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

Anonymous ICME submission

***Abstract*—Video anomaly detection (VAD) has witnessed significant advancements through the integration of large language models (LLMs) and vision-language models (VLMs), addressing critical challenges such as interpretability, temporal reasoning, and generalization in dynamic, open-world scenarios. This paper presents an in-depth review of cutting-edge LLM-/VLM-based methods in 2024, focusing on four key aspects: (i) enhancing interpretability through semantic insights and textual explanations, making visual anomalies more understandable; (ii) capturing intricate temporal relationships to detect and localize 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 by using semantic understanding and motion features for spatiotemporal coherence. We highlight their potential to redefine the landscape of VAD. Additionally, we explore the synergy between visual and textual modalities offered by LLMs and VLMs, highlighting their combined strengths and proposing future directions to fully exploit the potential in enhancing video anomaly detection.**

***Index Terms*—review, 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]–[7]. 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 [8]. However, traditional approaches [1]–[5], [9]–[14] 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 [7], [15].

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 [16]–[20]. 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.

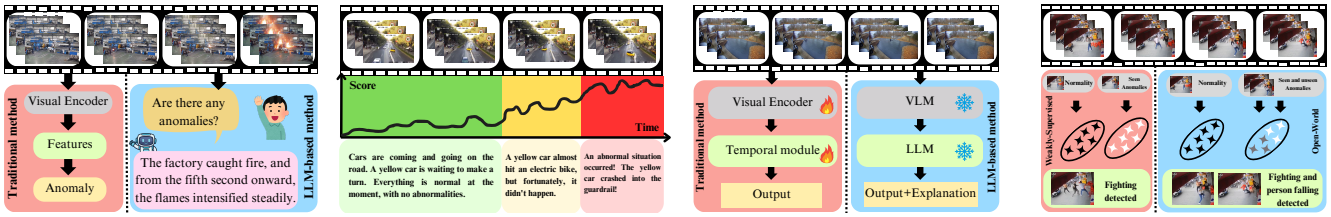
Despite these advancements [20]–[22], several challenges remain. First, most existing VAD methods struggle with capturing complex temporal relationships and context, which are often critical for understanding the evolution of anomalies over time [23]. Second, ensuring interpretability and explainability in anomaly detection is essential for real-world deployment, where transparency in decision-making is crucial [24], [25]. 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 [26], [27]. Finally, current methods tend to focus on class-specific anomalies, limiting their ability to generalize to open-world, class-agnostic settings [3], [28].

This work presents a comprehensive review and analysis of recent methods that integrate LLMs and/or VLMs for VAD. To align with current research trends, we examine 13 recently published works from 2024, exploring four critical aspects: temporal and contextual relationships, interpretability and explainability, training-free and few-shot learning approaches, and open-world/class-agnostic anomaly detection (illustrated in Figure 1). We evaluate the strengths and limitations of these approaches, offering valuable insights into how LLM and VLM integration can drive advancements in VAD. The key **contributions** of this work are as follows:

i. We identify the latest language model-driven methods, discussing 4 perspectives: temporal modeling, interpretability, training-free learning, and open-world detection.
ii. We conduct a comparative analysis of these methods, highlighting their strengths and weaknesses in addressing real-world challenges in VAD.
iii. We propose future research directions, emphasizing the integration of temporal context, fine-grained interpretability, and adaptive methods to detect new, unseen anomalies. We suggest that combining training-free approaches with fine-grained semantic supervision and open-world capabilities could enable more robust and scalable VAD solutions.

II. Related Work

Interpretability and semantic insights. Interpretability has become a crucial concern in VAD, especially in sensitive or high-stakes applications where it is essential to explain why a particular anomaly was flagged. Early methods [12]–[14] often relied on black-box models, which made it difficult to trust their predictions. Recent approaches [27], [29]–[34] have used semantic insights from LLMs and textual explanations from VLMs to generate intelligible reasoning for detected anomalies. These systems map detected visual anomalies to textual descriptions or semantic cues, making it easier for end-users to understand the nature of the anomalies. While this significantly improves transparency, the challenge remains in



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

Fig. 1: We present a systematic evaluation of 13 closely related works from 2024 that use large language models (LLMs) and vision-language models (VLMs) for video anomaly detection (VAD). The analysis is organized around four key perspectives: (a) temporal modeling, (b) interpretability, (c) training-free, and (d) open-world detection, each represented by a subfigure. For each perspective, we highlight the strategies used, key strengths, limitations, and outline promising directions for future research. The video frames used in the analysis are sourced from the MSAD [7] dataset.

balancing the granularity of these explanations with computational efficiency, especially for real-time systems.

