Journal of Artificial Intelligence Practice (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2026.090115
ISSN 2371-8412 Vol. 9 Num. 1

An RSPD-ECA-Enhanced YOLO Detector and Vision-Language Collaborative Assessment Framework for Traffic Accident Video Understanding

Pinxi Zhou, Xiaoyu Yu

Beijing National Day School, Beijing, 100088, China

Keywords: Traffic accident detection; YOLO; Vision-language model; Video understanding

Abstract: Automated traffic accident video analysis requires reliable object localization and concise semantic reporting under response-time constraints. This paper proposes YOLO-RSPD-ECA, a lightweight detector enhancement for accident scenes, and connects it with a vision-language reporting workflow. RSPDConv replaces selected stride-2 convolution paths with a residual spatial-preserving downsampling design, using a pooling detail branch to reduce local information loss. ECA attention is inserted after backbone and neck feature blocks to strengthen accident-relevant channels with limited parameter overhead. YOLO-RSPD-ECA outputs boxes, labels, and key-frame evidence, which are provided to Qwen3.5-397B-A17B for structured accident reports and then reviewed by domain experts. Experiments on 33,462 accident images and 101 test videos show consistent gains over YOLOv8, YOLO11, and YOLO26 baselines. The YOLOv8 mAP50-95 improves from 0.4438 to 0.4754; all videos are processed successfully, with an average expert correctness score of 0.833 and an acceptable-report rate of 85.15%.

1. Introduction

Road traffic accidents remain a major public-safety problem. The World Health Organization reports that about 1.19 million people die each year from road traffic crashes, and road traffic injuries remain a leading cause of death for people aged 5-29 years [1]. Global Burden of Disease results also show that road-injury burdens remain large and uneven across regions [2]. Therefore, accident video understanding should support not only post-event review but also rapid scene assessment, severity judgment, and dispatch prioritization.

Urban surveillance cameras, dashboard cameras, roadside cameras, and in-vehicle sensors produce large volumes of accident video evidence. Existing datasets and accident-recognition studies show that accident scenes often include small collision contact regions, occlusion, low-resolution frames, ambiguous temporal boundaries, and diverse anomaly types [3-6]. Manual frame-by-frame review is difficult to scale under these conditions, especially when local traffic-management departments need timely and traceable evidence.

Modern object detectors, including EfficientDet, DETR, and several YOLO variants, have improved accuracy and real-time deployment in general visual scenarios [7-12]. However, traffic

accident semantics may depend on very small visual cues, such as rear-end contact points, lateral impact areas, abnormal vehicle posture, or roadside hazards. Repeated downsampling can weaken these cues before the detection head recovers them.

Vision-language models provide a complementary route for semantic interpretation. Models such as CLIP, BLIP-2, and LLaVA demonstrate strong image-text alignment and visual instruction following, while multimodal retrieval-augmented generation emphasizes the use of structured evidence to reduce unsupported generation [13-16]. Used alone, however, a vision-language model may over-infer accidents from congestion or visually confusing scenes. Detection evidence can provide stronger anchors for report generation.

To address these issues, this paper proposes YOLO-RSPD-ECA and a collaborative accident video understanding workflow. RSPDConv retains the stride-2 output scale while adding a pooling-based detail branch. ECA recalibrates accident-relevant channels with limited computational cost. The detector outputs bounding boxes, class labels, and key-frame evidence, which are then used by Qwen3.5-397B-A17B to generate structured accident reports. Expert review is used to evaluate report consistency and correctness.

The main contributions of this paper are as follows:

- An RSPDConv residual spatial-preserving downsampling module is introduced to compensate for local information loss without changing the output spatial scale of stride-2 convolution.
- A lightweight ECA channel attention mechanism is combined with YOLOv8, YOLO11, and YOLO26 to form a plug-in RSPD-ECA enhancement.
- A detector-guided vision-language workflow is constructed, linking detection results, key frames, textual reports, and expert review into a traceable evidence chain for accident video assessment.

