# Quo Vadis, Anomaly Detection? LLMs and VLMs in the Spotlight

Xi Ding
*School of Computing*
*Australian National University*
Canberra, Australia
Xi.Ding1@anu.edu.au

Lei Wang*
*School of Computing*
*Australian National University*
Canberra, Australia
lei.w@anu.edu.au

***Abstract*—Video anomaly detection (VAD) has witnessed significant advancements through the integration of Large Language Models (LLMs) and Vision-Language Models (VLMs), offering innovative solutions to challenges such as interpretability, temporal reasoning, and generalization in dynamic and open-world scenarios. This paper provides an in-depth review of the most recent methods in 2024, emphasizing four pivotal aspects: (i) improving interpretability by using semantic insights and textual explanations, making visual anomalies more understandable, (ii) capturing complex temporal relationships for precise detection and localization of dynamic anomalies across video frames, (iii) enabling few-shot and zero-shot detection to minimize reliance on large, annotated datasets, and (iv) addressing open-world and class-agnostic anomalies through the use of semantic understanding and motion features to maintain spatio-temporal coherence. By analyzing these techniques, we highlight their potential to redefine the landscape of VAD. Additionally, we explore the synergy between visual and textual modalities, offering insights into their combined strengths and suggesting promising future directions to fully exploit the potential of LLMs and VLMs in enhancing video anomaly detection.**

***Index Terms*—anomaly detection, language models, multimodal, interpretability, open-world**

## I. Introduction

Video anomaly detection (VAD) is a critical problem with widespread applications in security surveillance, healthcare, autonomous driving, and content moderation [1]–[6]. The ability to automatically identify abnormal events or behaviors in video data is essential for real-time intervention, system optimization, and understanding complex dynamics in a variety of domains [7]. However, traditional approaches [8]–[10] to VAD face significant challenges due to the dynamic nature of video content, the complexity of detecting anomalies across various contexts, and the difficulty in obtaining labeled data for training robust models.

Recent advancements in deep learning have introduced powerful models such as large language models (LLMs) and vision-language models (VLMs), which show promising potential in enhancing VAD performance [11]–[13]. LLMs and VLMs enable a deeper understanding of both the visual and textual content of videos, offering new possibilities for detecting and explaining anomalies. These models can capture long-range temporal dependencies, understand contextual relationships, and even generate textual descriptions of video content, making them a versatile tool for improving anomaly detection in real-world, open-world scenarios.

Despite these advancements [14], [15], several challenges remain. First, most existing VAD methods struggle with capturing complex temporal relationships and context, which are often critical for understanding the evolution of anomalies over time [16]. Second, ensuring interpretability and explainability in anomaly detection is essential for real-world deployment, where transparency in decision-making is crucial [17], [18]. Third, the availability of labeled training data remains a bottleneck for many VAD systems, especially in open-world scenarios where new and previously unseen anomalies may arise [19], [20]. Finally, current methods tend to focus on class-specific anomalies, limiting their ability to generalize to open-world, class-agnostic settings [2], [21].

This work presents a comprehensive review and analysis of recent methods that combine LLMs and/or VLMs for video anomaly detection. We categorize 13 very recent (published in 2024) and closely related methods into four main areas of focus: temporal and contextual relationships, interpretability and explainability, training-free and few-shot learning approaches, and open-world/class-agnostic anomaly detection. By analyzing these categories, we highlight the strengths and limitations of each approach, offering insights into how the integration of LLMs and VLMs can advance VAD. The key **contributions** of this work are as follows:

i. We provide a categorization of the latest methods in VAD, classifying them based on their primary focus areas, including temporal modeling, interpretability, training-free learning, and open-world detection.
ii. We offer a comparative analysis of these methods, discussing the strengths and weaknesses of each approach in addressing real-world challenges in VAD.
iii. We propose future directions for research in VAD, particularly emphasizing the integration of temporal context, fine-grained interpretability, and methods for enhancing the adaptability of models to new, unseen anomalies.
iv. Finally, we identify potential synergies between existing approaches, suggesting that combining elements such as training-free methods with fine-grained semantic supervision or open-world capabilities could yield more robust

* Corresponding author.

and scalable solutions for anomaly detection.

