Two-Stage Optimization Framework for Traffic Anomaly Detection Based on Vision-Language Models

Siyu Wang
Key Laboratory of Road and Traffic Engineering of the Ministry of Education
Tongji University, Shanghai, China, 201804
Email: 2150644@tongji.edu.cn

Yishun Li, Ph.D.
Key Laboratory of Road and Traffic Engineering of the Ministry of Education
Tongji University, Shanghai, China, 201804
Email: lys_tongji@163.com

Yuchuan Du, Ph.D.
Key Laboratory of Road and Traffic Engineering of the Ministry of Education
Tongji University, Shanghai, China, 201804
Email: ycdu@tongji.edu.cn

George Yannis, Professor
Department of Transportation Planning and Engineering
National Technical University of Athens, Athens, Greece, GR-15773
Email: geyannis@central.ntua.gr

Chenglong Liu, Ph.D. , Corresponding Author
Key Laboratory of Road and Traffic Engineering of the Ministry of Education
Tongji University, Shanghai, China, 201804
Email: 14lcl_tj@tongji.edu.cn

Word Count: 5886 words + 3 table (250 words per table) = 6636 words

Submitted [July 30, 2025]

ABSTRACT

Timely detection of traffic anomalies is critical for ensuring road safety and enhancing traffic efficiency. While existing convolutional neural network-based unimodal and multimodal methods have shown promise, they face limitations in complex scene adaptability and deep semantic understanding. To address this, this study proposes an optimized vision-language model (VLM) framework for traffic anomaly detection, integrating static image semantic understanding with dynamic video multi-frame reasoning. The framework comprises two stages: enhancing recognition accuracy using image data and improving complex scene understanding with video data. The image dataset emphasizes concise, targeted information, while the video dataset incorporates cross-frame actions, trajectory relationships, and temporal semantics for comprehensive and accurate analysis. In the first stage, Supervised Fine-Tuning (SFT), Direct Preference Optimization (DPO) training were employed, but the model struggled with complex events like "accidents" and "abnormal parking". To address the lack of temporal context in static images, the second stage enhances dynamic scene reasoning through multi-frame temporal recognition. A unified evaluation system assesses model performance across stages. Training utilized traffic anomaly image data from Shanghai's highways and urban expressways, covering ten typical events, with semantic datasets focusing on event type, objects, location, and impact. Experimental results show the optimized model improved image understanding accuracy from 0.497 to 0.789, with redundancy reduced by 98.6% (from 0.519 to 0.007). In video reasoning tasks, accuracy for "accidents" and "abnormal parking" reached 0.59, a 64.8% improvement over static image scenarios. This study offers an efficient, viable approach for intelligent traffic systems.

Keywords: Traffic Anomaly Detection, Vision-Language Models, Multimodal Data, Fine-Tuning

INTRODUCTION

With the continuous development and increasing complexity of urban transportation systems, the rapid identification of anomalous events is critical for ensuring road safety, alleviating congestion, and improving traffic management efficiency. The widespread deployment of surveillance cameras across various levels of urban road networks has generated vast amounts of video data, enabling automated anomaly detection. Advanced image analysis and recognition techniques—particularly those based on computer vision—can extract spatiotemporal features from video streams to detect anomalies such as accidents, traffic jams, and violations. These approaches reduce the need for manual intervention and enhance the efficiency of Intelligent Transportation Systems (ITS). In recent years, computer vision-based traffic event analysis methods have made significant progress, particularly with the evolution from traditional deep learning models to large-scale multimodal models. This shift has greatly improved semantic understanding and reasoning in traffic scenes. However, real-world traffic environments present substantial challenges due to their high dynamism, semantic complexity, and heterogeneous data sources. Existing models still face notable limitations in terms of generalization, real-time performance, and deployment efficiency.

Large Vision-Language Models (VLMs) have demonstrated strong capabilities in image recognition, text generation, and multimodal alignment, offering new opportunities for anomaly detection in traffic scenarios. Nevertheless, three key bottlenecks remain in real-world applications: (1) a semantic gap between training data and task-specific demands, often leading to hallucinations and reasoning errors; (2) the requirement for robust temporal modeling and key-frame recognition in video reasoning, where existing models struggle to capture complex causal chains; and (3) the computational burden and latency of large models, which hinder real-time deployment in urban transportation systems.

To address the aforementioned challenges, this study proposes a multi-stage lightweight adaptation framework based on the Qwen2.5-VL-7B model. By integrating Supervised Fine-Tuning (SFT), Direct Preference Optimization (DPO)(*1*), and Low-Rank Adaptation (LoRA)(*2*), the framework enhances the model's semantic understanding, anomaly detection accuracy, and deployment efficiency in traffic scenarios. This approach aims to provide effective support for real-time event perception in urban traffic management. The main contributions of this study are as follows:

1. *Construction of a Localized Traffic Semantic Dataset*. A dataset is constructed around ten types of traffic events, with a focus on key elements such as event types, objects, and impacts. The design is approached from both image-text and video-text perspectives. The image dataset emphasizes simplicity and focus to improve event recognition accuracy, while the video dataset prioritizes comprehensiveness and accuracy to minimize the introduction of erroneous information.

2. *Proposal of a Multimodal Fusion Framework*. This framework combines image, video, and text data to comprehensively model event types, key objects, locations, and impacts in traffic scenarios. Through this fusion approach, the framework reduces computational overhead while enhancing semantic parsing capabilities in real-world deployments.

3. *Development of a Two-Stage Optimization Framework Based on LoRA*. A two-stage optimization framework is proposed, which combines SFT+DPO and LoRA with temporal multi-frame recognition. Different optimization strategies are applied to the image and video stages: image understanding enhances scene recognition accuracy, while video reasoning improves the model's semantic understanding ability, thus enabling precise identification of key elements in typical events.

RELATED WORKS

1. Traffic Anomaly Detection

Traffic anomaly detection has long been a key research area within ITS. Early approaches primarily relied on rule-based image processing techniques, such as optical flow estimation (*3, 4*) and background modeling (*5, 6*). However, these methods exhibit limited adaptability and high false detection rates in complex and dynamic traffic scenarios.

With the rise of deep learning, convolutional neural networks (CNNs) have become standard for

image feature extraction and temporal modeling. For example, Ijjina et al. (*7*) proposed an end-to-end framework using Mask R-CNN for object detection and centroid displacement for tracking and accident judgment, showing robustness in real-world traffic surveillance. Similarly, (*8*) introduced a hybrid model combining CNN, LSTM, and GNN to predict traffic accident risk from vehicle trajectories, leveraging CNN for spatial features, LSTM for temporal dynamics, and GNN for road network representation. While effective in detecting anomalies based on motion features like speed and trajectory, these approaches lack deep semantic understanding and contextual descriptions, limiting high-level reasoning.

To overcome the limitations of unimodal approaches, some studies have explored multimodal fusion strategies by incorporating auxiliary information such as textual descriptions, traffic flow data, or metadata to enhance traffic event classification (*9*). However, due to the complexity of data annotation and the ambiguity of task objectives, most existing fusion methods focus on shallow-level feature concatenation, such as early fusion of images with speed sensor data (*10*), or loosely aligned image-text models. While these approaches improve robustness to some extent, they fall short in enabling deep cross-modal semantic reasoning.

