Consensus Generation of Multi-Vehicle Accident Reports Based on Vision-Language Models and Graph Neural Networks

Yilin Yu[1*]

[1]School of Computer and Information Engineering, Henan University of Economics and Law, Zhengzhou, 450046, China

Abstract. As road traffic accidents are high-frequency incidents. Although there are sufficient and clear traffic control cameras at some perception-dense sites, in other places there is often a lack of enough cameras. In order to make full use of limited information to efficiently obtain an overview of the accident scene in a short time, assist on-site handling and investigation, and help reduce secondary casualties and restore traffic. This paper makes full use of existing technologies and builds a framework that can fully utilize the image and text information of the accident scene. First, the images uploaded by vehicles are used for VLM to generate objective descriptions, and the descriptions from multiple sensors within a single vehicle are fused. Then the network architecture is designed to carry out credibility calculation and severity classification. Finally, the summary is selected to form the final accident consensus report. Through the above work, semantic description and severity assessment of the event can be quickly achieved, while false reports are minimized to the greatest extent.

1 Introduction

To address the leading cause of death among children and young people - road traffic injuries [1], in recent periods, vision-language models based on end-to-end cross-modal alignment have significantly enhanced the level of semantic understanding in complex traffic scenarios. The CIRS framework proposed by Ayesha et al. adopts VideoLLaMA3 as the description Agent, achieving BLEU 0.0755 and METEOR 0.2258 text generation performance on a self-defined CCTV dataset, and can output natural language descriptions including vehicle types and accident details. However, VideoLLaMA3 may produce hallucination when processing static images [2].

In comparison, Qwen2-VL based on multimodal rotary position encoding (RoPE) and dynamic resolution processing of image patches achieves an accuracy of 96.5% in document-level visual question answering, demonstrating superior fine-grained visual understanding capability [3]. In the field of autonomous driving, the survey by Cui et al. shows that VLMs such as LLaVA and Qwen2-VL have been widely used for scene description, risk warning, and decision explanation, but current studies mostly focus on static image analysis from a single-vehicle perspective, lacking modeling of temporal coherence under multi-vehicle dynamic perspectives [4].

Multi-agent systems address real-time decision-making problems in complex traffic scenarios based on distributed collaboration. Mushtaq et al. propose a traffic signal optimization scheme based on multi-agent reinforcement learning, improving traffic efficiency through state sharing among agents. Kumar et al. establish an IoT-MAS hybrid system to realize cooperation between vehicles and roadside units based on wireless sensor networks. However, current MAS research mainly focuses on traffic flow optimization and signal control, and in accident detection scenarios, agents only perform simple information broadcasting, lacking quantitative evaluation of the credibility of multi-vehicle reports and methods for consistency verification [5].

The RoBERTa-large-mnli model proposed by Liu et al. achieves an accuracy of 90.2% on the MultiNLI benchmark and has been widely used in text semantic matching and factual consistency detection [6]. In the field of autonomous driving, NLI is mainly used to verify the consistency between perception model outputs and map data, but its application to cross-verification of multi-vehicle multi-perspective textual descriptions has not yet been systematically studied. Although the CIRS framework generates descriptions based on VLM, it does not conduct logical consistency checks on multi-source descriptions, which may allow potential erroneous information to directly enter the decision-making process.

Graph neural networks significantly improve the accuracy of group behavior prediction and risk warning by modeling complex interaction relationships among traffic entities. Sánchez-González et al. first applied graph networks to the simulation of physical scenarios, demonstrating their advantages in relational reasoning. In the field of transportation, graph attention networks can be used to model dynamic interactions among vehicles and achieve collision risk prediction [7].

* Corresponding author: Lynn353440662@outlook.com

Based on the above analysis, current studies have the following deficiencies: the limitation of single-vehicle perception: systems such as CIRS rely on fixed roadside CCTV and lack the complementary integration of vehicle-mounted first-person perspectives and roadside top-view perspectives, which may easily lead to detection blind spots under occlusion or low-light conditions. The contradiction at the semantic level has not been resolved, as current MAS frameworks only realize information transmission and do not establish a consistency verification mechanism for multi-source textual descriptions, making it difficult to handle conflicting statements such as the rear vehicle claiming sudden braking of the front vehicle and the front vehicle claiming being rear-ended. In addition, the consensus in decision-making is relatively insufficient, as current methods form decisions based on single-point perception, lacking a multi-vehicle collaborative global consensus mechanism, and making it difficult to evaluate the reliability of different information sources and generate a unified accident report.