Through this analysis, we aim to advance the understanding of how LLMs and VLMs can be used to tackle the complexities of video anomaly detection and open up new possibilities for real-world applications.

## II. Related Work

Below, we review closely related works published in 2024.

**VLMs in VAD.** The integration of visual and textual information has proven to be a powerful approach for VAD. VLMs, such as CLIP [22], have been applied in VAD to use the semantic alignment between visual features and textual descriptions. These models can capture rich contextual understanding, enhancing anomaly detection by considering both visual patterns and language-based context. One such example is `OVVAD` [23], which adopts an open-vocabulary approach by decoupling class-agnostic anomaly detection from class-specific classification. OVVAD uses image-level features from models like CLIP and excels in detecting both seen and unseen anomalies. However, its reliance on visual features alone limits its ability to effectively model temporal dependencies, which is critical for dynamic anomalies in video data. A more recent advancement, `VADor` [24], uses pretrained LLMs and VLMs for open-world anomaly detection, integrating Long-Term Context (LTC) modules to capture temporal dynamics. Although VADor successfully improves upon previous methods by adapting to unseen videos, its performance can be hindered by scalability issues when dealing with large video datasets. In contrast, `VLAVAD` [25] enhances the interpretability of VAD by combining semantic features with temporal inconsistencies through the Sequence State Space Module (S3M). This approach improves the model's ability to understand video anomalies in context, but like VADor [24], it faces challenges related to high-dimensional feature mapping, which complicates large-scale deployment.

**LLMs for VAD.** The application of large language models (LLMs) in VAD has gained traction due to their ability to aggregate contextual and temporal information from text [26]. `LAVAD` [27], a training-free model, uses LLMs to aggregate temporal descriptions and detect anomalies in video sequences. LAVAD is particularly attractive because it can adapt to new, unseen videos without requiring domain-specific training. However, its dependence on noisy or incomplete captions limits its performance, as inaccuracies in captioning can degrade the reliability of anomaly detection. `Holmes-VAD` [28] integrates multimodal LLMs to provide temporal supervision and generate multimodal instructions, offering improved anomaly localization and explanation. Holmes-VAD benefits from combining textual and visual cues, enabling it to better understand the context of anomalies. Despite these advantages, the approach struggles with scalability, especially when dealing with large-scale datasets that are common in real-world video anomaly detection scenarios.

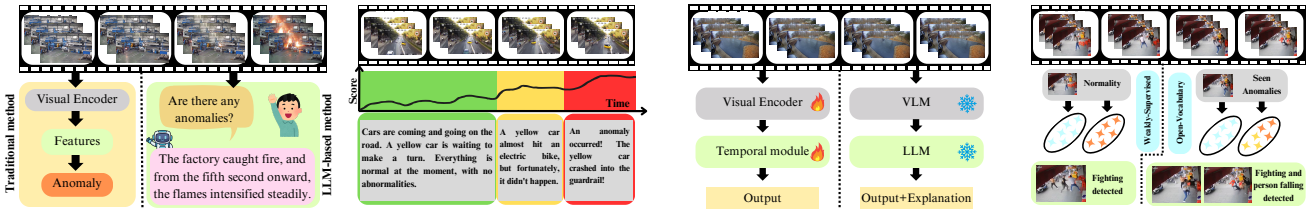
**Training-free and few-shot learning for VAD.** A key challenge in VAD is the need for large annotated datasets, which are often expensive and time-consuming to collect [29]–[31]. To address this, recent approaches [27], [32] have focused on training-free and few-shot learning techniques, which aim to detect anomalies with minimal supervision or adapt to new environments with a small amount of labeled data. `AnomalyRuler` [32] introduces a rule-based reasoning framework for few-shot learning, where a small number of normal samples are used to generate rules for anomaly detection. This approach shows promise in open-world scenarios where anomalies cannot be predefined. However, AnomalyRuler may lack robustness in handling more dynamic anomaly detection tasks, as it does not directly model temporal relationships or the finer nuances of complex video data.