The remainder of this paper is organized as follows. Section 2 reviews related work on traffic accident detection, multi-scale feature fusion, and vision-language models. Section 3 presents the YOLO-RSPD-ECA detection-enhancement method and the detection-guided vision-language assessment pipeline. Section 4 reports the dataset, ablation experiments, video assessment results, and representative cases. Section 5 discusses the implications, applicability, limitations of the method, and outlines future research directions.

2. Related Work

2.1 Traffic Accident Recognition Based on Visual Detection

Traffic accident recognition combines object detection, video anomaly detection, and traffic-scene understanding. Classical methods based on background modeling, trajectories, or hand-designed features are sensitive to weather, viewpoint change, occlusion, and complex traffic flow. Deep learning has improved spatial localization and temporal event representation in driving videos and surveillance videos [3,4,6].

Among the model of deep learning, the YOLO family is suitable for traffic deployment because it balances speed, accuracy, and implementation convenience. Nevertheless, accident-defining regions often occupy only a small part of the image. If early downsampling removes contact details, later feature fusion may not fully restore the evidence needed for reliable accident classification [9-12].

2.2 Multi-Scale Feature Fusion and Attention Mechanisms

Multi-scale feature fusion improves detection by combining high-level semantics with lower-level spatial details. EfficientDet uses bidirectional feature fusion and compound scaling, while

DETR shows the potential of end-to-end prediction [7,8]. YOLO-style detectors also depend on cross-layer aggregation to handle small and large objects within the same traffic scene [9-12].

Attention and residual design complement feature fusion. ECA-Net models local cross-channel interactions without dimensionality reduction, providing low-cost channel recalibration [17]. Residual learning supports stable compensation branches in deep networks [18]. The proposed RSPD-ECA design follows these principles by keeping the original stride-2 path and adding a gated detail branch plus lightweight channel attention.

2.3 Vision-Language Models and Traffic Accident Information

Vision-language models can convert visual evidence into natural-language descriptions. CLIP established transferable image-text alignment, BLIP-2 connects frozen visual encoders with large language models, and LLaVA demonstrates the value of visual instruction tuning [13-15]. These capabilities make such models useful for accident report generation, visual question answering, and scene explanation.

For accident assessment, open-ended generation must be constrained by evidence because police review and dispatch support require traceability. Multimodal retrieval-augmented generation studies emphasize structured external evidence for reducing unsupported generation [16]. In this paper, YOLO detections and key frames serve as evidence anchors before the language model produces reports.

3. Method

The method includes a detector-enhancement stage and a semantic-assessment stage. YOLOv8, YOLO11, and YOLO26 are used as base architectures; RSPDConv replaces eligible stride-2 convolutions, and ECA is inserted after backbone and neck feature blocks to form baseline, RSPD, ECA, and RSPD-ECA variants. The trained YOLO-RSPD-ECA model then samples videos, detects accident targets, selects key frames, aggregates class and confidence statistics, and provides this evidence to Qwen3.5-397B-A17B for report generation. Figure 1 shows the workflow.