2. Applications of Multimodal Foundation Models

In recent years, the emergence of Vision Foundation Models has provided more powerful solutions for semantic understanding and anomaly recognition in traffic scenarios. Unlike traditional approaches, these models are pre-trained on large-scale, heterogeneous datasets using self-supervised or weakly supervised methods, enabling stronger generalization in low-label or unseen environments. For instance, CLIP (*11*), trained on massive image-text pairs from the web, pioneered the contrastive learning paradigm and significantly improved zero-shot classification and semantic alignment capabilities. Models such as BLIP and BLIP-2 (*12, 13*), as well as Flamingo (*14*), have advanced joint image-text modeling and natural language understanding, substantially enhancing performance in visual question answering (VQA) and image captioning tasks.

Recent applications of large models in transportation focus on traffic management, safety, and autonomous driving, providing new paradigms for ITS. In traffic management, PromptGAT (*15*) uses large models and domain knowledge to optimize traffic signal control, bridging the gap between simulation and real-world deployment. TrafficGPT (*16*) supports traffic decision-making through natural language interaction, though its reliance on prompt design limits automation. ST-LLM (*17*) enhances traffic forecasting using spatiotemporal embeddings and attention mechanisms, but is computationally expensive and dependent on high-quality data. In traffic safety, TrafficSafetyGPT (*18*) fine-tuned LLaMA using the TrafficSafety-2K dataset for strong performance in safety-related question answering. AccidentGPT (*19*), a multimodal model built on V2X architecture, integrates comprehensive scene understanding for proactive accident prediction.

Despite their strong cross-modal perception capabilities, these models still face three major limitations: the lack of domain-specific semantic annotations, insufficient ability to model complex temporal events and causal relationships, and high deployment costs with limited real-time performance, which constrains their applicability in ITS scenarios.

3. Optimization Methods for Fine-Tuning Pre-Trained Models

In the field of large model fine-tuning, with the widespread application of pre-trained models, various optimization methods have been proposed to enhance model performance in specific task domains. Current fine-tuning methods primarily adjust the parameters of pre-trained models to help them better adapt to the requirements of different tasks. SFT is the most commonly used supervised learning fine-tuning method, which utilizes labeled data (such as image labels or text labels) to train the model, aligning its output as closely as possible with the true labels. However, this method heavily relies on large amounts of labeled data, and the model's performance is significantly influenced by the quality of the dataset. DPO(*1*) is a supervised learning optimization method that directly uses human preference data to optimize the behavior of language models, without relying on an explicit reward model. Compared to other methods such as PPO(*20*) and Reinforcement Learning from RLAIF(*21*), DPO is simpler to train, converges more

stably, and consumes fewer resources, making it particularly suitable for resource-constrained environments. However, DPO still depends on large amounts of preference data and faces challenges in the selection of preference samples when dealing with diverse and complex tasks.

Although these methods have shown excellent performance in generative tasks, they generally have limitations in the complex task of traffic anomaly detection. Traffic anomaly events involve dynamic spatiotemporal relationships, and the data is often imbalanced and scarce, making it difficult for existing single fine-tuning methods to effectively handle these challenges. Especially in processing time-series data and capturing subtle variations in traffic anomalies, the performance of existing methods is insufficient. Therefore, overcoming these limitations and proposing fine-tuning strategies better suited for traffic anomaly detection is crucial for improving model performance.

4. Hallucination Studies and Semantic Deviation in Traffic Scenarios

Recent advancements in multimodal large models have significantly improved their ability to generate coherent image-text pairs for various applications, including traffic event recognition. However, despite these improvements, the issue of "hallucination"—where models generate descriptions of nonexistent objects or inaccurately describe object attributes—remains a critical challenge. This phenomenon can severely affect the reliability of traffic event recognition systems, leading to potential misclassifications. Hallucinations are typically categorized as Unsupported Hallucination (nonexistent objects) and Image-Text Inconsistency (semantic misalignment). Evaluation methods include the CHAIR metric (*22*), which measures unsupported objects, and the POPE framework (*23*), which detects hallucination through binary classification. As multimodal evaluation evolves, traditional text similarity metrics like BLEU and ROUGE are being replaced by semantic embedding-based methods—e.g., BERTScore (*24*), CIDEr (*25*)—and image-text alignment metrics—e.g., CLIPScore (*26*), SPICE (*27*)—offering better semantic fidelity assessment. Human evaluation criteria such as factuality and completeness are also widely used to complement automated metrics.

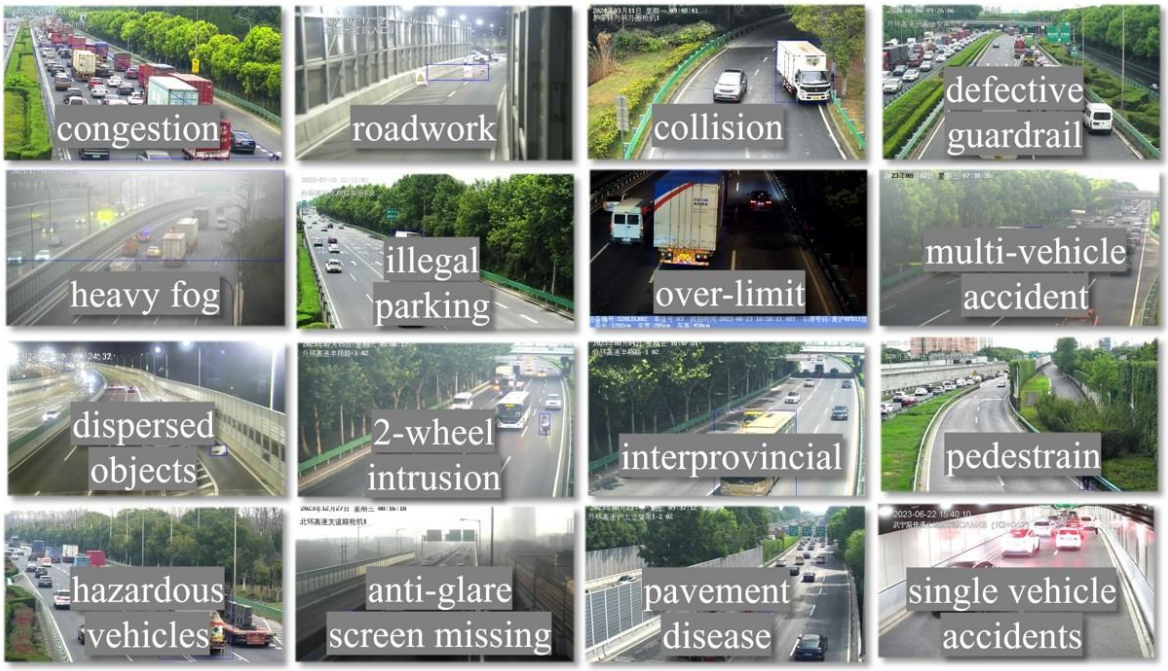
Despite progress in hallucination detection, challenges remain. Existing evaluation methods struggle to capture all forms of hallucinations, particularly in complex traffic scenarios. While semantic embedding and alignment metrics show promise, they often fail with fine-grained details in dynamic or cluttered environments, where hallucinations can be subtle. Additionally, automated metrics may not fully align with human understanding, causing models to perform well in evaluations but fall short in real-world applications.

Although this study does not propose a direct hallucination mitigation mechanism, we recognize that hallucination significantly impacts the reliability of traffic event recognition. Therefore, we incorporate dedicated evaluation metrics in our experiments to quantitatively assess hallucination-related issues, enabling a more comprehensive evaluation of generation quality and detail-level deviations in complex traffic scenarios.