Temporal reasoning in dynamic anomalies. Detecting and localizing dynamic anomalies that unfold over time remains one of the central challenges in VAD. Early methods [12]–[14] typically analyzed video frames independently, missing out on the temporal relationships that define many anomalies. Recent works [27], [29]–[32], [35], [36] integrating LLMs and VLMs have started to address this gap by modeling long-range dependencies between frames, enabling the detection of anomalies that span across temporal sequences. These models use advanced techniques such as motion and context modeling to improve the capture of temporal dynamics, which are crucial for identifying irregular behaviors in dynamic scenarios. However, scalability and handling noisy or incomplete data remain significant hurdles for these temporal reasoning methods.

Few-shot and zero-shot detection. The scarcity of labeled data is a persistent challenge in VAD, particularly for detecting anomalies in novel, unseen contexts [12]–[14]. Few-shot and zero-shot methods, powered by LLMs and VLMs, offer promising solutions by enabling generalization from limited labeled data [25], [30], [33]–[35], [37]. These models use pre-trained knowledge to recognize anomalies in unseen classes or with minimal training data. Methods that rely on semantic understanding of video content, combined with motion features, make it possible to identify anomalies without the need for large-scale annotated datasets. However, despite the potential, these methods often struggle with complex anomaly types that deviate significantly from the norm, especially in open-world environments where the nature of anomalies is unknown.

Open-world and class-agnostic anomalies. Traditional VAD approaches [12]–[14] typically operate in closed-world settings where predefined anomaly classes are assumed. However, real-world applications require models capable of detecting open-world, class-agnostic anomalies, which may involve previously unseen behaviors. Multimodal models [30], [33], [35], [38], [39] that combine semantic and motion reasoning are making strides in addressing these challenges by detecting anomalies without prior knowledge of the specific class. These systems are more robust in open-world settings, where they can detect unexpected anomalies, but issues related to scalability and dynamic adaptation remain unresolved, particularly when new types of anomalies appear over time.

Motivation and key differences. While existing methods have made significant strides in one or more of these areas [7], [19], [20], the integration of LLMs and VLMs offers a holistic approach to the challenges of video anomaly detection. Unlike previous works that tend to focus on isolated aspects of VAD (*e.g.*, temporal reasoning, interpretability, or class-specific detection), this paper emphasizes the synergy between visual and textual modalities to address all key challenges simultaneously. By focusing on semantic insights and motion features, this review highlights how multimodal models can provide a more comprehensive solution to video anomaly detection. Moreover, by exploring few-shot and zero-shot capabilities, this paper proposes a shift toward more generalizable systems that can perform well even with minimal training data.

III. Insights on Recent Advances

We offer a thorough analysis in VAD in 2024, with a focus on the integration of LLMs and VLMs. The methods reviewed include: `VADor` [29], `OVVAD` [35], `LAVAD` [30], `TPWNG` [36], `CALLM` [39], `Holmes-VAD` [32], `HAWK` [27], `VLAVAD` [31], `ALFA` [34], `AnomalyRuler` [37], `STPrompt` [33], `Holmes-VAU` [38], and `VERA` [25]. We structure our discussion around four key perspectives (see Figure 1) addressed by these recent advancements: temporal and contextual relationships, interpretability and explainability, training-free and few-shot learning approaches, and open-world/class-agnostic anomaly detection. For each perspective, we highlight the strategies used, evaluate the strengths and limitations of the methods, and suggest potential directions for future research.

