Video-Based Traffic Anomaly Detection with Vision-Language Models: A Survey *

Yuxin Zheng[1], Wei Zhou[2], *Graduate Student Member, IEEE*, Stefan Nastic[3] and Chen Wang[4*], *Member, IEEE*

***Abstract*— Traffic anomaly detection is a hot research topic in road safety. With the rapid growth of video data, Video-based Traffic Anomaly Detection (VTAD) becomes a core module in safe driving and the security of surveillance systems. Visual unimodal methods are usually limited to shallow modeling of visual features only and lack linguistic reference frames, facing inherent shortcomings such as limited semantic comprehension and insufficient cross-domain generalization. In recent years, the rapid development of Vision-Language Models (VLMs) provides a new paradigm for traffic anomaly detection, which significantly improves the detection robustness in complex scenes. This paper provides the first survey of traffic anomaly detection based on VLMs, focusing on the three dominant methodologies: prompt learning-based, end-to-end fine-tuning-based, and feature adapter-based. In addition, in order to support and promote further research in the field, we provide a critical review of the latest traffic anomaly datasets and related evaluation metrics. Through this survey, we hope to provide valuable references and open possible trends for researchers and practitioners in the field.**

I. Introduction

As road traffic accidents occur all over the world, the lives of approximately 1.19 million people worldwide are ended each year [1]. The cause of most of these accidents is traffic anomalies. This grim reality has driven a large number of researchers into the field of traffic anomaly research. In recent years, multi-sourced data have significantly advanced our understanding of driver behaviors [2], [3], traffic risk modeling [4], [5], [6] and so on. Furthermore, the rapid development of video surveillance technology has led to increasingly diverse video data sources. The abundant video data provide a foundation for traffic anomaly detection. Video-based Traffic Anomaly Detection (VTAD) technology has gradually become a hot direction of Intelligent Transportation System (ITS) research by virtue of its features of no physical contact and non-intrusive deployment.

Current mainstream VTAD methods primarily rely on unimodal visual analysis techniques. As shown in Figure 1. , the field has evolved from spatiotemporal feature extraction based on convolutional neural networks (CNNs) [7], [8], to leveraging generative adversarial networks (GANs) [9], [10] and memory modules [11] to improve sensitivity in anomaly detection, and further using long short-term memory networks (LSTM) [12], [13] for enhanced modeling of temporal dependencies in traffic scenes. With ongoing progress in research, spatiotemporal modeling approaches based on Vision Transformers (ViT) [14] have also emerged, further improving the global modeling capabilities for complex traffic anomaly detection.

However, traffic anomalies often involve significant scene changes, such as collisions or loss of vehicle control, which require richer background and prior knowledge than simple visual data to effectively represent the dynamic changes in driving scenarios. Moreover, the diversity and novelty characteristics of the anomaly types make it difficult for visual unimodal methods to distinguish between different traffic anomalies with similar visual features. Therefore, although visual unimodal methods have made good progress, identifying abnormal events only by analyzing visual features makes them lack the combination of visual features with meaningful context such as specific scene details and relevant linguistic information, and face certain limitations in complex traffic scenarios.

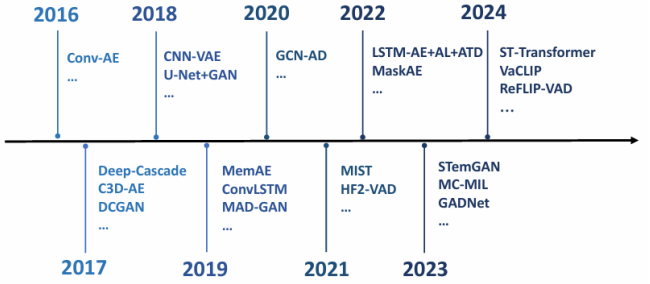

Figure 1. Development of VTAD Methods

In recent years, the rapid development of multimodal Vision-Language Models (VLMs) such as CLIP [15] has given new impetus to this field. VLMs can enhance the ability to characterize complex dynamic traffic scenes by combining textual features to obtain richer a priori visual language knowledge, achieve more accurate semantic differentiation, and also remain robust in the face of noise, occlusion, and missing information that are common in real video data.

*Research supported by the the National Key R&D Program of China [grant number 2023YFE0106800], Outstanding Youth Foundation of Jiangsu Province [grant number BK20231531], Frontier Technologies R&D Program of Jiangsu [grant number BF2024019], Austrian Research Promotion Agency (FFG) under the project RapidREC [grant number 903884].

[1]Yuxin Zheng is with Intelligent Transportation System Research Center, Southeast University, 2 Southeast University Road, Jiangsu District, Nanjing, 210096, P.R., China. (e-mail: 220243472@ seu.edu.cn).

[2]Wei Zhou is with School of Automation, Nanjing University of Science and Technology, Jiangsu District, Nanjing, 210094, P.R., China. (e-mail: vvgod@seu.edu.cn)

[3]Stefan Nastic is with Distributed Systems Group, TU Wien, Argentinierstraße 8, 1040 Vienna, Austria. (e-mail: snastic@dsg.tuwien.ac.at)

[4]Chen Wang is with Intelligent Transportation System Research Center, Southeast University, 2 Southeast University Road, Jiangsu District, Nanjing, 210096, P.R., China. (e-mail: wkobec@hotmail.com)

In addition, VTAD methods based on VLMs can provide clear semantic feedback on the information of anomalies detected in complex traffic scenarios, support autonomous driving systems or human drivers in making safer and more flexible driving decisions. Notably, further integration with the reasoning capabilities of large language models (LLMs) can offer linguistic descriptions of detected anomalies and conduct causal analyses, which can help vehicles respond to complex traffic scenarios in a timely and effective manner, further enhancing driving safety and reliability.