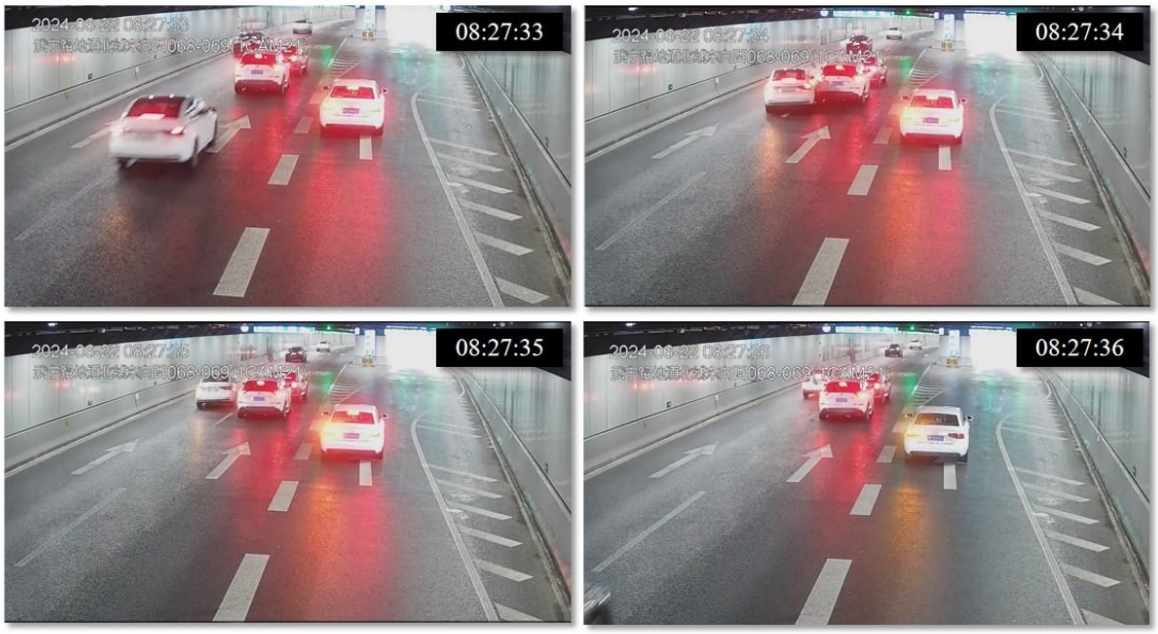
METHODS

1. Data Semantic Annotation

To enhance the visual understanding and semantic reasoning capabilities of large models for traffic anomaly events, we have developed a local traffic semantic dataset encompassing both static images and dynamic videos, supporting tasks in image comprehension and video reasoning. All image data is sourced from surveillance cameras on the expressways and highways of Shanghai, covering approximately 70,000 raw images and videos, involving more than ten event types. **Figure 1** presents local image examples, where (a) shows locally collected traffic anomaly image data, covering multiple event types such as congestion, roadwork, collisions, and illegal parking; (b) illustrates examples of uniform frame sampling from "anomalous" video keyframes, which capture the temporal evolution of dynamic traffic events.



(a) Local Collected Image Data Examples



(b) Example of Uniform Frame Sampling from 'Anomalous' Video Keyframes

FIGURE 1 Local Traffic Anomaly Image Data

Based on traffic engineering standards and urban traffic management needs, we have defined ten typical traffic semantic event types, including single-vehicle accidents, multi-collision, abnormal parking, two-wheeled vehicle intrusions, pedestrian intrusions, dispersed objects, overweight vehicles, intercity passenger vehicle supervision, road construction, and hazardous chemical vehicle supervision. For these

ten event types, we have constructed both image and video datasets and used them for model training, aiming to enhance the model's ability to recognize traffic events and develop deep semantic understanding.

To enable the model to better understand image content and produce outputs that align more closely with the needs of traffic managers, we manually annotated the traffic image data. The semantic annotation process, as shown in **Figure 2**, includes key stages such as data preprocessing, initial model annotation, manual verification, and semantic enhancement.

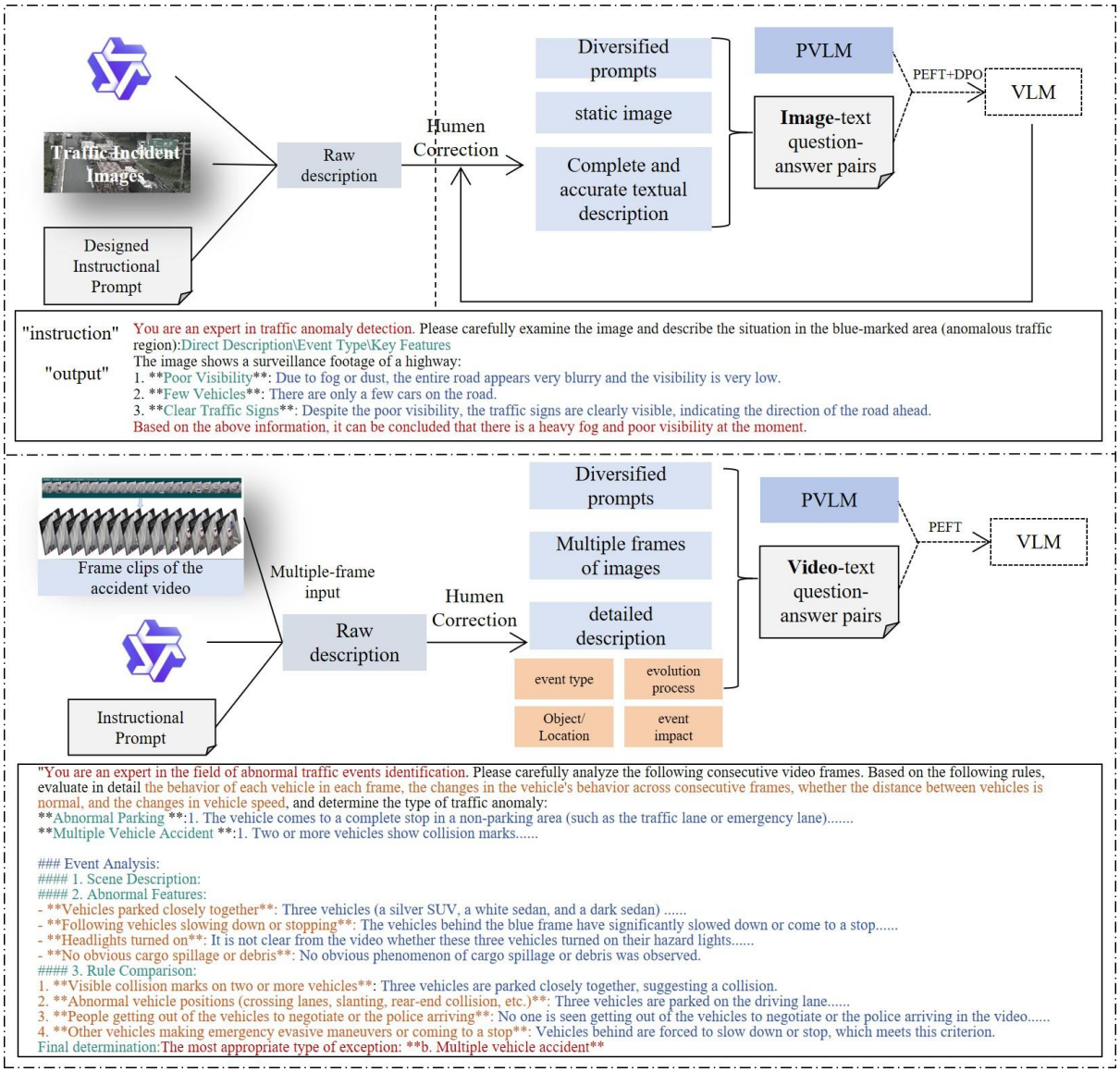


FIGURE 2 Semantic dataset construction pipeline for static images and dynamic video data in traffic anomaly detection