A. Temporal Modeling and Context

Temporal modeling is fundamental to video anomaly detection (VAD), as anomalies are often characterized by deviations in temporal patterns. The primary challenge lies in capturing intricate temporal dynamics while maintaining computational efficiency and scalability. Recent methods address these challenges with innovative modules and the integration of contextual reasoning [27], [29]–[32], [35], [36].

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

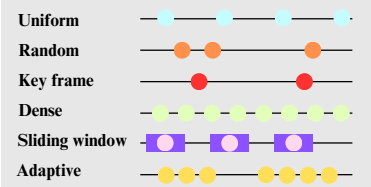


Fig. 2: Various sampling strategies.

VADor [29] introduces a Long-Term Context (LTC) module to address the limitations of open-sourced video LLMs in handling long-range context, effectively capturing temporal dynamics. However, scalability remains an issue for longer or more complex videos. LAVAD [30] uses a sliding window over frame-level captions to aggregate temporal information, achieving reasonable performance in structured scenarios but faltering with noisy or incomplete captions. OVVAD [35] uses a graph convolutional network (GCN) as a temporal adapter, bridging frozen CLIP encoders with sequential data for effective temporal reasoning without extensive retraining. However, it struggles to fully exploit fine-grained temporal cues. VLAVAD [31] integrates semantic inconsistencies with temporal information through a Sequence State Space Module (S3M), improving anomaly detection in unsupervised settings but facing scalability challenges due to high-dimensional state representations. Motion-centric approaches, such as HAWK [27], use motion-to-language mappings to connect dynamic patterns with textual descriptions, enhancing 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 [32] combines a lightweight temporal sampler with multimodal analysis, effectively identifying and explaining anomalies in complex scenarios.

These approaches showcase a diverse range of temporal modeling strategies. While VADor and OVVAD focus on predefined modules for long-term context, HAWK and Holmes-VAD emphasize motion dynamics and adaptive sampling. Future work could combine motion-based features [40]–[44] with advanced context-aware modules to address scalability and efficiency challenges in real-time anomaly detection.

B. Interpretability and Transparency

Interpretability is increasingly recognized as a critical factor in VAD systems, particularly for deployment in sensitive and high-stakes environments. Methods in this category focus on generating semantic and multimodal insights, making anomaly detection systems more comprehensible to end-users [27], [29]–[34].

VADor [29] enhances interpretability by fine-tuning Video-LLaMA's projection layer, blending anomaly detection with semantic reasoning. However, its reliance on instruction-tuned data limits adaptability to diverse anomaly types. LAVAD [30] increases transparency through scene descriptions, though noisy captions undermine reliability. In contrast, VLAVAD [31] simplifies semantic mappings to improve interpretability in unsupervised settings, sacrificing fine-grained detail for reduced complexity. Holmes-VAD [32] uses multimodal instruction tuning and temporal supervision to generate context-rich explanations of anomalies. HAWK [27] integrates motion-based reasoning via interactive visual-language models, enhancing interpretability. Similarly, STPrompt [33] aligns spatiotemporal regions with learned prompts, reducing background noise and improving spatial localization. ALFA [34] emphasizes pixel-level precision using image-text alignment but requires additional fine-tuning for effective generalization.

The emphasis on semantic and multimodal strategies marks a promising shift toward transparent VAD systems. While Holmes-VAD excels in providing contextual explanations, ALFA offers granular insights. Future research should balance granularity, semantic generalization, and computational efficiency to develop robust, interpretable VAD systems.

C. Training-Free and Few-Shot Detection

The scarcity of annotated datasets presents a significant challenge for VAD, especially in open-world scenarios. Training-free and few-shot approaches use pre-trained models and minimal annotations to facilitate anomaly detection in data-scarce environments [25], [30], [33]–[35], [37].

LAVAD [30] bypasses dataset-specific training by using pre-trained LLMs and VLMs for temporal aggregation. While adaptable, its lack of specialization hinders performance with complex anomaly types. AnomalyRuler [37] excels in static few-shot scenarios using rule-based reasoning with minimal normal samples but struggles with dynamic anomalies. OVVAD [35] 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 in straightforward cases but faltering with nuanced patterns. ALFA [34] dynamically adapts prompts at runtime for fine-grained detection, and VERA [25] introduces verbalized learning to enable training-free anomaly detection without modifying model parameters.