On the other hand, `STPrompt` [33] uses weak supervision to learn spatio-temporal prompt embeddings from pretrained models. It excels at localizing spatial anomalies and can be effective in scenarios with limited labeled data. However, STPrompt struggles with classifying more complex anomalies, particularly those requiring a deeper understanding of temporal relationships, which are often crucial for accurate detection in video data. `ALFA` [34] introduces an adaptive strategy that uses fine-grained image-text alignments to minimize ambiguity in anomaly detection. ALFA shows the potential to improve flexibility in weakly supervised and few-shot learning scenarios, but it requires further fine-tuning to handle a broader range of anomaly types effectively.

**Open-world and class-agnostic VAD.** The need for open-world and class-agnostic anomaly detection is increasingly important in real-world applications where anomalies may not belong to predefined classes [19]. `OVVAD` [23] and `LAVAD` [27] address this challenge by offering open-vocabulary approaches that enable the detection of both known and unknown anomalies. OVVAD's dual-task approach separates class-agnostic and class-specific detection tasks, allowing it to handle a broader range of anomaly types. However, both OVVAD and LAVAD struggle with temporal consistency, particularly in dynamic video contexts where the relationship between frames and anomalies is essential. `CALLM` [35] introduces an innovative pseudo-label generation approach using 3D deep autoencoders and multimodal models, aiming to bridge the gap between video data and language models. This approach offers potential for enhancing class-agnostic detection, though its scalability and ability to handle dynamic content in video sequences need further validation.

## III. Insights on Recent Advances

This section offers a thorough analysis and discussion of the advancements in VAD in 2024, focusing on the integration of LLMs and VLMs. The methods we discussed are: `VADor` [24], `OVVAD` [23], `LAVAD` [27], `TPWNG` [36], `CALLM` [35], `Holmes-VAD` [28], `HAWK` [20], `VLAVAD` [25], `ALFA` [34], `AnomalyRuler` [32], `STPrompt` [33], `Holmes-VAU` [37], and `VERA` [18]. We categorize these closely related approaches into four main groups (Figure 1), highlighting their strategies, strengths, limitations, and potential future directions.




(a) Temporal modeling. (b) Interpretability. (c) Training-free. (d) Open-world.

Fig. 1: To systematically analyze and evaluate 13 closely related works from 2024 that use VLMs and LLMs for video anomaly detection, we categorize them into four groups: (a) temporal modeling, (b) interpretability, (c) training-free approaches, and (d) open-world detection, with each group represented by a subfigure. For each category, we highlight the employed strategies, key strengths, limitations, and outline promising directions for future research. The video frames are sourced from MSAD [6].

### A. Temporal Modeling and Context

Temporal modeling remains a cornerstone of VAD, as anomalies are often defined by deviations in temporal patterns. The challenge lies in capturing intricate temporal dynamics while maintaining computational efficiency and scalability. Recent methods address these challenges through innovative modules and integration of contextual reasoning [20], [23]–[25], [27], [28], [36].

VADor [24] addresses the limitation of open-sourced video LLMs in handling long-range context by introducing Long-Term Context (LTC) modules. This allows the method to capture temporal dynamics effectively, though scalability remains a concern as the method may struggle with longer or more complex videos. Meanwhile, LAVAD [27] aggregates temporal information through sliding windows over frame-level captions, achieving reasonable performance in structured scenarios but faltering when captions are noisy or incomplete. OVVAD [23] takes a novel approach by using graph convolutional networks (GCNs) as temporal adapters, bridging frozen CLIP encoders with sequential data. This enables effective temporal reasoning without requiring extensive retraining, though the method struggles to fully exploit fine-grained temporal cues. VLAVAD [25] uses a Sequence State Space Module (S3M) to integrate semantic inconsistencies with temporal information, achieving improved anomaly detection in unsupervised settings. However, the computational burden associated with high-dimensional state representations limits its scalability. Motion-centric approaches such as HAWK [20] use motion-to-language mappings to associate dynamic patterns with textual descriptions, offering enhanced interpretability and precision in motion anomalies. Similarly, TPWNG [36] adapts to varying video durations using self-learning modules, excelling in weakly supervised settings. Finally, Holmes-VAD [28] combines a lightweight temporal sampler with multimodal analysis, effectively identifying and explaining anomalies in complex scenarios.