Specifically, during the image database construction phase, we extracted keyframe images and performed preprocessing, such as size normalization. We then designed a variety of prompt templates and used the Qwen series of visual-language model APIs for initial semantic annotation of the image content, covering elements such as event type, involved objects, event location, and impact range. After the model

generated the annotations, they were manually verified and corrected to ensure accuracy and completeness. The final dataset resulted in a structured, semantically rich image semantic dataset that covers ten types of events. To enhance the model's ability to adapt to natural language prompts, multiple sets of prompts with different expression styles but consistent semantics were assigned to image-text pairs in the dataset, strengthening the model's generalization ability.

To overcome the limitations of temporal understanding in static images for anomaly event recognition, we further constructed a video semantic dataset based on continuous frames. This dataset focuses on two high-frequency events, "multi-collision" and "abnormal parking." Several consecutive frames were extracted from surveillance videos, and event evolution information was incorporated in the temporal dimension. The data annotation process is similar to that of the image phase, relying on visual-language models for initial annotation, followed by manual correction. Additionally, semantic information , such as inter-frame actions, trajectory relations, and start-end times, was included, achieving a fusion of image, text, and time modalities. This significantly enhances the model's understanding and reasoning abilities regarding the development of traffic events in dynamic scenarios.

Ultimately, we obtained an image dataset consisting of over 876 image-instruction-output pairs, covering ten event types, and a video dataset containing 471 multi-frame image-instruction-output pairs focused on accidents and abnormal parking events. The distribution of event types in the datasets is shown in **Figure 3**.

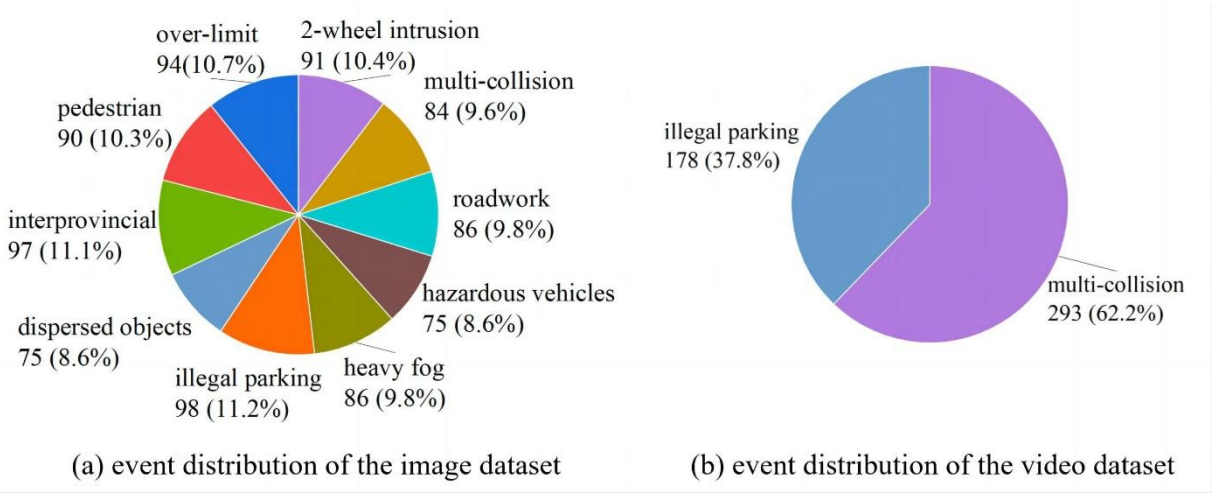


FIGURE 3 Comparison of Event Distribution in Image and Video Datasets

2. Two-Stage Optimization Framework Based on LoRA

To enhance the model's ability to recognize and describe various types of traffic anomalies, we propose a two-stage optimization framework based on LoRA, as illustrated in **Figure 4**. This framework is designed with differentiated modeling strategies targeting two distinct goals: accurate recognition of typical events and comprehensive description of critical events.

Stage 1: Image-Text Alignment for Improved Recognition of Typical Events

The first stage focuses on fast and reliable recognition of common traffic anomalies. We construct a local semantic corpus based on paired image-text data and perform SFT with LoRA to enhance the model's perception and identification of typical events, such as roadwork, pedestrian intrusion, and congestion. To further improve the structural consistency and stylistic coherence of the model's output, we introduce a locally annotated preference dataset and apply DPO. This enables the model to generate concise yet informative descriptions that align with the expectations and terminology commonly used by traffic management personnel.

Stage 2: Multi-Frame Video Semantic Modeling for Fine-Grained Description of Critical Events

In real-world traffic management scenarios, events such as "accidents" and "illegal parking" pose greater challenges in both recognition and semantic completeness. To address this, we introduce video-text multimodal datasets and develop a temporal reasoning module capable of modeling cross-frame semantic dependencies. The model is trained to incorporate key elements such as trajectory interactions, temporal boundaries, and evolving behaviors, using multiple fine-tuning strategies. The objective of this stage shifts from basic classification accuracy to achieving comprehensive, coherent, and detail-rich event descriptions, in line with the operational needs for structured incident reporting and decision-making.