In response to the above limitations, this study proposes a consensus report scheme integrating VLM, NLI, and heterogeneous GNN, achieving end-to-end generation from multi-source heterogeneous perception to global consistent consensus based on in-vehicle NLI verification and cross-vehicle GNN fusion.

2 Method design

In response to the limitations of existing systems in perspective fusion and consensus generation, this section proposes a multi-vehicle accident consensus reporting framework based on heterogeneous graph neural networks (Multi-Vehicle Accident Consensus Reporting Framework, MACRF). The framework contains four core modules: VLM-based Perception, Intra-Vehicle Fusion, Inter-Vehicle Consensus, and Report Generation.

Different from the five-agent centralized architecture of CIRS, MACRF adopts a hierarchical design of "edge perception - local verification - global consensus". First, natural language inference (NLI) consistency verification of multi-sensor descriptions within a single vehicle is completed at each vehicle edge node, then explicit modeling of multi-vehicle relationships and credibility weighting are realized through heterogeneous graph neural networks (GNN) at the cloud/edge server, and finally a structured consensus report is generated.

2.1 Multi-Source visual perception and description generation

2.1.1 Heterogeneous sensor configuration

The accident scenario processed by MACRF contains three types of heterogeneous perception sources, forming a complement to the scheme that CIRS only relies on roadside closed-circuit television (CCTV). The vehicle-mounted front-view camera (c_i,front) is responsible for recording the road scene in front of the vehicle, the vehicle-mounted rear-view camera (c_i,rear) is responsible for recording the rear and rear-side blind areas of the vehicle, and the roadside top-view monitoring (c_road) is fixedly deployed above intersections or road sections to provide a global top-view perspective.

This configuration corresponds to the multi-view setting in the DeepAccident dataset, where each vehicle is equipped with five sensors (front/rear/left/right/surround view), and this study selects the front/rear cameras and the nearest roadside monitoring to form the minimum heterogeneous observation set.

2.1.2 VLM description generation

The single-vehicle perception module serves as the baseline comparison of the system and is responsible for processing image data from a single roadside perspective (Camera_Front) to generate structured accident reports. This module is built based on the Qwen2-VL-7B-Instruct vision-language large model and adopts a local deployment method.

The input data is a single vehicle-mounted camera image (JPG/PNG format), which is sent into the VLM model after image preprocessing. The model output is strictly constrained to JSON format and contains three key fields. The accident description uses natural language text to describe the accident type and on-site situation in detail; the severity is divided into three levels: minor, moderate, and severe; the confidence is a value between 0 and 1, reflecting the certainty of the model's judgment.

2.2 Intra-Vehicle fusion module

2.2.1 NLI consistency detection

The RoBERTa-large-MNLI model is used to calculate the semantic entailment relationship between sensor descriptions. For two sensor descriptions d_i,m and d_i,n of vehicle i, the NLI model outputs the probabilities of three relations, namely entailment, contradiction, and neutral. The consistency score is calculated as follows:

$$\text{Consistency\_Score} = \text{P(Entailment)} - \text{P(Contradiction)} \quad (1)$$

2.2.2 Weighted fusion strategy

Based on the NLI consistency score, a weighted fusion strategy is adopted to generate vehicle-level descriptions. First, a consistency threshold (such as 0.3) is set to filter low-credibility descriptions. Then, the descriptions that pass the verification are fused by weighted averaging according to confidence. Finally, majority voting is performed on the severity judgments of multiple sensors.

2.3 Inter-Vehicle consensus reasoning module

The inter-vehicle consensus reasoning module is the core innovation of MACRF, which realizes explicit modeling of multi-vehicle information and credibility weighting through heterogeneous graph neural networks.