These approaches highlight a rich diversity in temporal modeling strategies. VADor and OVVAD prioritize predefined modules for long-term context, while HAWK and Holmes-VAD focus on integrating motion dynamics and adaptive sampling. Future work could benefit from combining motion-based features [38]–[42] with advanced context-aware modules to address scalability and efficiency in real-time anomaly detection.

### B. Interpretability and Transparency

Interpretability has emerged as a critical factor in VAD systems, particularly for their deployment in sensitive and high-stakes environments. Methods in this category emphasize explainability by generating semantic and multimodal insights, making anomaly detection systems more comprehensible to end-users [20], [24], [25], [27], [28], [33], [34].

VADor [24] integrates textual explanations by fine-tuning Video-LLaMA's projection layer, blending anomaly detection with semantic reasoning. While effective, the reliance on instruction-tuned data may limit its adaptability to diverse anomaly types. LAVAD [27] generates scene descriptions for increased transparency, yet noisy captions undermine its reliability. By contrast, VLAVAD [25] uses low-dimensional semantic mappings to reduce complexity, sacrificing fine-grained detail for improved interpretability in unsupervised scenarios. Innovative approaches like Holmes-VAD [28] use multimodal instruction tuning and temporal supervision to generate context-rich explanations of anomalies. HAWK [20] uses interactive visual-language models, incorporating motion-based reasoning to enhance interpretability. Similarly, STPrompt [33] aligns spatiotemporal regions with learned prompts, reducing background noise and improving spatial localization. ALFA [34] emphasizes pixel-level precision through image-text alignment, though it requires additional fine-tuning to generalize effectively.

The growing focus on semantic and multimodal strategies demonstrates a promising shift towards transparent VAD systems. Methods like Holmes-VAD excel in providing contextual explanations, while ALFA offers granular insights. However, balancing granularity, semantic generalization, and computational efficiency remains an open challenge for future research.

### C. Training-Free and Few-Shot Detection

The scarcity of annotated datasets poses a significant barrier in VAD, particularly for open-world scenarios. Training-free and few-shot approaches use pre-trained models and minimal

TABLE I: Comparison of different sampling strategies for temporal reasoning.

| Sampling | Interval | Frame count | Redundancy | Target use case | Cost |
|---|---|---|---|---|---|
| **Uniform** | Fixed | Medium | Medium | Global trend | High |
| **Random** | Random | Medium | Low | Data augmentation | High |
| **Key frame** | Adaptive | Low to Med. | Low | Key event extraction | Medium |
| **Dense** | One | High | High | Fine-grained modeling | Low |
| **Sliding window** | Adaptive | Medium | Medium | Local temporal details | Medium |
| **Adaptive** | Dynamic | High | Low | Comprehensive modeling | Medium |

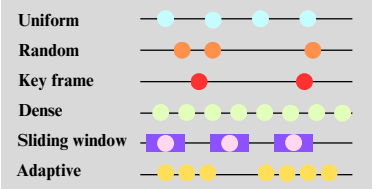


Fig. 2: Various sampling strategies.

annotations to enable anomaly detection in data-scarce environments [18], [23], [27], [32]–[34].

LAVAD [27] bypasses dataset-specific training by using pre-trained LLMs and VLMs for temporal aggregation. While adaptable, it struggles with complex anomaly types due to its lack of specialization. AnomalyRuler [32] employs rule-based reasoning with minimal normal samples, excelling in static few-shot scenarios but underperforming with dynamic or evolving anomalies. OVVAD [23] decouples anomaly detection from classification, enabling robust detection of unseen anomalies but lacking temporal depth. STPrompt [33] aligns spatiotemporal prompts to localize anomalies under weak supervision, performing well with straightforward scenarios but struggling with nuanced patterns. ALFA [34] adapts prompts dynamically at runtime for fine-grained detection, while VERA [18] introduces verbalized learning, completely avoiding model parameter modifications to enable training-free anomaly detection.

Future systems could combine the adaptability of VERA and the fine-grained capabilities of ALFA while incorporating temporal reasoning as seen in OVVAD. This synthesis could unlock solutions that address the complexities of open-world anomaly detection.

### D. Open-World and Class-Agnostic Detection

In real-world applications, anomalies are rarely predefined, necessitating open-world and class-agnostic approaches. These methods focus on detecting unseen anomalies and adapting to unexpected scenarios [23], [27], [33], [35], [37].