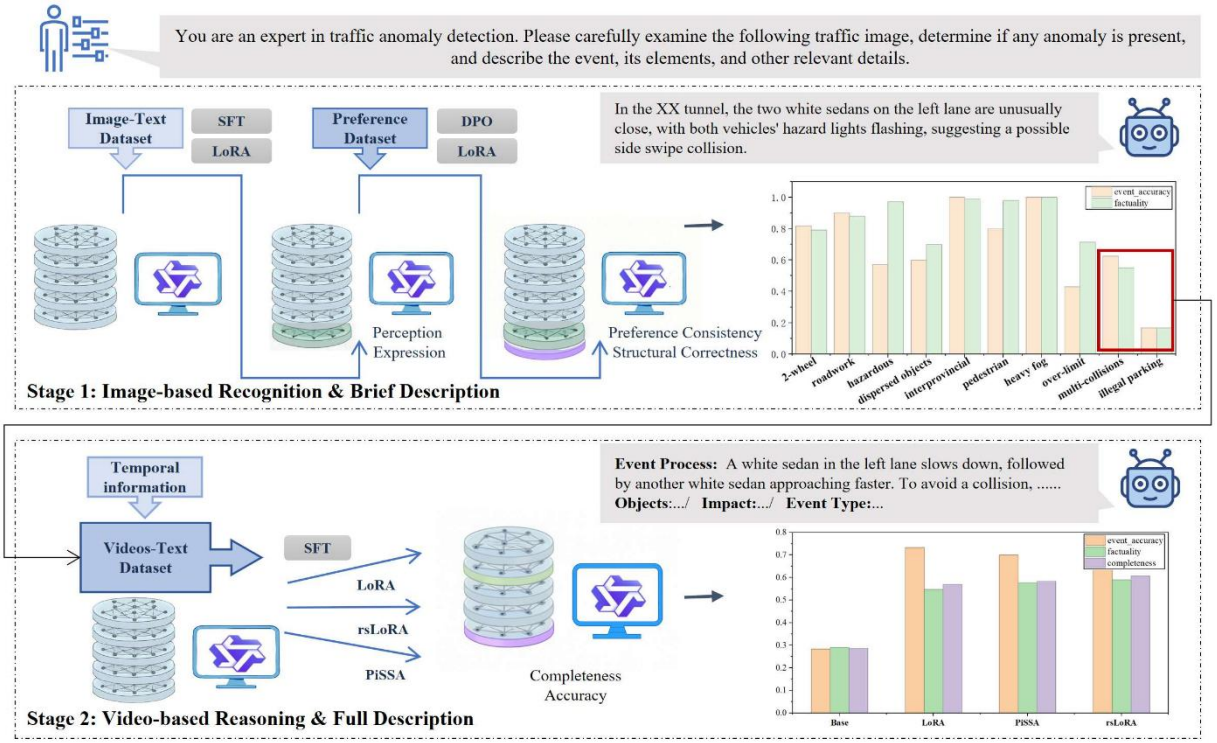


FIGURE 4 Two-stage optimization framework based on LoRA

Each stage contributes to distinct yet complementary modeling goals. The effectiveness of this design will be thoroughly evaluated in Section: **Experiments and Results**.

3. Evaluation Metrics

As Vision-Language Models (VLMs) are increasingly applied to multimodal tasks, evaluation frameworks for model outputs have evolved from early-stage language similarity metrics to more comprehensive approaches that integrate semantic understanding and visual-text alignment. In recent years, image-text consistency metrics such as CLIPScore (*26*) and SPICE (*27*) have been widely adopted for assessing the quality of multimodal generation. In this study, considering the semantic complexity of traffic image and video tasks, we adapt and extend existing evaluation frameworks to design a multi-level evaluation system for both image understanding and video reasoning. The system is divided into two main categories: structural metrics and semantic metrics. Structural metrics—such as BERTScore (*24*)—are computed automatically by models or scripts and are suitable for evaluating basic semantic similarity and element extraction capabilities. In contrast, semantic metrics rely on human judgment to perform interpretive evaluations, focusing on detecting hallucinations and assessing the completeness and factual accuracy of generated event descriptions.

TABLE 1 Evaluation Metrics Design for Multimodal Semantic Understanding Tasks

Task Stage	Metric Type	Metric Name	Evaluation Objective	Scoring Scale
Image Understanding	Structured Evaluation	BERTScore	Semantic similarity between generated text and reference	0–1
		Precision	Accuracy of keyword recognition	Percentage
		Recall	Coverage of relevant keywords	Percentage
		F1-score	Overall keyword extraction performance	Percentage
	Semantic Evaluation	Factuality	Consistency of model output with image/video content regarding event type, location, and quantity; indirectly reflects hallucination degree	0–1
		Completeness	Coverage of key semantic elements; assesses whether core objects or states are omitted	0–1
		redundency	Quantitative assessment of the occurrence of invalid information generation	0-1
		Event Accuracy	Accuracy of event type classification in images	0/1
Video reasoning	Semantic Evaluation	Factuality	Consistency of model output with image/video content regarding event type, location, and quantity; indirectly reflects hallucination degree	0–1
		Completeness	Coverage of key semantic elements; assesses whether core objects or states are omitted	0–1
		Event Accuracy	Accuracy of event type classification in images	0/1

Given the differing task objectives of image understanding and video reasoning, we adopt differentiated evaluation strategies for each stage. **Table 1** above summarizes the evaluation framework adopted in this study:

(1) In the image understanding stage, the model is required to identify anomalous events and their key elements—such as involved objects, spatial locations, and impact scope—within a static image. This aligns with the structured interpretation needs of traffic administrators. Therefore, both structural and semantic metrics are applied at this stage, balancing basic element recognition with factual accuracy.

(2) In the video reasoning stage, the focus shifts to assessing the model's comprehension of the temporal evolution of dynamic events and the coherence of its descriptive outputs. Accordingly,

only semantic metrics are used, emphasizing the evaluation of output completeness, accuracy, and logical consistency.

EXPERIMENTS AND RESULTS

1. Overview of Experimental Settings

To comprehensively evaluate the adaptability of large models in traffic semantic tasks, the experimental section is divided into two stages: image understanding and video reasoning, corresponding to event recognition in static scenes and dynamic inference across consecutive frames, respectively. Both stages are based on the pre-trained Qwen2.5-VL-7B model but adopt different training strategies and evaluation methods. For the image understanding task, we apply LoRA for parameter-efficient fine-tuning, combined with DPO to achieve human preference alignment. For the video reasoning task, we focus on two core event types—multi-collision and illegal parking—and explore the performance differences among various LoRA variants.

Due to the differing computational requirements of the two tasks, the experiments for each stage were conducted on separate hardware configurations. To ensure comparability, all model comparisons were performed under identical hardware and task settings. The evaluation metrics were designed with a consistent structural framework across both stages, as detailed in **Table 1** in the Methods section.

2. Image Understanding

1. Model Configuration and Training Procedure

To enhance the model's adaptability to the specific task, this section applies Parameter-Efficient Fine-Tuning (PEFT) (*28*) to the Qwen2.5-VL-7B model using the annotated scenario dataset described above, with the fine-tuning implemented specifically through the LoRA method. During training, the initial learning rate is set to 3e-5, while a lower learning rate of 1e-6 is applied to the embedding layers. A cosine annealing strategy is adopted to dynamically adjust the learning rate, aiming to accelerate convergence and improve training stability. The entire training process is carried out using the LLaMA-Factory open-source framework and executed on a system equipped with an Intel(R) Xeon(R) Platinum 8352V CPU (2.10GHz), 1008 GB of RAM, and a single NVIDIA vGPU with 32 GB of memory. To balance model convergence with overfitting risk, the batch size is set to 2, and the total number of training epochs is set to 10.

To ensure that the model outputs are more concise and aligned with the needs of traffic administrators, we conducted DPO training on top of the supervised fine-tuning stage. The hardware environment used for this experiment is consistent with that of the SFT stage. Since DPO is based on gradient-based optimization, this training process also employs the LoRA method for parameter-efficient fine-tuning. The low-rank dimension is set to r=8r = 8, consistent with the SFT stage, with a scaling factor α=16\alpha = 16 and a dropout rate of 0.05 to balance the influence of the adapters and the regularization effect. During training, the initial learning rate is set to 1e-5 with cosine annealing scheduling, the per-device batch size is 1, and the effective batch size is 8. To accelerate computation, the training process adopts mixed-precision training (fp16) and gradient clipping (max_grad_norm = 1). Additionally, the visual encoder is frozen (freeze_vision_tower = 1) to reduce computational overhead.

2. Incremental Ablation Analysis of Training Enhancement Strategies on Model Performance

Figure 5 presents a comparison of multiple performance metrics across different training stages. The results show that the model exhibits significant improvements on all metrics after supervised fine-tuning compared to the original pre-trained model. Specifically, the BERTScore increases to 0.7843, representing an improvement of approximately 21.6%, indicating a substantial enhancement in semantic similarity between the model outputs and the reference texts. At the keyword level, the precision improves to 0.5335, recall reaches 0.4938, and the F1 score increases to 0.4798, reflecting a systematic improvement in the model's ability to extract core semantic elements. Meanwhile, event recognition accuracy rises from 0.5139 to 0.7639, confirming a notable enhancement in the model’s understanding of traffic events.