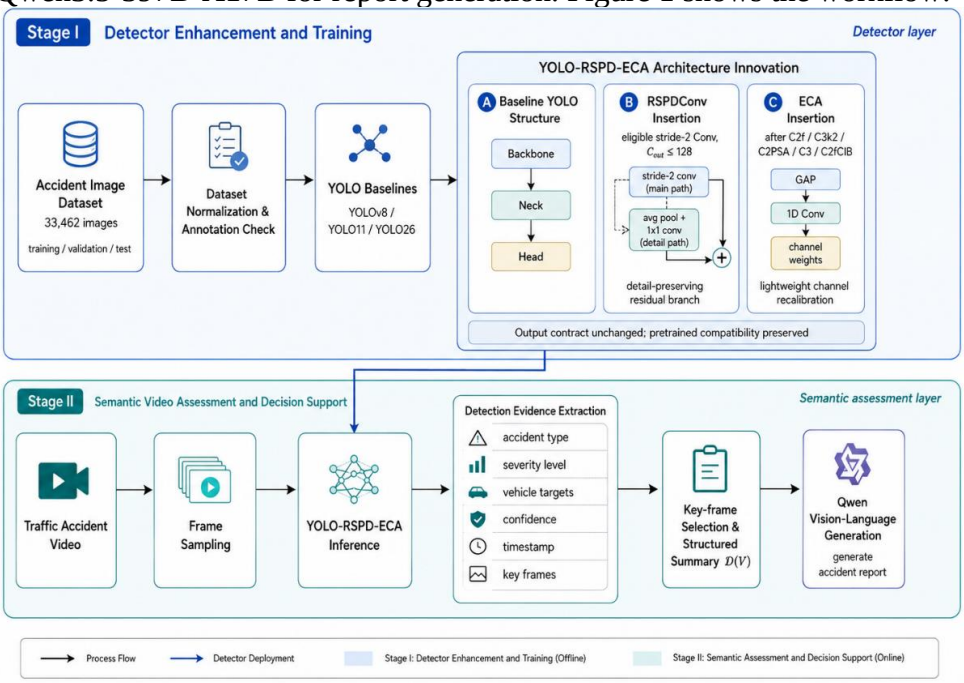

Figure 1. Overall technical workflow for traffic accident video understanding.

3.1 Detection Data Preprocessing

For an input video V = {I_t | t = 1, ..., T}, the sampling strategy selects frame indices S. The detector f_θ predicts boxes, labels, and confidence scores for each sampled frame. Key frames are ranked by target number, confidence, and accident-related class coverage, producing the evidence summary D(V):

$$B_t = f_\theta(I_t) = \{(b_i, c_i, p_i);\ i = 1, \ldots, n_t\},\quad t \in S. \tag{1}$$

$$D(V) = \{count(c), max(p), keyframe(K), timestamp(K)\}. \tag{2}$$

The summary and key frames are sent to the vision-language model so that reports are anchored by box locations, classes, confidence values, and timestamps. Figure 2 shows the plug-in detector architecture.