OVVAD [23] uses a dual-task strategy for both class-agnostic and class-specific detection but could improve with enhanced temporal modeling. LAVAD [27] uses textual descriptions for anomaly scoring but is limited by noisy captions. STPrompt [33] excels in weak supervision by localizing anomalies, though its robustness against complex patterns is limited. Holmes-VAU [37] employs hierarchical annotations for broader coverage, while CALLM [35] innovates with pseudo-labeling using multimodal features, albeit requiring further validation in dynamic contexts.

Integrating Holmes-VAU's hierarchical approach with the multimodal innovation of CALLM could pave the way for systems capable of addressing real-world complexities. Additionally, enhancing temporal and textual reasoning frameworks could improve detection reliability in open-world scenarios.

## IV. Analysis and Discussion

**Frame sampling strategies in VAD.** Table I summarizes common frame sampling strategies, while Figure 2 provides a visual comparison. In VAD, selecting the right frame sampling strategy is pivotal for balancing temporal resolution, computational cost, and model performance. Dense sampling provides the highest temporal granularity, essential for detecting subtle, complex anomalies. However, its high redundancy and computational cost make it less feasible for large-scale or real-time applications. On the other hand, strategies like uniform (*e.g.*, VERA) and random sampling (*e.g.*, randomly sample a video clip of consecutive frames as in VADor) can capture broad trends or augment data but may miss critical local temporal dependencies that are key to anomaly detection. Adaptive sampling strikes a balance by dynamically adjusting to focus on important temporal events (*e.g.*, temporal sampler in Holmes-VAD and Holmes-VAU), offering both high performance and lower redundancy, but at moderate cost. Ultimately, the choice of frame sampling strategy should align with the nature of the anomalies being detected, whether they are global trends or fine-grained, time-sensitive events, and the available computational resources.

**Quantitative evaluation and comparative analysis.** Table II provides a summary of the evaluation results. Among the methods evaluated, VLAVAD, VADor, Holmes-VAD, and STPrompt stand out for their high interpretability and temporal modeling, though they perform differently across benchmark datasets. VLAVAD excels in capturing fine-grained temporal features through fine-tuning and is highly effective on datasets such as *UCSD Ped2* (99.0%), but it lacks adaptability to open-world anomalies. In contrast, LAVAD offers interpretability with semantic explanations, but its performance on datasets like *UCF-Crime* (80.3%) and *XD-Violence* (62.0%) is limited due to its insufficient handling of temporal dynamics. This contrast highlights the importance of balancing interpretability with strong temporal modeling for real-world anomaly detection.

In terms of temporal modeling, methods such as Holmes-VAD and Holmes-VAU are more successful in addressing the temporal dependencies inherent in video anomaly detection. LAVAD offers a training-free solution with temporal aggregation, but it struggles to compete with methods like TPWNG that use spatio-temporal prompt learning. Despite AnomalyRuler achieving solid performance on the *ShanghaiTech* (85.2%) dataset, it lags behind STPrompt (97.2%), demonstrating that STPrompt's ability to adapt to temporal

TABLE II: Comparison of recent methods in video anomaly detection (VAD), focusing on key aspects: interpretability, temporal modeling, few-shot learning, and open-world detection. Performance is evaluated on five benchmark datasets: UCSD Ped2 (Ped2), CUHK Avenue (CUHK), ShanghaiTech (ShT), UCF-Crime (UCF), XD-Violence (XD), and UBnormal (UB). Datasets evaluated using AUC include Ped2, CUHK, ShT, UCF, and UB, while the dataset evaluated using AP is XD.