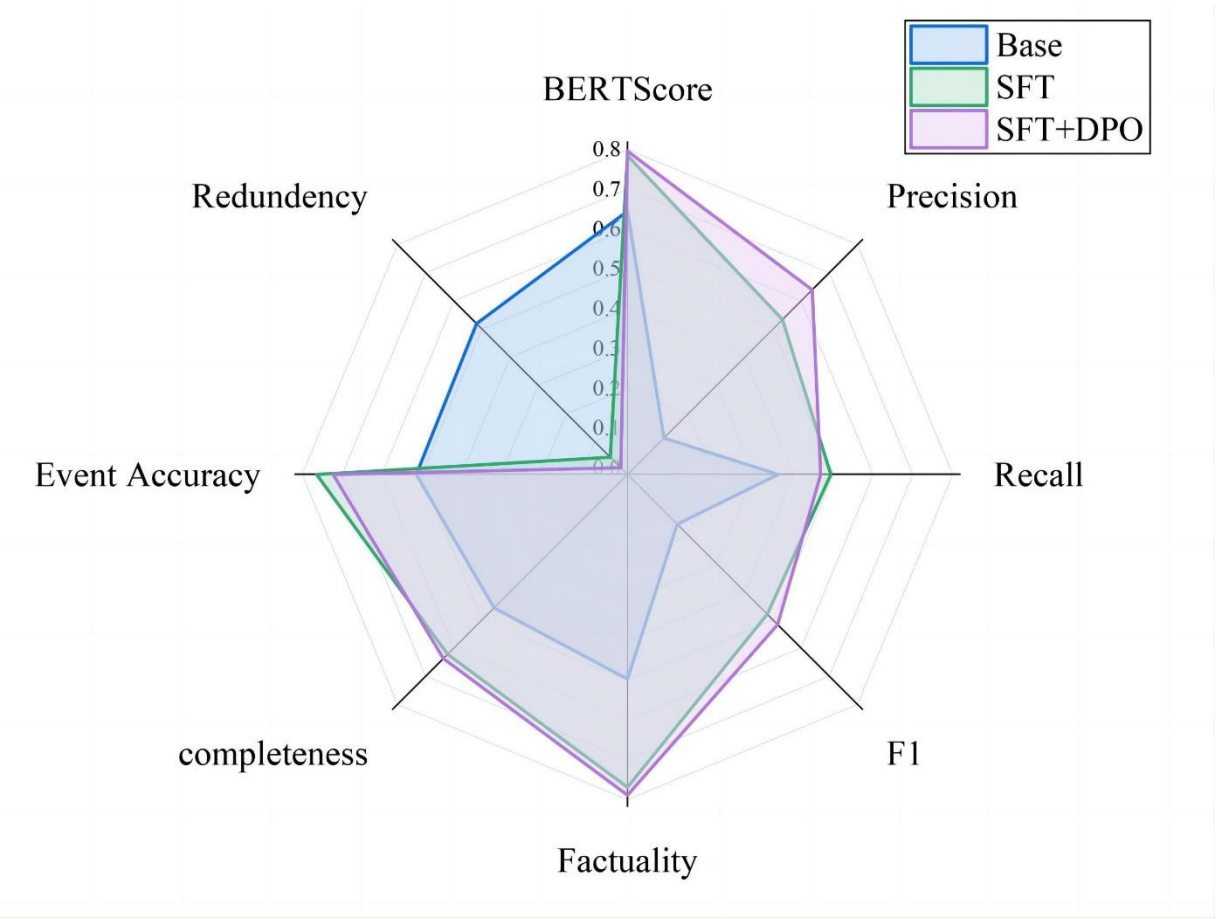


FIGURE 5 Comparative performance of the base model, SFT, and SFT+DPO across eight evaluation dimensions.

Building upon this, we introduced DPO training to further optimize the model's generation preferences. After DPO, the model's BERTScore reaches 0.7957, keyword F1 increases to 0.5165, and although event accuracy slightly decreases to 0.7222, human-evaluated factuality improves to 0.7889 and completeness rises to 0.6375, demonstrating advantages in overall text quality and consistency with visual information. Notably, the redundancy score drops significantly from 0.0444 to 0.0069, highlighting the effectiveness of DPO in reducing unnecessary repetition and compressing generated content.

3. Analysis of Recognition Performance Across Fine-Grained Traffic Event Types

Figure 6 illustrates the recognition and generation performance of the model—after supervised fine-tuning and preference optimization—across different event types. The following analysis focuses on several key scenarios related to the identification of anomalous events:

In the multi-collision scenario, the model achieves relatively strong performance on certain metrics, with a BERTScore of 0.736 and an event accuracy of 0.625. However, the keyword recall is notably low at 0.232, and the completeness score drops to just 0.028. These results indicate that the model struggles to capture key information related to multi-collision—such as the accident context, involved vehicles, and environmental factors—which leads to generated texts lacking sufficient detail. The poor completeness suggests that the model's outputs fail to comprehensively describe the accident scene, often omitting critical elements such as the cause of the incident, its impact, and surrounding environmental conditions.