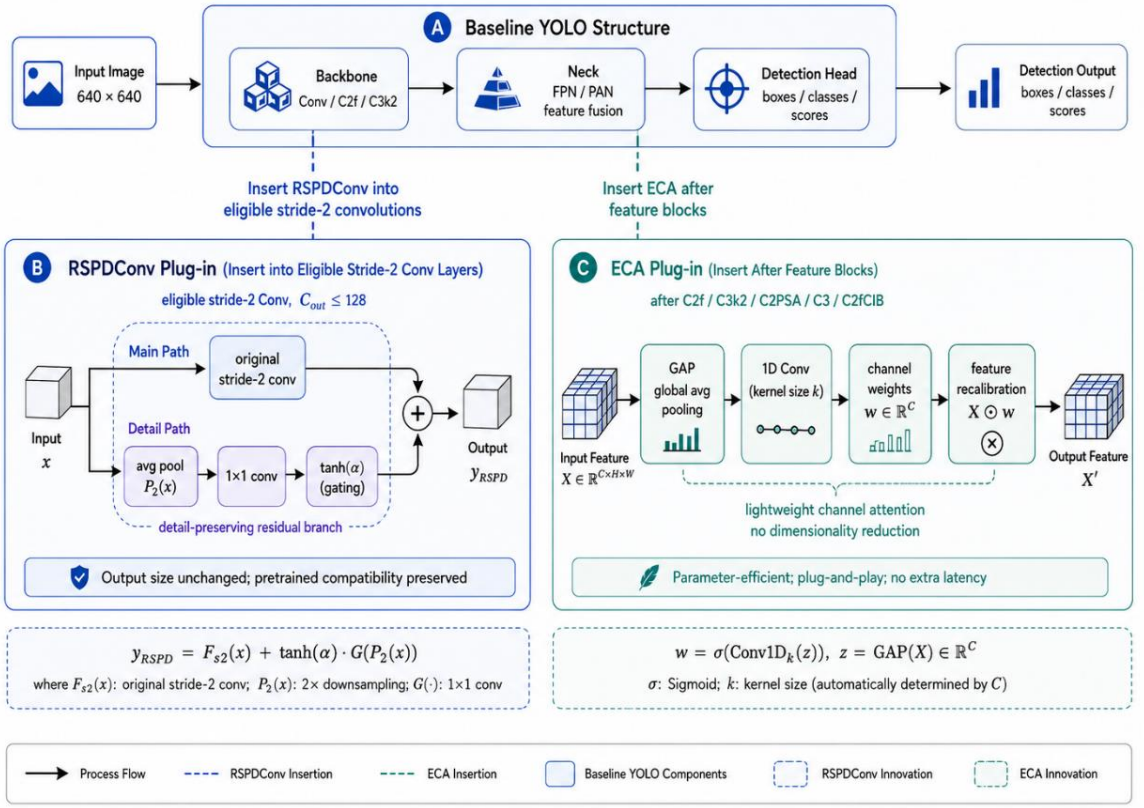

Figure 2. Plug-in architecture of YOLO-RSPD-ECA.

3.2 RSPDConv Residual Spatial-Preserving Downsampling Module

Standard stride-2 convolution reduces resolution and may discard collision-contact details. RSPDConv preserves the original stride-2 path and adds a pooled detail branch. This residual design recovers local structure while keeping the detector output scale unchanged.

Let x be the input feature, F_s2 the original stride-2 convolution, P_2 the average-pooling detail branch, G a 1×1 channel projection, and α a learnable gate:

$$y_{RSPD} = F_{s2}(x) + tanh(\alpha) \cdot G(P_2(x)). \tag{3}$$

The tanh gate limits the residual branch, and G aligns channel dimensions. RSPDConv is applied only to stride-2 layers with no more than 128 output channels, limiting overhead in deeper layers.

3.3 ECA Channel Attention Module

ECA uses global average pooling to obtain a channel descriptor and one-dimensional convolution to model local cross-channel interaction [17]. This lightweight recalibration is suitable for accident detection because it improves channel selectivity with small parameter cost.

$$z_c = \left(\frac{1}{H \cdot W}\right) \cdot \Sigma_i \Sigma_j x_c(i,j). \quad (4)$$

A one-dimensional convolution with kernel size k generates the channel weights:

$$w = \sigma(Conv1D_k(z)). \quad (5)$$

A learnable residual scale β is used to preserve training stability after module insertion:

$$y_{ECA} = x \cdot [1 + tanh(\beta) \cdot 2(w - 0.5)]. \quad (6)$$

When β is close to zero, the module approximates identity mapping. During training, it can emphasize channels associated with vehicle contact areas and suppress background interference.

3.4 Detection-Guided Vision-Language Assessment

During semantic assessment, sampled video frames are processed by YOLO-RSPD-ECA. The system records accident classes, Vehicle detections, confidence distributions, and timestamps. Frames are ranked by detection number, maximum confidence, and class coverage to select key frames and build D(V_i). Figure 3 shows the collaborative workflow.