| Method | LLM/VLM | Property | | | | Dataset | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Interpret. | Temporal | Few-shot | Open-world | Ped2 | CUHK | ShT | UCF | XD | UB |
| **VLAVAD** [25] | Fine-tuning | ✓ | ✓ | | | 99.0 | 87.6 | 87.2 | – | – | – |
| **VADor** [24] | Fine-tuning | ✓ | ✓ | | | – | – | – | 88.1 | – | – |
| **OVVAD** [23] | Fine-tuning | | ✓ | | ✓ | – | – | – | 86.4 | 66.5 | 62.9 |
| **LAVAD** [27] | Training-free | ✓ | ✓ | ✓ | | – | – | – | 80.3 | 62.0 | – |
| **TPWNG** [36] | Fine-tuning | | ✓ | | | – | – | – | 87.8 | 83.7 | – |
| **Holmes-VAD** [28] | Fine-tuning | ✓ | ✓ | | | – | – | – | 89.5 | 90.7 | – |
| **AnomalyRuler** [32] | Fine-tuning | | | ✓ | | 97.9 | 89.7 | 85.2 | – | – | 71.9 |
| **STPrompt** [33] | Fine-tuning | ✓ | ✓ | | | – | – | 97.8 | 88.1 | – | 64.0 |
| **Holmes-VAU** [37] | Fine-tuning | | ✓ | | ✓ | – | – | – | 89.0 | 87.7 | – |
| **VERA** [18] | Training-free | ✓ | | | | – | – | – | 86.6 | 88.2 | – |

dynamics in video sequences provides a significant advantage. However, while STPrompt shows strong performance in time-sensitive anomaly detection, its dependence on fine-tuning limits its scalability and applicability to unseen anomaly types, which is a key drawback (*e.g.*, 64.0% on *UBnormal*).

Few-shot and open-world detection capabilities are critical for handling emerging or previously unseen anomalies, and methods such as OVVAD and AnomalyRuler perform well in this regard. OVVAD demonstrates the ability to detect both seen and unseen anomalies, especially with its open-vocabulary approach and class-agnostic detection. However, its performance is suboptimal in scenarios requiring temporal modeling, as seen with its results on *XD-Violence* (66.5%). On the other hand, AnomalyRuler achieves strong performance on both *UCSD Ped2* (97.9%) and *CUHK Avenue* (89.7%), showcasing its robustness in few-shot learning. Its rule-based approach, however, may struggle with more complex, dynamic anomalies, suggesting that while AnomalyRuler is effective in controlled settings, it may need further refinement for broader use cases.

Lastly, the Holmes-VAD and STPrompt methods excel in terms of interpretability, temporal modeling, and adaptability. Holmes-VAD stands out as one of the top performers, especially on the *UCF-Crime* (89.5%) and *XD-Violence* (90.7%) dataset, thanks to its combination of anomaly-aware supervision and fine-tuning, which allows it to capture both temporal and semantic features effectively. Similarly, STPrompt uses spatio-temporal prompt learning and fine-tuning to achieve excellent results on datasets like *ShanghaiTech* (97.8%) and *UCF-Crime* (88.1%). However, both methods are limited by their reliance on fine-tuning, which reduces their generalization ability across different anomaly types and datasets.

In conclusion, the results indicate that a multi-faceted approach is needed to optimize VAD systems. Methods like Holmes-VAD and STPrompt show that combining fine-tuned temporal and semantic modeling with interpretability and adaptability to new anomalies can lead to high performance across multiple datasets. However, the challenges of scalability, the need for robust temporal models, and handling noisy captions or incomplete annotations remain significant hurdles. The combination of training-free solutions with fine-tuning, as demonstrated in LAVAD, could provide a more versatile framework for open-world anomaly detection.

**Fine-tuning *vs*. training-free.** The choice between fine-tuning and training-free approaches significantly impacts performance and adaptability. Fine-tuning methods like Holmes-VAD and TPWNG excel in capturing temporal and semantic details, yielding high accuracy for known anomalies. However, they struggle with generalization to new or unseen anomaly types and are limited by the cost of continuous retraining. In contrast, training-free methods, *e.g.*, LAVAD, offer greater scalability and adaptability to diverse anomaly types without the need for retraining. While these methods are more flexible, they often fall short in capturing complex temporal dynamics, reducing accuracy in more intricate VAD tasks (*e.g.*, 80.3% on *UCF-Crime*, 62.0% on *XD-Violence*).

Future research could benefit from hybrid models that combine the scalability of training-free approaches with the accuracy of fine-tuned models. Additionally, exploring few-shot learning and open-world detection techniques could help bridge the gap, enabling VAD systems to better handle emerging anomalies with minimal retraining (*e.g.*, Holmes-VAU).