Existing surveys [16], [17], [18], [19] focus on the application of traditional vision methods in general-purpose traffic anomaly detection, but fail to fully explore the unique value of visual-linguistic modeling in VTAD and its domain-specific challenges. The purpose of this paper is to fill the above research gaps by providing a comprehensive overview of VLM-based traffic anomaly detection techniques, focusing on their methodological innovations, evaluation challenges and future directions. The rest of this paper is structured as follows: Section II provides a brief overview of the development of the VLMs. Section III specifically discusses the VLM-based traffic anomaly detection methodology. Experimental open-source data and evaluation metrics are outlined in Section IV. Section V dissects the challenges and looks forward to the future directions, and Section VI summarizes the research insights.

II. DEVELOPMENT OF VISION-LANGUAGE MODELS

The development of VLMs has undergone an evolution from early simple modal fusion methods to current large-scale pre-trained models. Initial research mainly used methods such as embedding splicing and bilinear pooling to simply combine image and text features for tasks such as image description generation and Visual Question Answering (VQA) [20]. With the introduction of deep learning structures such as Transformer [21], the information interaction between vision and language has become closer, and mechanisms such as cross-attention have significantly improved cross-modal understanding.

Since 2021, large-scale end-to-end pre-training has become the mainstream in the VLM field. Represented by CLIP [15] proposed by OpenAI, the model's performance on tasks such as open-domain target recognition and zero-sample inference was substantially improved by training on hundreds of millions of sets of graphical data for comparative learning, there is also a growing number of derived models based on CLIP [22], [23], [24], [25]. Series of models such as ALIGN [26] and BLIP [27] continue to innovate in data size and pre-training strategies, further promoting the generalization ability of VLMs in multimodal tasks.

The latest generation of large-scale VLMs continues to innovate in multimodal reasoning, visual understanding approaches, and model capability extensions, thus enabling them to handle more complex semantic understanding tasks. For example, Flamingo [28] enables models to reason about and generate image content given textual contexts by modeling cross-modal sequences. BLIP-2 [29] pioneered the introduction of the Q-Former mechanism based on BLIP, which efficiently embed visual features into linguistic models, enabling multimodal capabilities. LLaVA [30] introduces visual command fine-tuning for enhanced visual and verbal synergistic understanding. Video-LLaMA [31] extends the processing and comprehension capabilities of large language models for video data, especially in dynamic content comprehension, broadening its applications in multimodal scenarios. Since 2024, the latest multimodal large models such as GPT-4V [32] and DeepSeek-VL [33] exhibit stronger cross-modal understanding abilities in a variety of tasks, including image captioning, visual question answering, and complex reasoning.

However, these models rely on large-scale generalized datasets for training with rich parameterized knowledge, their generality makes them not directly applicable in specific traffic anomaly detection scenarios, and often require transfer learning.

III. VISION-LANGUAGE MODEL-DRIVEN TECHNIQUES

In recent years, VTAD based on Vision-Language Models has gradually focused on the transfer learning paradigm, which significantly reduces the computational and annotation costs of domain adaptation by migrating the cross-modal comprehension ability of the pre-trained models and adapting the parameters of the models with the characteristics of the downstream tasks. As shown in Figure 2. , current research mainly centers on three types of migration strategies: Prompt Learning, End-to-End Fine-Tuning and Feature Adapter [34]. Based on the investigation, we will review the VTAD works using VLMs from these three aspects.

A. Prompt Learning

Existing mainstream prompt learning methods can be categorized into two paradigms based on modal interaction forms: Text Prompt Learning and Visual Prompt Learning.

1) Text Prompt Learning

Text Prompt Learning utilizes the semantic understanding capability of VLMs to extract spatial and temporal features from video sequences that are highly correlated with textual prompts, which can be further categorized into three types: hand-crafted prompts, learnable prompts, and knowledge-based prompts.

Hand-crafted text prompts, i.e., manually designing text prompts or questions. Several researchers [35], [36], [37], [38] based on the chain-of-thought reasoning mechanism, drive the LLM to gradually generate fine-grained traffic anomaly detection results and analyses through staged prompts, which enhances the credibility and interpretability of the detection results.

Further, R. Liang *et al.* [39] devised a Linguistic Focusing Strategy (LFS) to enhance the understanding of traffic anomalies by using fine-grained text prompts specific to traffic events, to guide the model to adaptively focus on the visual context of interest. J. Fang *et al.* [40] on the other hand, reconstructed the textual descriptions by using antithetical verbs (e.g., “do not”) to enhance the understanding of the semantics of accidents. In addition, to address the limitation of traditional methods to solidify the anomaly categories, some recent work [41], [42] dynamically defines the semantic boundaries of anomalies by inputting user prompts into the

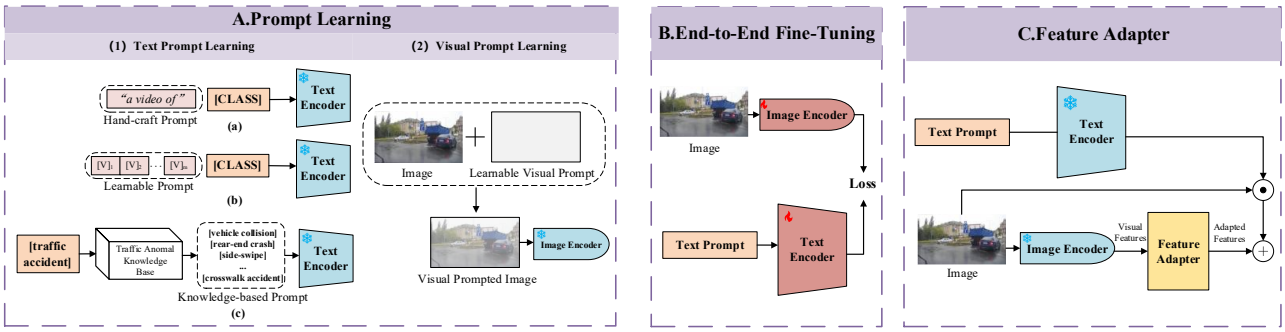

Figure 2. Vision-Language Model-Driven Techniques

text encoder, to adapt the challenge of the dynamics of the anomaly concepts in real-world scenarios.