Combining the adaptability of VERA with the fine-grained capabilities of ALFA, alongside temporal reasoning as seen in OVVAD, could provide a pathway to more robust solutions for open-world anomaly detection.

D. Open-World and Class-Agnostic Detection

Real-world applications demand VAD systems capable of detecting unseen anomalies and adapting to unpredictable

TABLE II: Comparison of recent methods in video anomaly detection (VAD). We compare recent approaches in VAD, highlighting key aspects such as interpretability, temporal modeling, few-shot learning, and open-world detection. Performance is evaluated across six benchmark datasets: UCSD Ped2 (Ped2) [45], CUHK Avenue (CUHK) [46], ShanghaiTech (ShT) [47], UCF-Crime (UCF) [12], XD-Violence (XD) [13], and UBnormal (UB) [14]. Datasets evaluated using Area Under the Curve (AUC) include Ped2, CUHK, ShT, UCF, and UB, while the XD dataset is evaluated using Average Precision (AP).

Method	LLM/VLM	Property				Dataset					
		Interpret.	Temporal	Few-shot	Open-world	Ped2	CUHK	ShT	UCF	XD	UB
VLAVAD [31]	Fine-tuning	✓	✓			99.0	87.6	87.2	–	–	–
VADor [29]	Fine-tuning	✓	✓			–	–	–	88.1	–	–
OVVAD [35]	Fine-tuning		✓		✓	–	–	–	86.4	66.5	62.9
LAVAD [30]	Training-free	✓	✓	✓		–	–	–	80.3	62.0	–
TPWNG [36]	Fine-tuning		✓			–	–	–	87.8	83.7	–
Holmes-VAD [32]	Fine-tuning	✓	✓			–	–	–	89.5	90.7	–
AnomalyRuler [37]	Fine-tuning			✓		97.9	89.7	85.2	–	–	71.9
STPrompt [33]	Fine-tuning	✓	✓			–	–	97.8	88.1	–	64.0
Holmes-VAU [38]	Fine-tuning		✓		✓	–	–	–	89.0	87.7	–
VERA [25]	Training-free	✓				–	–	–	86.6	88.2	–

scenarios. Open-world and class-agnostic approaches aim to address these challenges [30], [33], [35], [38], [39].

OVVAD [35] uses a dual-task strategy for both class-agnostic and class-specific detection, though its temporal modeling could be enhanced. LAVAD [30] uses textual descriptions for anomaly scoring but is limited by noisy captions. STPrompt [33] excels in weak supervision, localizing anomalies effectively, though its robustness against complex patterns is limited. Holmes-VAU [38] uses hierarchical annotations for broader coverage, while CALLM [39] innovates with pseudo-labeling using multimodal features, though further validation in dynamic contexts is needed.

Integrating the hierarchical annotation approach of Holmes-VAU with the multimodal innovation of CALLM could address real-world complexities. Further advancements in temporal and textual reasoning frameworks are essential to enhance detection reliability in open-world scenarios.

IV. Analysis and Discussion

Frame sampling strategies. Frame sampling strategies play a pivotal role in balancing temporal resolution, computational efficiency, and overall model performance. Table I summarizes common strategies, while Figure 2 visually compares them. Dense sampling offers the highest temporal granularity, essential for detecting nuanced, rapid anomalies such as sudden behavioral changes or fleeting events. However, the redundancy of densely sampled frames increases computational costs, making this strategy less practical for large-scale or real-time applications. Uniform sampling, used in methods like VERA, provides a simpler alternative by sampling frames at fixed intervals. This approach balances computational overhead and temporal coverage but often misses critical local temporal patterns. Similarly, random sampling, such as in VADor, introduces variability, augmenting training data by exposing the model to diverse temporal patterns. However, this strategy risks overlooking key anomaly-defining frames, reducing its effectiveness in scenarios requiring precise temporal modeling. Adaptive sampling, used in Holmes-VAD and Holmes-VAU, dynamically focuses on regions of interest in time. This method prioritizes frames likely to contain anomalies, enabling both fine-grained detection and computational efficiency. Adaptive strategies strike an optimal balance, excelling in scenarios where anomalies are temporally sparse or context-dependent. Nonetheless, their reliance on additional heuristic or learning mechanisms introduces moderate costs.