## V. Conclusion

We have explored the integration of LLMs and VLMs in VAD, focusing on their ability to address key challenges such as temporal modeling, interpretability, and few-shot learning. We categorized recent methods into four groups based on their strategies for capturing temporal relationships, enhancing interpretability, handling limited supervision, and supporting open-world anomaly detection. Our analysis highlights the strengths and limitations of current approaches, emphasizing the need for more robust temporal modeling, the importance of fine-grained interpretability, and the potential of training-free and few-shot learning methods.

We propose that combining the advantages of these strategies, such as temporal consistency, semantic feature alignment, and adaptive learning, could lead to more effective, flexible,

and interpretable VAD systems. This work sets the foundation for future research in advancing VAD by further refining these models, enhancing their scalability, and addressing the complexities of dynamic video data.

## ACKNOWLEDGMENT

Xi Ding, a Research Assistant with the Temporal Intelligence and Motion Extraction (TIME) Lab at ANU, contributed to this work. TIME Lab is a dynamic research team comprising master's and honours students focused on advancing video processing and motion analysis. This research was conducted under the supervision of Lei Wang.

## REFERENCES

[1] Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu, and Xian-Sheng Hua, "Spatio-temporal autoencoder for video anomaly detection," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1933–1941.

[2] Trong-Nguyen Nguyen and Jean Meunier, "Anomaly detection in video sequence with appearance-motion correspondence," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1273–1283.

[3] Sijie Zhu, Chen Chen, and Waqas Sultani, "Video anomaly detection for smart surveillance," in *Computer Vision: A Reference Guide*, pp. 1315–1322. Springer, 2021.

[4] Jing Ren, Feng Xia, Yemeng Liu, and Ivan Lee, "Deep video anomaly detection: Opportunities and challenges," in *2021 international conference on data mining workshops (ICDMW)*. IEEE, 2021, pp. 959–966.

[5] Yau Alhaji Samaila, Patrick Sebastian, Narinderjit Singh Sawaran Singh, Aliyu Nuhu Shuaibu, Syed Saad Azhar Ali, Temitope Ibrahim Amosa, Ghulam E Mustafa Abro, and Isiaka Shuaibu, "Video anomaly detection: A systematic review of issues and prospects," *Neurocomputing*, p. 127726, 2024.

[6] Liyun Zhu, Lei Wang, Arjun Raj, Tom Gedeon, and Chen Chen, "Advancing video anomaly detection: A concise review and a new dataset," in *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[7] Rashmika Nawaratne, Damminda Alahakoon, Daswin De Silva, and Xinghuo Yu, "Spatiotemporal anomaly detection using deep learning for real-time video surveillance," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 393–402, 2019.

[8] Venkatesh Saligrama and Zhu Chen, "Video anomaly detection based on local statistical aggregates," in *2012 IEEE Conference on computer vision and pattern recognition*. IEEE, 2012, pp. 2112–2119.

[9] Yi Hao, Jie Li, Nannan Wang, Xiaoyu Wang, and Xinbo Gao, "Spatiotemporal consistency-enhanced network for video anomaly detection," *Pattern Recognition*, vol. 121, pp. 108232, 2022.

[10] Roberto Leyva, Victor Sanchez, and Chang-Tsun Li, "Video anomaly detection with compact feature sets for online performance," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3463–3478, 2017.

[11] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[12] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[13] Xi Ding and Lei Wang, "Do language models understand time?," *arXiv preprint arXiv:2412.13845*, 2024.

[14] 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*, vol. 43, no. 3, pp. 1070–1084, 2019.

[15] Huu-Thanh Duong, Viet-Tuan Le, and Vinh Truong Hoang, "Deep learning-based anomaly detection in video surveillance: A survey," *Sensors*, vol. 23, no. 11, pp. 5024, 2023.

[16] Dan Xu, Rui Song, Xinyu Wu, Nannan Li, Wei Feng, and Huihuan Qian, "Video anomaly detection based on a hierarchical activity discovery within spatio-temporal contexts," *Neurocomputing*, vol. 143, pp. 144–152, 2014.

[17] Chongke Wu, Sicong Shao, Cihan Tunc, Pratik Satam, and Salim Hariri, "An explainable and efficient deep learning framework for video anomaly detection," *Cluster computing*, pp. 1–23, 2022.