Different from traditional static manual design, learnable text prompts adaptively learn representative video event text prompts to extract spatio-temporal features in the video that are strongly associated with the text. Several research [43], [44], [45], [46] put the anomaly category embedding and the learnable prompts in tandem into the text encoder, which are able to dynamically adapt to the semantic expressions of different anomaly categories, and make up for the domain gap of the original text encoder of CLIP in the video anomaly task. Instead, TCVADS [47] combined three types of information, namely, video-based generated text descriptions, corresponding original category labels in the dataset as well as learnable prompts, and fed them into a text encoder. Unlike the above researchers who directly concatenate learnable prompts and labels, C. Xu et al. [48] designed a learnable Domain-specific prompt module and an Anomaly-specific prompt module that generates anomaly category characterization using LLM, fusing generalized domain knowledge with fine-grained semantic descriptions specific anomalous scenarios to improve the model's accuracy in recognizing complex anomaly patterns. A reparameterized prompt encoder (DistilBERT) [49] was designed to re-parameterize input prompt embeddings to generate task-specific context-rich templates. Specifically, M. Ye et al. [50] proposed a two-stage collaborative optimization mechanism: determining video anomalies based on a guiding question by a learner agent, and dynamically adjusting the guiding question using an optimizer, which effectively reduces the induction bias of artificial rules in complex anomaly scenarios.

Knowledge-based text prompts further introduce an external knowledge base to build context-rich prompt templates, and enhances the model with structured semantic constraints on the fine-grained attributes of traffic anomalies. Y. Pu *et al.* [51] utilize an external knowledge base [52], and S. Hu *et al.*[53] introduce an external dataset of high-quality instructions [54], to construct prompts templates that provide a nuanced understanding of the specific semantics of an exception, enhancing fine-grained discriminability and inter-class separability.

2) Visual Prompt Learning

Unlike text prompt tuning, visual prompt tuning transfers VLMs by modulating the input of image encoder [34]. M. Zhang *et al.* [55] obtained implicit knowledge from training a Visual Relationship Recognition (VRR) task on the Visual Genome dataset [56] and embedded it into a frame prediction network, enabling the model to capture key object-context and object-object relationships associated with anomalies more effectively. Y. Su *et al.* [57] on the other hand, enabled the model to adaptively learn the contextual relevance of visual representations by integrating external scene-aware examples and designing a prompt likelihood learning mechanism, thereby utilizing scene prior knowledge to effectively guide the model to focus on specific anomaly types.

B. End-to-End Fine-Tuning

The end-to-end fine-tuning strategy adapts the pre-trained VLMs to the traffic anomaly detection task by directly optimizing all of its parameters. This strategy is able to fully utilize the model's representational capabilities and significantly improve the detection performance with the support of sufficient labeled data.

M. Shoman *et al.* [58] introduced the PDVC dense description model and fine-tuned its domain adaptation on the WTS Normal and Event datasets, significantly improving the model's adaptability to traffic scenarios. A. Lohner *et al.* [59] fine-tuned a classification header after multimodal alignment (visual, textual, scene graph) for fusing the three-modal embeddings and outputting incident classification results. To further reduce computational costs, Low-Rank Adaptation (LoRA) [60] has been adopted by some researchers [61], [62], [63], [64], enabling lightweight fine-tuning of low-rank matrix parameters in pre-trained Large Vision Language Models (LVLMs) without modifying the full model weights, thereby significantly reducing resource demands while preserving detection performance.

C. Feature Adapter

The feature adapter-based transfer learning approach enables the model to flexibly adapt to specific spatio-temporal anomalous features in the traffic scene without changing the original feature extraction capability by introducing a lightweight adaptation module after the feature encoder of the VLM, and training the adaptation module only. Compared with fine-tuning, this strategy reduces the computational overhead while effectively enhancing the adaptation capability of image or text features, which is especially suitable for traffic monitoring scenarios with limited data or resources.

P. Wu *et al.* [65] proposed a lightweight Temporal Adapter Module (TA) to model the contextual positional dependencies

between video frames by constructing adjacency matrices to enhance temporal dynamic feature capture. In the follow-up study [43], the authors further introduced the spatial attention aggregation module to dynamically screen key spatial anomaly regions by fusing the motion a priori and the attention weighting mechanism. In addition, P. Wu *et al.* [46], P. P. Dev *et al.* [49], and Y. Wu *et al.* [66] designed a two-stage adapter to extract local time-series features and global time-series representations, which further enhances the model's ability to characterize multi-scale anomalies. On the other hand , H. Lv *et al.* [67] and J. Tang *et al.* [68] designed an adapter for receiving original visual features with context-enhanced features and generating the input embedding of LLaMA through linear transformation, which effectively enhances the dynamic interaction of multimodal features in anomaly detection tasks.

IV. Metrics and Datasets

In this section, we systematically review publicly available traffic anomaly detection datasets (TABLE I.) for the last five years (2020-2024), and provide a comprehensive comparison in terms of dimensions such as data size, class diversity, and annotation granularity. Further, the mainstream evaluation metrics in this field are analyzed and the performance results of representative methods for the traffic anomaly detection task are summarized in TABLE II.

A. Datasets

The performance of traffic anomaly detection is highly dependent on the diversity and annotation quality of datasets. We summarize the mainstream traffic video anomaly datasets and their characteristics in the last five years through TABLE I. Some of them are depicted as follows.

DoTA[69], built by the Robotics Institute of the University of Michigan, contains 4,677 on-board camera videos, and provides temporal boundaries of anomalous events, spatial locations as well as 9 anomaly category annotations. Compared with previous datasets [70], [71] that only contain temporal annotations, DoTA is much larger in scale and realizes the assessment of anomaly localization capability for the first time. However, DoTA primarily collects data from dashboard perspectives in specific urban settings, which may bring geographical and scene diversity limitations. Consequently, models trained solely on DoTA may face generalization challenges when deployed in rural, highway, or adverse weather scenarios.