The choice of sampling strategy should align with the nature of anomalies and the operational constraints. For global trends, uniform or random sampling suffices, while dense or adaptive sampling is indispensable for fine-grained, time-sensitive detection tasks. Integrating adaptive mechanisms into training-free frameworks, as a future direction, could enhance both scalability and precision in VAD systems.

Fine-tuning *vs*. training-free approaches. Fine-tuning-based methods dominate in datasets requiring detailed temporal reasoning. For example, Holmes-VAD achieves 89.5% on *UCF-Crime* and 90.7% on *XD-Violence*, thanks to anomaly-aware fine-tuning that captures temporal and semantic patterns. STPrompt, using spatio-temporal prompts, performs exceptionally on *ShanghaiTech* (97.8%) but requires retraining for each dataset, limiting scalability. Training-free methods like LAVAD and VERA excel in scalability, avoiding the overhead of retraining while maintaining competitive performance. VERA achieves 88.2% on *XD-Violence*, demonstrating its adaptability to new scenarios. However, they may struggle with complex temporal dynamics, *e.g.*, LAVAD's lower performance on *XD-Violence* (62.0%) and *UCF-Crime* (80.3%).

A hybrid approach combining training-free scalability with fine-tuning precision could address these limitations. For instance, integrating temporal sampling techniques from fine-tuning-based methods into training-free frameworks may enhance their temporal reasoning without compromising scalability. Future research should also explore few-shot learning and open-vocabulary techniques to bridge gaps in generalization and adaptability, as demonstrated by Holmes-VAU's promising results (89.0% on *UCF-Crime*). This direction can enable systems to handle emerging anomalies with minimal

retraining while maintaining high accuracy.

Quantitative evaluation and comparative analysis. Table II highlights the performance and properties of recent VAD methods across benchmark datasets. 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 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 shows 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. 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.

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.

V. Conclusion

This work explores the integration of large language models (LLMs) and vision-language models (VLMs) in video anomaly detection (VAD), focusing on key challenges such as temporal modeling, interpretability, few-shot learning, and open-world anomaly detection. We examine how recent advances seek to address these challenges, highlighting both the strengths and limitations of current methods. Our analysis emphasizes the need for more robust temporal modeling to capture complex dependencies within video data, as well as the importance of fine-grained interpretability to better understand anomaly detection decisions. Additionally, we recognize the potential of training-free and few-shot learning methods, which show promise for improving scalability and adaptability in scenarios with limited supervision or previously unseen anomalies. We propose that future VAD systems could benefit from combining these approaches, such as improving temporal consistency, aligning semantic features, and incorporating adaptive learning strategies. This work lays the foundation for advancing VAD by refining these models, enhancing their scalability, and addressing the complexities inherent in dynamic video data.

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] Lei Wang, Du Q Huynh, and Moussa Reda Mansour, "Loss switching fusion with similarity search for video classification," in *2019 IEEE international conference on image processing (ICIP)*. IEEE, 2019, pp. 974–978.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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*.

[8] 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.

[9] 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.

[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] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] Dexuan Ding, Lei Wang, Liyun Zhu, Tom Gedeon, and Piotr Koniusz, "Lego: Learnable expansion of graph operators for multi-modal feature fusion," *arXiv preprint arXiv:2410.01506*, 2024.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

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

[27] 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.

[28] 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.

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

[30] 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.

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

[32] 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.

[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] 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.

[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] 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.

[38] 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.

[39] 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.

[40] 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.

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

[42] 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)*.

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

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

[45] Shu Wang and Zhenjiang Miao, "Anomaly detection in crowd scene," in *IEEE 10th International Conference on Signal Processing Proceedings*. IEEE, 2010, pp. 1220–1223.

[46] 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*, 2013, pp. 2720–2727.

[47] Weixin Luo, Wen Liu, and Shenghua Gao, "A revisit of sparse coding based anomaly detection in stacked rnn framework," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 341–349.