2.3.1 Graph structure construction

The multi-vehicle scenario is modeled as a heterogeneous graph G = (V, E, X)

For nodes (V), each vehicle is treated as a node, and the node feature x_i includes: CLIP visual features (512 dimensions), vehicle position encoding (2 dimensions: x, y coordinates), and severity one-hot encoding (3 dimensions: minor/moderate/severe), with a total feature dimension of 517. For edges (E), the logical relationship edges between vehicles include: spatial proximity edges (constructed based on Euclidean distance), perspective complementarity edges (connections between different perspective types), and full connections (to ensure sufficient information propagation). In the experiment, the graph composed of 4 vehicles contains 4 nodes and 18 edges (including self-loop edges).

2.3.2 Graph attention network

A 2-layer Graph Attention Network (GAT) is adopted for message passing [8]. The attention mechanism enables the model to automatically learn the importance weights of information among vehicles, and roadside monitoring nodes usually obtain higher attention weights due to their global perspective.

2.3.3 Severity decision logic

The severity determination adopts a fusion strategy combining rules and neural networks, in which a no-accident interceptor forcibly determines the case as minor level through regular matching of keywords such as "no collision" and "normal driving" to prevent hallucination, and the keyword scoring mechanism assigns scores to keywords. According to the total score and the number of reporting vehicles, a three-level severity classification is performed.

2.4 Report generation module

The report generation module is responsible for integrating the consensus results into a structured accident report.

According to credibility ranking, high-credibility vehicles (usually >0.5) are selected as consensus members, and low-credibility outlier reports are removed. Based on the selected consensus member descriptions, a concise summary is generated: event type identification (determining the accident type based on keyword matching), timeline reconstruction (integrating temporal information from multi-vehicle descriptions), and key information extraction (extracting key elements such as the number of involved vehicles, damage degree, and road impact).

The final output is a structured JSON report, including: accident summary (description), severity level (severity), confidence score (confidence), consensus vehicle list (consensus_vehicles), and consistency score (consistency_score).

3 Experimental setup

3.1 Experimental environment and configuration

3.1.1 Hardware environment & software environment

The experiment is conducted on a computer equipped with an NVIDIA RTX 5880 Ada 48GB graphics card, an Intel Core i9-13900K processor, and 32GB DDR5 memory, with the operating system being Windows 10 Professional.

The software environment adopts Python 3.12 as the development language, the deep learning framework uses PyTorch 2.0.1 (CUDA 12.1), natural language processing uses Transformers 4.35.0, and graph neural networks use PyTorch Geometric 2.4.0.

3.1.2 Model configuration

The visual encoder adopts Qwen2.5-VL-7B-Instruct with FP16 half-precision local deployment; the text encoder adopts RoBERTa-large-MNLI (HuggingFace version); the GNN adopts a 2-layer Graph Attention Network, with hidden dimensions of 517→256→64→3.

3.2 Dataset

The DeepAccident dataset is the first large-scale accident dataset for autonomous driving, released by the University of Hong Kong in 2022 [9]. The dataset includes: 131k annotated LiDAR samples (3 times nuScenes), 791k annotated camera images, and six categories, supporting multi-vehicle collaborative autonomous driving. This study selects the front/rear cameras and the nearest roadside monitoring to form the minimum heterogeneous observation set.

The VeRi-776 dataset is a large-scale vehicle re-identification benchmark dataset in urban surveillance scenarios proposed by Liu et al. in 2016. The dataset is collected from real urban road surveillance cameras and contains 776 different vehicles, covering various lighting conditions in daytime and nighttime, and provides rich vehicle attribute annotations [10]. The dataset has become a widely used standard benchmark in the field of vehicle re-identification, used to verify the effectiveness of multi-view vehicle feature extraction and matching algorithms.

3.3 Comparison Methods

The single-vehicle VLM only uses single-vehicle single-view (Camera_Front) image and directly generates accident reports through Qwen2-VL, serving as the system baseline comparison. The complete MACRF framework includes: multi-vehicle VLM feature extraction, intra-vehicle NLI consistency verification, cross-vehicle GNN consensus fusion, and structured report generation.