MP-RAD[72] is a synthetic dataset proposed by the Artificial Intelligence and Robotics Research Institute (AIRI) of KIST, which simulates and generates 400 unique road accidents through the gaming platform GTA-V, and each event is recorded from five independent camera angles, containing a total of 2,000 high-precision videos. Compared with real datasets [69], MP-RAD fills the research gaps such as missing data and insufficient samples from multiple viewpoints. Nevertheless, as a synthetic dataset, MP-RAD exhibits a domain gap compared to real-world scenarios, notably in visual appearance, event dynamics, and traffic participant variety, models require additional domain adaptation when applied to real-world data.

MM-AU [40] is a large-scale dataset for multimodal traffic accident understanding, which consists of several publicly available self-view accident datasets [69], [70], [71] and video clips from various video platforms, containing 11,727 videos of accidents. It offers rich event diversity, with 58 accident categories, over 2.23 million bounding boxes, and multi-level annotations including bounding boxes, accident causes, preventive suggestions, and aligned textual descriptions. This scale supports research on complex scene understanding, multi-label learning, and vision-language alignment. Yet, due to the heterogeneous video sources, MM-AU may suffer from inconsistent video quality, annotation noise, and potential regional or temporal biases.

B. Metrics and Performance Evaluation

Since video data is temporal and each frame represents a state at a point in time, the moment and duration of anomalies can be captured more accurately by detecting and evaluating each frame, so frame-level evaluation metrics are commonly used in VTAD studies, the most common being the frame-level Receiver Operating Characteristic (**ROC**) curve, Area under the ROC curve (**AUC**) and frame-level Average Precision (**AP**).

The AUC curve is a plot of true positive rate (**TPR**) versus false positive rate (**FPR**). This metric measures the overall performance of the VTAD model at different thresholds. VTAD models with higher AUC values are considered superior to models with lower AUC values. Some researchers [46], [49] have further used anomalous AUC (**Ano-AUC**) as an evaluation metric. Ano-AUC focus only on the detection performance of the anomalous category, which can more objectively reflect the model's ability to recognize anomalous events, reduce the interference of the normal category in the evaluation, and at the same time.

In addition to the ROC family, **Precision**, **Recall**, and **F1 test** values are also popular in VTAD tasks. Precision is the proportion of instances predicted by the model to be positive samples that are actually positive samples, Recall refers to the proportion of actual positive instances that are correctly predicted as positive by the model. And the frame-level average precision (**AP**), which is calculated by averaging the precision of each prediction result matched with the true label, is used to measure the average performance of the model under different recall rates. It is especially suitable for evaluating the classification and localization performance in video anomaly detection tasks.

Certainly, there are some works advocating the use of intersection and union ratio (**IoU**) and mean Average Precision (**mAP**) based on different IoU thresholds, to comprehensively evaluate the model's ability to recognize multiple anomaly types.

It's worth mentioning that H. Du *et al.* [73] proposed a multimodal metric, **MMEval**. Compared to a single-modal metric, this metric integrates multimodal inputs (video, text, and contextual annotations) to holistically evaluate causal relationships in anomaly comprehension. By designing natural language prompts and temporal importance curves, it focuses on anomalous clips to emulate human-like analysis of temporal severity shifts and causal dependencies.

TABLE I.

THE REPRESENTATIVE DATASETS FOR THE PAST FIVE YEARS FOR VTAD WITH SYNTHETIC(S) OR REAL(R), NUMBER OF SEQUENCES(SEQ.), OBSERVATION VIEWS(DASHCAM, SURVEILLANCE, AND BEV), NUMBER OF ANOMALOUS CATEGORIES(CATEG.) , AND DATASET LINKS

Datasets	Years	S/R	Seq.num	View	Categ.num	URL
RetroTrucks[74]	2020	R	474	Sur.	4	*https://drive.google.com/drive/folders/1VxFG1jHBiep4R3i_MmvMfKWH11AEFFhu*
TAD-1[75]	2021	R	500	Dash., Sur.	7	*https://github.com/ktr-hubrt/WSAL*
TaskFix[76]	2021	R	1436F*	Sur.	-	*https://bit.ly/TaskFixDataset*
USDC[77]	2022	R	122	Dash.	-	*https://public.roboflow.com/object-detection/self-driving-car*
DADA-2000[78]	2022	R	2,000	Dash.	54	*https://github.com/JWFangit/LOTVS-DADA*
DoTA[69]	2022	R	4,677	Dash.	9	*https://github.com/MoonBlvd/Detection-of-Traffic-Anomaly*
MP-RAD[72]	2023	S	2,000	Sur.	-	*https://github.com/draxler1/MP-RAD-Dataset-ITS-*
UIT-ADrone[79]	2023	R	14,021F*	BEV	-	*https://uittogether.github.io/datasets/UIT-ADrone*
CTAD[80]	2023	S	1,100	Sur.	-	*https://github.com/hankluo2/UrbanTrafficAccidentDetection*
MM-AU[40]	2024	R	11,727	Dash.	58	*https://github.com/jeffreychou777/LOTVS-MM-AU*
SO-TAD[81]	2024	R	2,186	Sur.	4	*https://github.com/cccxy-299/so-tad.*
TADS[82]	2024	R	966	Sur.	13	*https://github.com/cyc-gh/TADS*