[18] Muchao Ye, Weiyang Liu, and Pan He, "Vera: Explainable video anomaly detection via verbalized learning of vision-language models," *arXiv preprint arXiv:2412.01095*, 2024.

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

[20] Jiaqi Tang, Hao LU, RUIZHENG WU, Xiaogang Xu, Ke Ma, Cheng Fang, Bin Guo, Jiangbo Lu, Qifeng Chen, and Ying-Cong Chen, "HAWK: Learning to understand open-world video anomalies," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[21] Joey Tianyi Zhou, Jiawei Du, Hongyuan Zhu, Xi Peng, Yong Liu, and Rick Siow Mong Goh, "Anomalynet: An anomaly detection network for video surveillance," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 10, pp. 2537–2550, 2019.

[22] 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*. PMLR, 2021, pp. 8748–8763.

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

[24] Hui Lv and Qianru Sun, "Video anomaly detection and explanation via large language models," *arXiv preprint arXiv:2401.05702*, 2024.

[25] Yalong Jiang and Liquan Mao, "Vision-language models assisted unsupervised video anomaly detection," *arXiv preprint arXiv:2409.14109*, 2024.

[26] Yujiang Pu, Xiaoyu Wu, Lulu Yang, and Shengjin Wang, "Learning prompt-enhanced context features for weakly-supervised video anomaly detection," *IEEE Transactions on Image Processing*, vol. 33, pp. 4923–4936, 2024.

[27] 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*, 2024, pp. 18527–18536.

[28] Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Chuchu Han, Xiaonan Huang, Changxin Gao, Yuehuan Wang, and Nong Sang, "Holmes-vad: Towards unbiased and explainable video anomaly detection via multi-modal llm," *arXiv preprint arXiv:2406.12235*, 2024.

[29] Waqas Sultani, Chen Chen, and Mubarak Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6479–6488.

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

[31] Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah, "Ubnormal: New benchmark for supervised open-set video anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 20143–20153.

[32] Yuchen Yang, Kwonjoon Lee, Behzad Dariush, Yinzhi Cao, and Shao-Yuan Lo, "Follow the rules: reasoning for video anomaly detection with large language models," in *European Conference on Computer Vision*. Springer, 2025, pp. 304–322.

[33] Peng Wu, Xuerong Zhou, Guansong Pang, Zhiwei Yang, Qingsen Yan, Peng Wang, and Yanning Zhang, "Weakly supervised video anomaly detection and localization with spatio-temporal prompts," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 9301–9310.

[34] Jiaqi Zhu, Shaofeng Cai, Fang Deng, Beng Chin Ooi, and Junran Wu, "Do llms understand visual anomalies? uncovering llm's capabilities in zero-shot anomaly detection," in *Proceedings of the 32nd ACM International Conference on Multimedia*, New York, NY, USA, 2024, MM '24, p. 48–57, Association for Computing Machinery.

[35] Apostolos Ntelopoulos and Kamal Nasrollahi, "Callm: Cascading autoencoder and large language model for video anomaly detection,"

in *2024 IEEE Thirteenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE, 2024, pp. 1–6.

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

[37] Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, and Nong Sang, "Holmes-vau: Towards long-term video anomaly understanding at any granularity," *arXiv preprint arXiv:2412.06171*, 2024.

[38] Lei Wang and Piotr Koniusz, "Flow dynamics correction for action recognition," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 3795–3799.

[39] Lei Wang, Xiuyuan Yuan, Tom Gedeon, and Liang Zheng, "Taylor videos for action recognition," in *Forty-first International Conference on Machine Learning*.

[40] Qixiang Chen, Lei Wang, Piotr Koniusz, and Tom Gedeon, "Motion meets attention: Video motion prompts," in *The 16th Asian Conference on Machine Learning (Conference Track)*.

[41] Huilin Chen, Lei Wang, Yifan Chen, Tom Gedeon, and Piotr Koniusz, "When spatial meets temporal in action recognition," *arXiv preprint arXiv:2411.15284*, 2024.

[42] Arjun Raj, Lei Wang, and Tom Gedeon, "Tracknetv4: Enhancing fast sports object tracking with motion attention maps," *arXiv preprint arXiv:2409.14543*, 2024.