3.4 Evaluation metrics

The severity classification metrics evaluate the accuracy of severity classification through accuracy, use MAE (Mean Absolute Error) to measure the average absolute error of severity levels, and combine F1 score (F1-Score) to calculate macro-average F1 score to comprehensively evaluate model performance. The confidence calibration dimension includes two core indicators, namely evaluating the consistency between predicted confidence and actual accuracy through confidence calibration, and quantifying the semantic consistency level among multi-vehicle reports using consistency score. The efficiency metrics focus on system real-time performance and resource consumption, specifically recording the processing time of a single scenario through average inference time.

4 Experimental results and analysis

4.1 Main experimental results

The comparison experimental results on part of the DeepAccident dataset are shown in Table 1.

Table 1. Comparison Results of Baseline and Consensus Methods.

Metrics	Baseline (Single-Vehicle VLM)	Consensus (GNN+NLI)	Relative Improvement
Severity Accuracy	0.5000	0.6000	+20.00%
Severity MAE	0.5000	0.4000	-20.00%
Severity F1 Score	0.2222	0.2500	+12.50%
Confidence Calibration	1.0000	0.8235	-17.65%
Average Inference Time	2.586s	10.205s	+294.63%
Average Consistency	-	0.6750	-

From Table 1, it can be seen that the Consensus method improves the severity classification accuracy by 20% compared with the Baseline (from 50% to 60%), reduces MAE by 20% (from 0.5 to 0.4), and increases the F1 score by 12.5%. This indicates that the multi-vehicle consensus mechanism can effectively integrate multi-view information and improve the accuracy of accident severity judgment.

In the test scenarios where conflicts in severity judgments among vehicles are detected, the Consensus method effectively resolves these conflicts through the GNN+NLI mechanism, with an average consistency score reaching 0.675. In terms of efficiency, the Consensus method requires processing multi-vehicle multi-view data, and the inference time (10.205s) is about 4 times that of the Baseline (2.586s), but it still meets real-time requirements (<15s).

4.2 Ablation study

In order to verify the independent contribution of each module, ablation experiments are conducted, and the results are shown in Table 2.

Table 2. Ablation Experiment Results.

Method	Severity Accuracy	Severity F1 Score	Average Consistency
Single-View (Baseline)	0.50	0.2222	-
Multi-View Average	0.55	0.2333	0.50
Multi-View + GNN	0.58	0.2417	0.62
Multi-View + GNN + NLI (Full)	0.60	0.2500	0.675

The ablation experiment results show that, compared with single-view, multi-view aggregation improves the accuracy by 5%, verifying the value of multi-view information. Compared with multi-view averaging, GNN fusion improves the accuracy by 3% and the consistency by 12%, verifying the effectiveness of graph neural networks in cross-vehicle information fusion. Compared with using GNN alone, NLI verification further improves the accuracy by 2% and the consistency by 5.5%, verifying the supplementary role of natural language inference in consistency constraints.

4.3 VeRi-776 vehicle Re-Identification verification experiment

In order to verify the effectiveness of multi-view feature extraction, supplementary experiments are conducted on the VeRi-776 vehicle re-identification dataset, and the results are shown in Table 3.

Table 3. VeRi-776 Vehicle Re-Identification Experimental Results.

Method	Rank-1	Rank-5	mAP
Single-View	0.72	1.00	0.7457

Method	Rank-1	Rank-5	mAP
Multi-View	1.00	1.00	1.00
Relative Improvement	+38.89%	-	+34.11%

The experimental results show that multi-view aggregation improves the Rank-1 accuracy from 72% to 100%, verifying the effectiveness of multi-view visual feature fusion. Achieving saturated performance on VeRi-776 indicates that the dataset is relatively simple, and larger or more challenging datasets (such as VehicleID or VERI-Wild) are required to further verify the generalization ability of the method.