1436F: 1436 frames. 14,021F:14,021 frame

TABLE II. PERFORMANCE OF CURRENT REPRESENTATIVE VLM-BASED METHODS FOR TRAFFIC ANOMALY DETECTION

Category			Supervision	Approach	Feature	Years	Benchmark: Metrics
Prompt Learning	Text	Hand-craft	Weakly-Supervised	TTHF[39]	CLIP	2024	DADA-2000: AUC=71.7% DoTA: AUC=84.7%
			Open World	LaGoVAD[42]	CLIP	2025	TAD-1: AUC=89.56% DoTA: AUC=62.60%
			Weakly-Supervised	STPrompt[43]	CLIP	2024	UBnormal: AUC=63.98% UCF-Crime: AUC=88.08%
		Learnable	Weakly-Supervised	TPWNG[44], DWFF[45], VadCLIP[46], TCVADS[47], ReFLIP-VAD[49]	CLIP	2024	UCF-Crime:AUC=87.79%,88.39%, 88.02%, 88.58%, 88.57% XD-Violence:AP=83.68%, 85.27%, 84.51%, 85.58%, 85.81%
			Open Vocabulary	PLOVAD[48]	CLIP	2025	UBnormal: AUC=64.35% UCF-Crime: AUC=86.78%
			Weakly-Supervised	VERA[50]	-	2024	UCF-Crime: AUC=86.55% XD-Violence: AUC=88.26%
		Knowledge-based	Weakly-Supervised	Y. Pu et al.[51]	I3D	2023	UCF-Crime: AUC=86.76% XD-Violence: AP=85.59%
	Visual		Unsupervised	CG-VAD[55]	Swin Transformer	2024	UBnormal: AUC=67.00%
			Weakly-Supervised	VPE-WSVAD[57]	Prompt	2024	UCSDped2: AUC: 99.86% Shanghai Tech: AUC= 96.88%
End-to-End Fine-Tuning			Weakly-Supervised	Holmes-VAD[63]	CLIP	2024	UCF-Crime: AUC=89.51% XD-Violence: AP=90.67%
Feature Adapter			Weakly-Supervised	VadCLIP[46], Y. Wu et al.[64]	CLIP	2024	UCF-Crime:AUC=84.51%, 87.42% XD-Violence: AP=84.51%, 82.39%
			Open Vocabulary	OVVAD[65]	CLIP	2024	UCF-Crime: AUC=86.40% XD-Violence: AP=62.94%
			Weakly-Supervised	VADor[67]	CLIP	2024	UCF-Crime: AUC=88.13% TAD-1: AUC=91.77%

V. CHALLENGES AND FUTURE DIRECTIONS

Although VLMs provide a new paradigm for traffic anomaly detection and have shown significant potential in research, their application in complex dynamic traffic scenarios still faces multiple challenges. This section systematically examines the three major challenges faced by the field of video-based traffic anomaly detection driven by VLMs, and further explores potential directions for future research.

A. Long-tail Characteristics

Traffic anomalies have a significant long-tail characteristic, reflected in the fact that high-frequency normal events constitute the majority of samples (head category), while anomalous events are diverse but each has only a small number of samples and occurs infrequently (tail category). Abnormal events in traffic scenarios include rear-end collisions, vehicle loss of control, wrong-way driving, and the incursion of non-motor vehicles or pedestrians into motor lanes. These events are diverse in type, with significant variations in their frequency within datasets, and some datasets do not cover rare anomaly types. Most existing VLM-based approaches are trained primarily on massive conventional data training, and their ability to detect traffic anomalies after migration learning is still limited, making it difficult to effectively identify critical but extremely rare traffic incidents. Therefore, enhancing the sensitivity and generalization ability of models to long-tailed or rare traffic anomalies in real-world roads and varying environments remains an urgent challenge.

In the future, meta-learning approaches can be explored to model anomaly detection tasks under varying environmental conditions, enabling VLMs to rapidly adapt the parameters upon receiving a small number of novel anomaly samples from specific traffic scenarios. This would allow for efficient transfer and generalization to new road networks, special weather conditions, or newly deployed infrastructure, thereby enhancing the detection of rare traffic anomalies in complex settings.

B. *Lack of Explainability*

In intelligent transportation systems, detecting anomalies alone does not effectively support safety decisions. Compared to general vision tasks, VTAD has a higher need for interpretability, because the interpretation results not only enhance model transparency, but also serve as a key basis for traffic management and accident accountability. Although existing VLM-based approaches have improved detection performance by fusing visual features with semantic descriptions, current research is still dominated by shallow cross-modal alignment. Only a few studies [42], [83], [84] have preliminarily explored the potential of VLMs for causal inference and interpretable representations.

Subsequent work can explore the deep integration of attribution-based approaches into VLM-based VTAD workflows. SHAP is a machine learning model interpretability tool based on game theory, which can help to locate key spatio-temporal segments and actors by calculating the mean value of marginal contribution among all possible feature combinations after feature extraction, providing a basis for traffic event traceability and key roadway supervision. Counterfactual analysis, by comparison, provides a causal level expansion for model interpretation. By constructing hypotheses in a specific traffic situation that are contrary to the current scenario (e.g., adjusting for specific vehicle speeds, positions, or pedestrian behaviors), it can help to distinguish between real anomalies and misjudgments brought about by environmental disturbances.

C. *Domain Shift*

Rare abnormal events in many traffic scenarios are difficult to collect extensively in real-world environments; as a result, researchers often rely on simulation platforms or synthetic data to augment sample sets. However, there are considerable differences between simulation and real-world roads in terms of scene complexity, participant behaviors, traffic flow patterns, and details such as urban versus rural road structures. These disparities lead to a significant decline in model generalization performance when applied to real-world traffic environments, resulting in a typical domain shift problem. To address this, on one hand, domain adaptation techniques such as adversarial networks can be used to introduce adversarial losses when training the VLM, so that the model produces as similar feature representations as possible on the source and actual data. On the other hand, locally collected real-world multi-source traffic data can be integrated, utilizing fine-grained scene attribute alignment and multi-factor simulation event reconstruction to better reflect real-world features such as varying road types, weather conditions, and traffic signal layouts, thereby improving the fidelity of simulation data to complex real scenarios.

TABLE III. summarizes the main challenges and their corresponding representative approaches.

TABLE III. CHALLENGES AND SOLUTIONS IN VLM-BASED VTAD

Challenges	Potential Solutions
Long-tail Characteristics	Meta-Learning (Model-Agnostic Meta-Learning, Prototypical Networks)
Lack of Explainability	Attribution-based approaches (SHAP, Counterfactual Analysis, CAM, Diffusion-based Attribution, GAN-CAM)
Domain Shift	Domain Adaptation (Distribution distance metric constraint-based, Adversarial learning-based), Joint Training with Multiple Sources of Data

VI. CONCLUSIONS

This paper focus on a systematic review of VLM-based traffic anomaly detection methods. After briefly reviewing the development of VLMs, the paper provides in-depth explorations around three major technological paradigms: prompt learning, end-to-end fine-tuning, and feature adapter. Meanwhile, we systematically comb through the latest key publicly available datasets, and further point out some challenges encountered in the existing research work as well as potential research directions for further investigation. From this survey, we hope VTAD problems can bring springing progress from effective models, new benchmarks, insights, and practical applications.