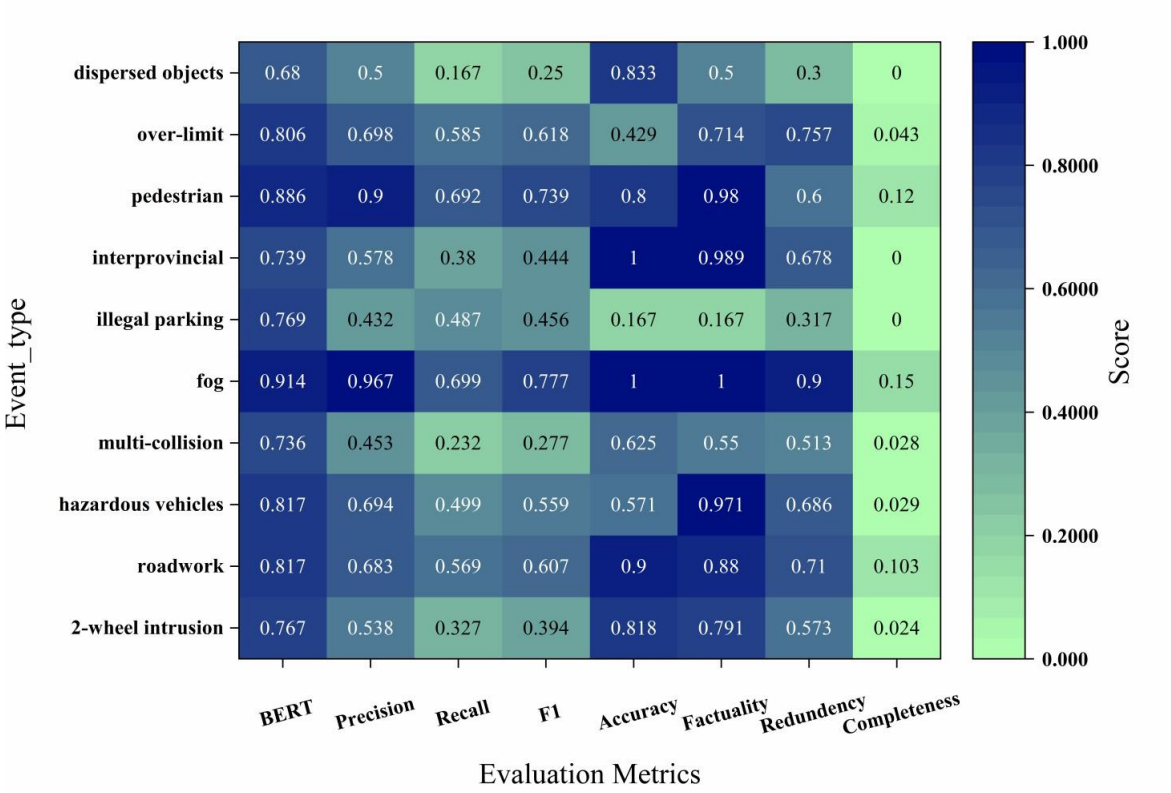


FIGURE 6 Heatmap of Structured and Semantic Evaluation Scores for Traffic Events

The model performs weakest in the illegal parking scenario, with both event accuracy and factuality scoring only 0.167, indicating a substantial deviation in the model's understanding of this event type. Moreover, the completeness score is 0, suggesting that the generated texts entirely lack detailed descriptions of the event. This poor performance can likely be attributed to the subtle or ambiguous visual cues present in static scenes—illegal parking may not exhibit clear distinctions from normal parking based on image features alone, making it difficult for the model to reliably differentiate between the two. Additionally, the low completeness score reflects the model's inability to provide informative descriptions, often resulting in superficial outputs that lack contextual or surrounding details.

In contrast, the model performs exceptionally well in identifying condition-based events such as heavy fog. The keyword precision reaches as high as 0.967, recall is 0.699, and event accuracy achieves a perfect score of 1. These results indicate that the model possesses strong capabilities in understanding and describing scenes with distinct visual features.

3. Video Reasoning

1. Comparative Design of Fine-Tuning Strategies (LoRA vs. rsLoRA / PiSSA)

To overcome the limitations of static image semantic understanding in traffic anomaly detection and further enhance the ability of vision-language models to perceive and reason about event evolution in dynamic scenes, we extended the previous static image-based work by conducting multimodal video reasoning research based on consecutive frames. Focusing on two core event types—multi-collision and illegal parking—we constructed a multimodal video semantic dataset and integrated the Qwen2.5-VL vision-language model to establish a closed-loop framework for semantic understanding and event response generation in dynamic traffic scenarios.

To better adapt the model for anomaly detection in traffic scenarios, we fine-tuned it using LoRA

and its variants, and compared their training efficiency. During training, the initial learning rate was uniformly set to 5e-5, with a cosine annealing schedule applied to dynamically adjust the learning rate. The entire training process was conducted on a system equipped with an AMD EPYC 9K84 96-Core Processor, 1239 GB of RAM, and a single NVIDIA H20 GPU with 96 GB of memory. To balance convergence and overfitting, the batch size was set to 1, and the total number of training epochs was set to 1.5.

The entire training process was conducted on a video-text dataset constructed based on the diagram, containing a total of 680 samples. The dataset was split into training, validation, and test sets in a ratio of 8:1:1.

2. Performance Comparison and Multi-Event Analysis

To evaluate the performance differences among fine-tuning strategies for traffic video anomaly detection, we compared three representative methods: LoRA (*18*), rsLoRA (*29*), and PiSSA (*30*). The comparison focuses on multiple dimensions, including the factuality and completeness of generated content, the accuracy of event type recognition, and training efficiency. The detailed performance results of each model are presented in **Table 2**.

TABLE 2. Multi-Metric Comparison of LoRA, rsLoRA, and PiSSA in Video Event Detection

Fine-Tuning Strategies	factuality	completeness	Event Accuracy	Convergence Time	Epoch
Base	0.29	0.2866	0.283	——	——
LoRA	0.5466	0.57	**0.733**	3h40m44s	1.46
rsLoRA	**0.59**	**0.6066**	0.633	4h13m19s	1.49
PiSSA	0.5766	0.5834	0.7	3h50m33s	1.26

The results show that all three fine-tuning strategies significantly outperform the original model. Among them, rsLoRA achieves the best performance on the two subjective metrics—factuality (0.59) and completeness (0.6066)—indicating that its rank-stabilization mechanism contributes to higher informational consistency and expression completeness. However, it also incurs the longest training time, approximately 4 hours and 13 minutes. LoRA attains the highest event recognition accuracy (0.733) and converges more quickly, demonstrating strong generalization ability in low-rank parameter tuning. PiSSA performs comparably to LoRA in overall metrics, with slightly longer training time but the lowest number of epochs (1.26), suggesting an advantage in convergence efficiency and potential suitability for rapid deployment scenarios.

（a） illegal parking　　（b） multi-collision　　（c） illegal parking

FIGURE 7 Visual Illustrations of Model Anomaly Detection Issues under rsLoRA Training Strategy

TABLE 3. Model Anomaly Detection Analysis under rsLoRA Training Strategy

Visual Examples	expect_output	rsLoRA_output	analysis
Figure 7 (a)	Rule Comparison: Abnormal Parking Criteria: Truck is stationary in the emergency lane — Met. No signals permit parking in the emergency lane — Met. Hazard lights are on — Met. Driver's status is unclear — Uncertain. Vehicles behind are not affected — Not met. Final Judgment: The most fitting abnormal type: Abnormal Parking	Rule Comparison: Abnormal Parking Criteria: Vehicle is stationary in a non-parking area (emergency lane) — Met. No signals permit parking — Met. Hazard lights are on — Not met. Driver's presence is unconfirmed, and vehicle status is unclear — Met. Final Judgment: Most fitting abnormal type: Abnormal Parking	√ Incident Type √ Incident Location √ Driver Status × Hazard Lights Status
Figure 7 (b)	Incident Analysis: Location: Wu Ning Road North Tunnel, east to west direction. Involved Vehicles: White sedan, parked in the far-right lane with hazard lights on.White SUV, parked in the middle lane with hazard lights on.Red sedan, parked in the middle lane with hazard lights on. Other vehicles (including a white and black sedan) parked in the left lane with hazard lights on. Behavior: Multiple vehicles are parked with hazard lights on, indicating a traffic accident or emergency.	Incident Analysis: Abnormal Features: Multiple vehicles are stopped or moving slowly in the lanes, indicating a possible collision. Vehicle positions are abnormal, with vehicles crossing multiple lanes, suggesting a collision has occurred. Some individuals have exited their vehicles, possibly handling the accident or checking the situation.	The event description is not clear enough and lacks an understanding of dynamic time series
Figure 7 (c)	Abnormal Features: Rightmost lane (the emergency lane) has two large trucks parked on the roadside, both stationary. The front truck has its hazard lights on, and people have exited the vehicle to inspect the situation. The second truck has no hazard lights on, and there are no people nearby. Rule Comparison: Based on the provided abnormal parking judgment criteria: Vehicle is completely stationary in a non-parking area (e.g., driving lanes, emergency lanes): The truck is parked in the emergency lane, which is a non-parking area.	Scene Description: Vehicles are moving slowly, with some using hazard lights, possibly due to rainy weather or wet roads. The distance between vehicles is normal, with no severe congestion or stalling. Vehicle speeds remain low, and most vehicles are concentrated in the left lanes. No vehicles are stationary in the emergency lane or non-driving lanes, and no signs of a multiple vehicle accident are observed.	Rainy weather / target distance is too far, visual features are not obvious, and the model is unable to detect the anomaly

We further analyzed the model's performance under specific event types, using the output results from the rsLoRA training, which demonstrated the best performance in terms of content factuality, as shown in **Table 3**. Despite fine-tuning, the model still exhibits the following issues:

(1) Inaccurate recognition of hazard light status in illegal parking events: Although the model can correctly identify the event type, it often exhibits significant errors in determining vehicle position and type, and even generates severe hallucinations. This indicates a lack of stable semantic control at the detail level.

