843–13853.