ACKNOWLEDGMENT

This research is supported by the National Key Research and Development Program of China (2023YFE0106800), Outstanding Youth Foundation of Jiangsu Province (BK20231531), Frontier Technologies Program of Jiangsu (BF2024019), and Austrian Research Promotion Agency (FFG) under the project RapidREC (903884).

REFERENCES

[1] World Health Organization.Road safety Accessed: Feb. 16, 2025. [Online].Available:https://www.who.int/gho/health-topics/road-safety/

[2] Y. Chen, Y. Xie, C. Wang, L. Yang, N. Zheng, and L. Wu, "Time-dependent effect of advanced driver assistance systems on driver behavior based on connected vehicle data," *Analytic Methods in Accident Research*, vol. 45, p. 100370, Mar. 2025.

[3] D. Yang, K. Ozbay, J. Gao, and F. Zuo, "A Functional Approach for Analyzing Time-Dependent Driver Response Behavior to Real-World Connected Vehicle Warnings," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 3, pp. 3438–3447, Mar. 2023.

[4] Y. Chen, Y. Xie, C. Wang, S. Xu, and L. Wu, "Temporal Dependency of Forward Collision Warning Effectiveness: A Functional Framework for Speed Profiles After Receiving Warnings," in *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*, Sep. 2024, pp. 1793–1798.

[5] Y. Chen, C. Wang, and Y. Xie, "Modeling the risk of single-vehicle run-off-road crashes on horizontal curves using connected vehicle data," *Analytic Methods in Accident Research*, vol. 43, p. 100333, Sep. 2024.

[6] L. Zhao, W. Zhou, S. Xu, Y. Chen, and C. Wang, "Multi-agent trajectory prediction at unsignalized intersections: An improved generative adversarial network accounting for collision avoidance behaviors," *Transportation Research Part C: Emerging Technologies*, vol. 171, p. 104974, Feb. 2025.

[7] S. K. Kumaran, D. P. Dogra, P. P. Roy, and A. Mitra, "Video Trajectory Classification and Anomaly Detection Using Hybrid CNN-VAE," Dec. 18, 2018, *arXiv*: arXiv:1812.07203.

[8] W. Zhou, L. Wen, Y. Zhan, and C. Wang, "An Appearance-Motion Network for Vision-Based Crash Detection: Improving the Accuracy in Congested Traffic," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 12, pp. 13742–13755, Dec. 2023.

[9] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery," Mar. 17, 2017, *arXiv*: arXiv:1703.05921.

[10] D. Li, D. Chen, L. Shi, B. Jin, J. Goh, and S.-K. Ng, "MAD-GAN: Multivariate Anomaly Detection for Time Series Data with Generative Adversarial Networks," Jan. 15, 2019, *arXiv*: arXiv:1901.04997.

[11] D. Gong *et al.*, "Memorizing Normality to Detect Anomaly: Memory-Augmented Deep Autoencoder for Unsupervised Anomaly Detection," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1705–1714.

[12] Z. Lu, W. Zhou, S. Zhang, and C. Wang, "A New Video-Based Crash Detection Method: Balancing Speed and Accuracy Using a Feature Fusion Deep Learning Framework," *Journal of Advanced Transportation*, vol. 2020, no. 1, p. 8848874, 2020.

[13] W. Zhou, Y. Yu, Y. Zhan, and C. Wang, "A vision-based abnormal trajectory detection framework for online traffic incident alert on freeways", *Neural Computing & Applications*, vol. 34, no. 17, pp. 14945–14958, Sep. 2022.

[14] T. M. Tran, D. C. Bui, T. V. Nguyen, and K. Nguyen, "Transformer-Based Spatio-Temporal Unsupervised Traffic Anomaly Detection in Aerial Videos", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 9, pp. 8292–8309, Sep. 2024.

[15] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," *Proceedings of the 38th International Conference on Machine Learning*, PMLR, Jul. 2021, pp. 8748–8763.

[16] Y. Djenouri, A. Belhadi, J. C.-W. Lin, D. Djenouri, and A. Cano, "A Survey on Urban Traffic Anomalies Detection Algorithms," *IEEE Access*, vol. 7, pp. 12192–12205, 2019.

[17] K. K. Santhosh, D. P. Dogra, and P. P. Roy, "Anomaly Detection in Road Traffic Using Visual Surveillance: A Survey," *ACM Computing Surveys*, vol. 53, no. 6, pp. 1–26, Nov. 2021.

[18] J. Fang, J. Qiao, J. Xue, and Z. Li, "Vision-Based Traffic Accident Detection and Anticipation: A Survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 4, pp. 1983–1999, Apr. 2024.

[19] W. Zhou *et al.*, "Vision Technologies with Applications in Traffic Surveillance Systems: A Holistic Survey," Nov. 30, 2024, *arXiv*: arXiv:2412.00348.

[20] S. Antol *et al.*, "VQA: Visual Question Answering," *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2425–2433.

[21] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017.

[22] L. Yao *et al.*, "FILIP: Fine-grained Interactive Language-Image Pre-Training," Nov. 09, 2021, *arXiv*: arXiv:2111.07783.

[23] N. Mu, A. Kirillov, D. Wagner, and S. Xie, "SLIP: Self-supervision Meets Language-Image Pre-training," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., Cham: Springer Nature Switzerland, 2022, pp. 529–544.

[24] B. Ni *et al.*, "Expanding Language-Image Pretrained Models for General Video Recognition," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., Cham: Springer Nature Switzerland, 2022, pp. 1–18.

[25] Y. Li *et al.*, "Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm," Mar. 14, 2022, *arXiv*: arXiv:2110.05208.

[26] C. Jia *et al.*, "Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision," *Proceedings of the 38th International Conference on Machine Learning*, PMLR, Jul. 2021, pp. 4904–4916.