(2) Insufficient perception of event progression in multi-collision: Most models fail to describe the causal chain of accident development, relying instead on surface-level cues such as vehicle clustering and changes in speed. This limitation is likely related to the frame sampling strategy during training. In particular, the use of evenly spaced frames for training data may result in the omission of critical moments, thereby impairing the model's ability to understand and reason over dynamic event sequences.

(3) Performance degradation under weak visual signals: In low-visibility conditions—such as poor lighting or inconspicuous vehicle positioning—the accuracy and completeness of generated content drop significantly. This suggests that current models still lack sufficient robustness when dealing with visually complex environments.

Overall, although various fine-tuning strategies significantly improve the performance of large models in video-based event recognition, challenges remain in detailed characterization of multiple events, causal reasoning, and adaptation to low-quality visual scenarios. Future research may focus on enhancing keyframe selection mechanisms, low-light image enhancement, and multi-scale semantic modeling to further improve the model’s practicality and robustness in real-world traffic environments.

CONCLUSIONS

This study presents a two-stage optimization framework that combines static image semantic analysis with dynamic video reasoning for traffic anomaly detection. By constructing a localized image-text dataset and employing a multi-stage training strategy, including SFT and DPO, we significantly enhance the performance of visual language models in complex traffic environments. Experimental results demonstrate that, after two-stage optimization, the Qwen2.5-VL-7B model's recognition accuracy in static image understanding tasks improved from 0.497 to 0.789, with a significant reduction in text redundancy. In dynamic video reasoning tasks, for events such as "multi-collision" and "illegal parking," the model's recognition accuracy increased from 0.358 to 0.59, achieving a 64.8% improvement. These results validate that the proposed framework, combined with domain-specific datasets, effectively enhances the model's semantic understanding and event detection capabilities in traffic-related applications.

This approach not only demonstrates the immense potential of visual language models in the intelligent transportation field but also provides an efficient and cost-effective solution for multimodal anomaly detection. Future research will focus on exploring key frame selection mechanisms, low-light image enhancement techniques, and more refined causal modeling of events, to further improve the model's generalization ability and robustness in complex, multi-source environments.

ACKNOWLEDGMENTS

This work was funded by research grants from the National Natural Science Foundation of China(52472327, 52372305), the Belt and Road Cooperation Program under the 2023 Shanghai Action Plan for Science, Technology and Innovation(23210750500), and Science and Technology Commission of Shanghai Municipality(23QB1404900).

REFERENCES

1. Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 53728–53741.

2. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., et al. 2022. LoRA: Low-rank adaptation of large language models. Presented at International Conference on Learning Representations (ICLR).

3. Horn, B. K., and Schunck, B. G. 1981. Determining optical flow. *Artificial Intelligence*.

4. Farnebäck, G. 2003. Two-frame motion estimation based on polynomial expansion. Presented at Scandinavian Conference on Image Analysis, Berlin, Heidelberg. Springer, pp. 363–370.

5. Stauffer, C., and Grimson, W. E. L. 1999. Adaptive background mixture models for real-time tracking. Presented at IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 2, pp. 246–252. IEEE.

6. Elgammal, A., Duraiswami, R., Harwood, D., and Davis, L. S. 2002. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. *Proceedings of the IEEE*, 90(7), 1151–1163.

7. Li, H., and Chen, L. 2025. Traffic accident risk prediction based on deep learning and spatiotemporal features of vehicle trajectories. *PLoS One*, 20(5), e0320656.

8. Ijjina, E. P., Chand, D., Gupta, S., and Goutham, K. 2019. Computer vision-based accident detection in traffic surveillance. Presented at 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), IEEE, pp. 1–6.

9. Ma, X., Dai, Z., He, Z., Ma, J., Wang, Y., and Wang, Y. 2017. Learning traffic as images: A deep convolutional neural network for large-scale transportation network speed prediction. *Sensors*, 17(4), 818.

10. Cui, Z., Ke, R., Pu, Z., and Wang, Y. 2018. Deep bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction. *arXiv preprint arXiv:1801.02143.*

11. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... and Sutskever, I. 2021. Learning transferable visual models from natural language supervision. Presented at International Conference on Machine Learning, PMLR, pp. 8748–8763.

12. Li, J., Li, D., Xiong, C., and Hoi, S. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. Presented at International Conference on Machine Learning, PMLR, pp. 12888–12900.

13. Li, J., Li, D., Savarese, S., and Hoi, S. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. Presented at International Conference on Machine Learning, PMLR, pp. 19730–19742.

14. Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., et al. 2022. Flamingo: A visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35, 23716–23736.

15. Da, L., Gao, M., Mei, H., and Wei, H. 2024. Prompt to transfer: Sim-to-real transfer for traffic signal control with prompt learning. Presented at AAAI Conference on Artificial Intelligence, Vol. 38, No. 1, pp. 82–90.

16. Zhang, S., Fu, D., Liang, W., Zhang, Z., Yu, B., Cai, P., and Yao, B. 2024. TrafficGPT: Viewing, processing and interacting with traffic foundation models. *Transport Policy*, 150, 95–105.
17. Liu, C., Yang, S., Xu, Q., Li, Z., Long, C., Li, Z., and Zhao, R. 2024. Spatial-temporal large language model for traffic prediction. Presented at 25th IEEE International Conference on Mobile Data Management (MDM), IEEE, pp. 31–40.
18. Zheng, O., Abdel-Aty, M., Wang, D., Wang, C., and Ding, S. 2023. TrafficSafetyGPT: Tuning a pre-trained large language model to a domain-specific expert in transportation safety. *arXiv preprint arXiv:2307.15311.*
19. Wang, L., Ren, Y., Jiang, H., Cai, P., Fu, D., Wang, T., et al. 2023. AccidentGPT: Accident analysis and prevention from V2X environmental perception with multi-modal large model. *arXiv preprint arXiv:2312.13156.*
20. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. 2017. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347.*
21. Lee, H., Phatale, S., Mansoor, H., Mesnard, T., Ferret, J., Lu, K., et al. 2023. RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. *arXiv preprint arXiv:2309.00267.*
22. Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., and Saenko, K. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156.*
23. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and Wen, J. R. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355.*
24. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. 2019. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675.*
25. Vedantam, R., Zitnick, C. L., and Parikh, D. 2015. CIDEr: Consensus-based image description evaluation. Presented at IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4566–4575.
26. Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., and Choi, Y. 2021. CLIPScore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718.*
27. Anderson, P., Fernando, B., Johnson, M., and Gould, S. 2016. SPICE: Semantic propositional image caption evaluation. Presented at European Conference on Computer Vision (ECCV), Springer, pp. 382–398.
28. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., et al. 2019. Parameter-efficient transfer learning for NLP. Presented at International Conference on Machine Learning (ICML), *PMLR*, pp. 2790–2799.
29. Kalajdzievski, D. 2023. A rank stabilization scaling factor for fine-tuning with LoRA. *arXiv preprint arXiv:2312.03732.*
30. Meng, F., Wang, Z., and Zhang, M. 2024. PiSSA: Principal singular values and singular vectors adaptation of large language models. *Advances in Neural Information Processing Systems*, 37, 121038–121072.