4.4 Discussion

The MACRF system significantly improves the accuracy of accident severity classification through a multi-vehicle collaborative mechanism. The experiments show that the Consensus method improves the accuracy by 20% compared with the single-vehicle Baseline and realizes conflict detection and resolution in test scenarios, the visualization of attention weights can show the information flow among vehicles, and the NLI output can explain conflicts. At the same time, multi-view feature fusion improves the Rank-1 recognition rate from 72% to 100% on the VeRi-776 dataset; in addition, the system has strong interpretability, as the visualization of GNN attention weights can intuitively show the information flow among vehicles, the NLI output clearly explains the source of conflicts, and the confidence scoring mechanism quantifies decision uncertainty. The system adopts a fully localized deployment scheme and can run offline without relying on cloud APIs, with a single-scenario processing time of about 10 seconds, meeting the real-time requirements of edge computing; at the same time, a four-stage pipeline architecture (perception → fusion → reasoning → generation) is adopted to realize modular design, each module can be independently optimized, supports flexible configurations such as threshold adjustment and model replacement, and is easy to extend to more sensor types and accident scenarios. However, the system still has certain limitations, as it involves multimodal feature extraction and graph neural network reasoning, resulting in relatively high computational complexity, and the single-scenario processing time is about 4 times that of the Baseline, and the computational efficiency needs to be further optimized.

5 Conclusion

This paper proposes a multi-vehicle accident consensus reporting framework (MACRF) based on vision-language models and heterogeneous graph neural networks for the problem of multi-vehicle collaborative accident report generation in autonomous driving scenarios. By jointly using the Qwen2-VL-7B-Instruct vision-language model, the RoBERTa-large-MNLI natural language inference model, and a two-layer graph attention network, end-to-end processing is achieved from multi-source heterogeneous visual perception such as vehicle-mounted front/rear cameras and roadside monitoring to structured consensus reports, effectively solving the technical problems of sensor noise interference and report conflict resolution in multi-vehicle accident scenarios. The experimental results show that the system verifies the effectiveness of the multi-vehicle consensus mechanism on the DeepAccident dataset, improves the severity classification accuracy by 20% compared with the single-vehicle Baseline, achieves an average consistency score of 0.675, successfully detects and resolves conflicts among vehicle judgments in test scenarios, and realizes a single-scenario processing time of about 10 seconds on a single RTX 5880 GPU, meeting the real-time requirements under edge computing environments. This study constructs an open-source verification scheme that can be locally deployed, providing a technical reference for accident report generation in distributed vehicle systems. Future work will face challenges such as insufficient utilization of temporal information and the need to optimize computational efficiency, and it is planned to further improve system performance and practicality by introducing temporal graph networks and model quantization acceleration.

References

1. World Health Organization, Global status report on road safety 2023. (2024) Available: https://www.who.int/health-topics/road-safety
2. S. Ayesha, A. Aslam, M. H. Zaheer, and M. B. Khan, CIRS: A multi-agent machine learning framework for real-time accident detection and emergency response. *Sensors*, 25, 5845 (2025)
3. P. Wang, S. Bai, S. Tan, et al., Qwen2-vl: Enhancing vision-language model‘s perception of the world at any resolution. *arXiv:2409.12191* (2024)
4. C. Cui, Y. Ma, X. Cao, et al., A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 958-979 (2024)
5. A. Mushtaq, I. U. Haq, M. A. Sarwar, A. Khan, W. Khalil, and M. A. Mughal, Multi-agent reinforcement learning for traffic flow management of autonomous vehicles. *Sensors*, 23(5), 2373 (2023)
6. Y. Liu, M. Ott, N. Goyal, et al., Roberta: A robustly optimized BERT pretraining approach. *arXiv:1907.11692* (2019)
7. A. Sanchez-Gonzalez, J. Godwin, T. Pfaff, et al., Learning to simulate complex physics with graph networks. In *Proceedings of the International Conference on Machine Learning (PMLR)*, 8459-8468 (2020)

8. Z. Hu, Y. Dong, K. Wang, and Y. Sun, Heterogeneous graph transformer. In *Proceedings of The Web Conference 2020 (WWW '20)*, ACM, 2704-2710 (2020)
9. T. Wang, S. Kim, J. Wenxuan, et al., Deepaccident: A motion and accident prediction benchmark for V2X autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6), 5599-5606 (2024)
10. X. Liu, W. Liu, H. Ma, and H. Fu, Large-scale vehicle re-identification in urban surveillance videos. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, Seattle, WA, USA, 1-6 (2016)