[27] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," *Proceedings of the 39th International Conference on Machine Learning*, PMLR, Jun. 2022, pp. 12888–12900.

[28] J.-B. Alayrac *et al.*, "Flamingo: a Visual Language Model for Few-Shot Learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23716–23736, Dec. 2022.

[29] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," *Proceedings of the 40th International Conference on Machine Learning*, PMLR, Jul. 2023, pp. 19730–19742.

[30] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 34892–34916, Dec. 2023.

[31] H. Zhang, X. Li, and L. Bing, "Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding," Oct. 25, 2023, *arXiv*: arXiv:2306.02858.

[32] Z. Yang *et al.*, "The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision)," Oct. 11, 2023, *arXiv*: arXiv:2309.17421.

[33] H. Lu *et al.*, "DeepSeek-VL: Towards Real-World Vision-Language Understanding," Mar. 11, 2024, *arXiv*: arXiv:2403.05525.

[34] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-Language Models for Vision Tasks: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5625-5644, Aug. 2024.

[35] M. Abu Tami, H. I. Ashqar, M. Elhenawy, S. Glaser, and A. Rakotonirainy, "Using Multimodal Large Language Models (MLLMs) for Automated Detection of Traffic Safety-Critical Events," *Vehicles*, vol. 6, no. 3, pp. 1571–1590, Sep. 2024.

[36] T. Ren *et al.*, "CoT-VLM4Tar: Chain-of-Thought Guided Vision-Language Models for Traffic Anomaly Resolution," Mar. 03, 2025, *arXiv*: arXiv:2503.01632.

[37] H. Ding, Y. Du, and Z. Xia, "Urban Road Anomaly Monitoring Using Vision–Language Models for Enhanced Safety Management," *Applied Sciences*, vol. 15, no. 5, p. 2517, Feb. 2025.

[38] L. Shi, B. Jiang, T. Zeng, and F. Guo, "ScVLM: Enhancing Vision-Language Model for Safety-Critical Event Understanding," *Proceedings of the Winter Conference on Applications of Computer Vision*, 2025, pp. 1061–1071.

[39] R. Liang, Y. Li, J. Zhou, and X. Li, "Text-Driven Traffic Anomaly Detection With Temporal High-Frequency Modeling in Driving Videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 9, pp. 8684–8697, Sep. 2024.

[40] J. Fang *et al.*, "Abductive Ego-View Accident Video Understanding for Safe Driving Perception," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22030–22040.

[41] S. Ahn, Y. Jo, K. Lee, S. Kwon, I. Hong, and S. Park, "AnyAnomaly: Zero-Shot Customizable Video Anomaly Detection with LVLM," Mar. 06, 2025, *arXiv*: arXiv:2503.04504.

[42] Z. Liu, X. Wu, J. Wu, X. Wang, and L. Yang, "Language-guided Open-world Video Anomaly Detection," Mar. 17, 2025, *arXiv*: arXiv:2503.13160.

[43] P. Wu *et al.*, "Weakly Supervised Video Anomaly Detection and Localization with Spatio-Temporal Prompts," *Proceedings of the 32nd ACM International Conference on Multimedia*, Melbourne VIC Australia: ACM, Oct. 2024, pp. 9301–9310.

[44] Z. Yang, J. Liu, and P. Wu, "Text Prompt with Normality Guidance for Weakly Supervised Video Anomaly Detection," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18899–18908.

[45] H. Lim, D. Kim, M. Kim, C. Park, D. Kang, and S. Lee, "Weakly Supervised Video Anomaly Detection Using Dynamic-Weighted Feature Fusion," in *2024 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, Danang, Vietnam: IEEE, Nov. 2024, pp. 1–4.

[46] P. Wu *et al.*, "VadCLIP: Adapting Vision-Language Models for Weakly Supervised Video Anomaly Detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, Art. no. 6, Mar. 2024.

[47] W.-D. Jiang, C.-Y. Chang, H.-C. Chang, J.-Y. Chen, and D. S. Roy, "Injecting Explainability and Lightweight Design into Weakly Supervised Video Anomaly Detection Systems," Dec. 28, 2024, *arXiv*: arXiv:2412.20201.

[48] C. Xu, K. Xu, X. Jiang, and T. Sun, "PLOVAD: Prompting Vision-Language Models for Open Vocabulary Video Anomaly Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2025.

[49] P. P. Dev, R. Hazari, and P. Das, "ReFLIP-VAD: Towards Weakly Supervised Video Anomaly Detection via Vision-Language Model," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2024.

[50] M. Ye, W. Liu, and P. He, "VERA: Explainable Video Anomaly Detection via Verbalized Learning of Vision-Language Models," *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 8679–8688.

[51] Y. Pu, X. Wu, L. Yang, and S. Wang, "Learning Prompt-Enhanced Context Features for Weakly-Supervised Video Anomaly Detection," Jan. 23, 2024, *arXiv*: arXiv:2306.14451.

[52] R. Speer, J. Chin, and C. Havasi, "ConceptNet 5.5: An open multilingual graph of general knowledge," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, Feb. 2017, pp. 4444–4451.

[53] S. Hu *et al.*, "Accident LLM: A multimodal large language model of accident description based on Inter-Frame Attention Mechanism," in *2025 4th International Symposium on Computer Applications and Information Technology (ISCAIT)*, Mar. 2025, pp. 345–350.

[54] R. Goel *et al.*, "PRESTO: A Multilingual Dataset for Parsing Realistic Task-Oriented Dialogs," Mar. 17, 2023, *arXiv*: arXiv:2303.08954.

[55] M. Zhang, J. Wang, Q. Qi, Z. Zhuang, H. Sun, and J. Liao, "Cognition Guided Video Anomaly Detection Framework for Surveillance Services," *IEEE Transactions on Services Computing*, vol. 17, no. 5, pp. 2109–2123, Sep. 2024.

[56] R. Krishna *et al.*, "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, May 2017.

[57] Y. Su, Y. Tan, M. Xing, and S. An, "VPE-WSVAD: Visual prompt exemplars for weakly-supervised video anomaly detection," *Knowledge-Based Systems*, vol. 299, p. 111978, Sep. 2024.

[58] M. Shoman, D. Wang, A. Aboah, and M. Abdel-Aty, "Enhancing Traffic Safety with Parallel Dense Video Captioning for End-to-End Event Analysis," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7125–7133.

[59] A. Lohner, F. Compagno, J. Francis, and A. Oltramari, "Enhancing Vision-Language Models with Scene Graphs for Traffic Accident Understanding," in *2024 IEEE International Automated Vehicle Validation Conference (IAVVC)*, Oct. 2024, pp. 1–7.

[60] E. J. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," Oct. 16, 2021, *arXiv*: arXiv:2106.09685.

[61] J. Wu *et al.*, "A Multimodal Large Model-based Approach for Analyzing Traffic Anomalies in Highway Scenarios," *Journal of Graphics*, vol. 45, no. 6, pp. 1266–1276, 2024.

[62] K. T. Xuan, K. Nguyen Nguyen, B. H. Ngo, V. Dinh Xuan, M.-H. An, and Q.-V. Dinh, "Divide and Conquer Boosting for Enhanced Traffic Safety Description and Analysis with Large Vision Language Model," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2024, pp. 7046–7055.

[63] H. Zhang *et al.*, "Holmes-VAD: Towards Unbiased and Explainable Video Anomaly Detection via Multi-modal LLM," Jun. 29, 2024, *arXiv*: arXiv:2406.12235.

[64] Y. Yang, B. Tian, F. Yu, and Y. He, "An Anomaly Detection Model Training Method Based on LLM Knowledge Distillation," in *2024 International Conference on Networking and Network Applications (NaNA)*, Yinchuan City, China: IEEE, Aug. 2024, pp. 472–477.

[65] P. Wu *et al.*, "Open-Vocabulary Video Anomaly Detection," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18297–18307.

[66] Y. Wu, Z. Mao, C. Yu, G. Liu, and J. Shen, "Enhancing Weakly Supervised Anomaly Detection in Surveillance Videos: The CLIP-Augmented Bimodal Memory Enhanced Network," in *2024 18th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, Dec. 2024, pp. 756–762.

[67] H. Lv and Q. Sun, "Video Anomaly Detection and Explanation via Large Language Models," Jan. 11, 2024, *arXiv*: arXiv:2401.05702.

[68] J. Tang *et al.*, "HAWK: Learning to Understand Open-World Video Anomalies," *Advances in Neural Information Processing Systems*, vol. 37, pp. 139751–139785, Dec. 2024.

[69] Y. Yao, X. Wang, M. Xu, Y. Wang, E. Atkins, and D. Crandall, "DoTA: Unsupervised detection of traffic anomaly in driving videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 444–459, Jan. 2023.

[70] Y. Yao, M. Xu, Y. Wang, D. J. Crandall, and E. M. Atkins, "Unsupervised traffic accident detection in first-person videos," *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 273–280, Nov. 2019.

[71] E. P. Ijjina and S. K. Sharma, "Accident detection from dashboard camera video," *Proceedings of the 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Jul. 2019, pp. 1–4.

[72] T. K. Vijay, D. P. Dogra, H. Choi, G. Nam, and I.-J. Kim, "Detection of Road Accidents Using Synthetically Generated Multi-Perspective Accident Videos," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 2, pp. 1926–1935, Feb. 2023.

[73] H. Du *et al.*, "Uncovering What Why and How: A Comprehensive Benchmark for Causation Understanding of Video Anomaly," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18793–18803.

[74] S. Haresh, S. Kumar, M. Z. Zia, and Q.-H. Tran, "Towards Anomaly Detection in Dashcam Videos," in *2020 IEEE Intelligent Vehicles Symposium (IV)*, Oct. 2020, pp. 1407–1414.

[75] H. Lv, C. Zhou, Z. Cui, C. Xu, Y. Li, and J. Yang, "Localizing anomalies from weakly-labeled videos," *IEEE Transactions on Image Processing*, vol. 30, pp. 4505–4515, 2021.

[76] C. D. Juan, J. R. Bat-Og, K. Wan, and M. Cordel II, "Investigating visual attention-based traffic accident detection model," *Philippine Journal of Science*, vol. 150, no. 2, pp. 515–525, Jan. 2021.

[77] H. A. Yawovi, M. Kikuchi, and T. Ozono, "Who was wrong? An object detection based responsibility assessment system for crossroad vehicle collisions," *AI*, vol. 3, no. 4, pp. 844–862, Oct. 2022.

[78] J. Fang, D. Yan, J. Qiao, J. Xue, and H. Yu, "DADA: Driver Attention Prediction in Driving Accident Scenarios," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 4959–4971, Jun. 2022.

[79] T. M. Tran, T. N. Vu, T. V. Nguyen, and K. Nguyen, "UIT-ADrone: A Novel Drone Dataset for Traffic Anomaly Detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 5590–5601, 2023.

[80] H. Luo and F. Wang, "A simulation-based framework for urban traffic accident detection," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2023, pp. 1–5.

[81] X. Chen, H. Xu, M. Ruan, M. Bian, Q. Chen, and Y. Huang, "SO-TAD: A surveillance-oriented benchmark for traffic accident detection," *Neurocomputing*, vol. 618, p. 129061, Feb. 2025.

[82] Y. Chai, J. Fang, H. Liang, and W. Silamu, "TADS: a novel dataset for road traffic accident detection from a surveillance perspective," *The Journal of Supercomputing*, vol. 80, no. 18, pp. 26226–26249, Dec. 2024.

[83] Y. Yang, K. Lee, B. Dariush, Y. Cao, and S.-Y. Lo, "Follow the Rules: Reasoning for Video Anomaly Detection with Large Language Models," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., Cham: Springer Nature Switzerland, 2025, pp. 304–322.

[84] L. Zanella, W. Menapace, M. Mancini, Y. Wang, and E. Ricci, "Harnessing Large Language Models for Training-free Video Anomaly Detection," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18527–18536.