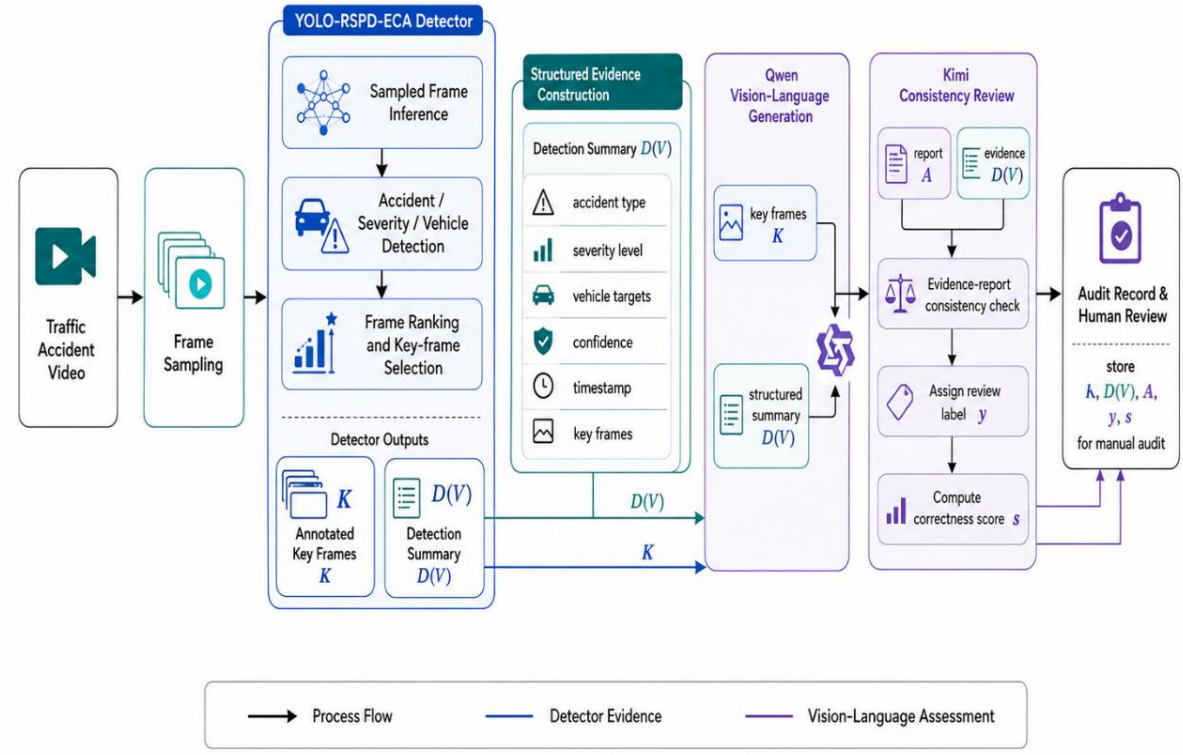

Figure 3. Collaborative assessment workflow combining YOLO-RSPD-ECA and a vision-language model.

For report generation, D(V_i) and K_i are provided to Qwen3.5-397B-A17B with a fixed task prompt. The prompt requests accident presence, severity level, risk location, vehicle involvement,

traffic impact, emergency suggestions, and supporting visual evidence, while prioritizing boxes, class statistics, and visible key-frame content.

The process is formalized as:

$$A_i = Q_\varphi(P_{task}, D(V_i), K_i), \quad D(V_i) = \{B_t, count(c), max(p), timestamp(K_i)\}. \tag{7}$$

4. Case Study

4.1 Dataset Description

The dataset is collected with local traffic police and traffic-management agencies, covering highways, mountainous roads, daytime scenes, and nighttime scenes. It contains 33,462 images and 43,826 YOLO-format target annotations, with no empty labels, missing labels, or illegal rows. Each image has 1.31 boxes on average, and the maximum is 38. The dataset includes Accident, Mild, Moderate, Noaccident, Severe, and Vehicle classes, and is split as shown in Table 1.

Table 1. Dataset split statistics.

Split	Number of images	Proportion
Training set	28,135	84.08%
Validation set	3,513	10.50%
Test set	1,814	5.42%
Total	33,462	100.00%

The distribution is long-tailed. Severe has 24,932 boxes (56.89%) and appears in 23,875 images; Moderate and Noaccident account for 17.92% and 16.63%, respectively; Vehicle contributes 7.65%. Accident and Mild together account for less than 1% of boxes, indicating sample scarcity for low-severity accident cases. Figure 4 shows representative samples.

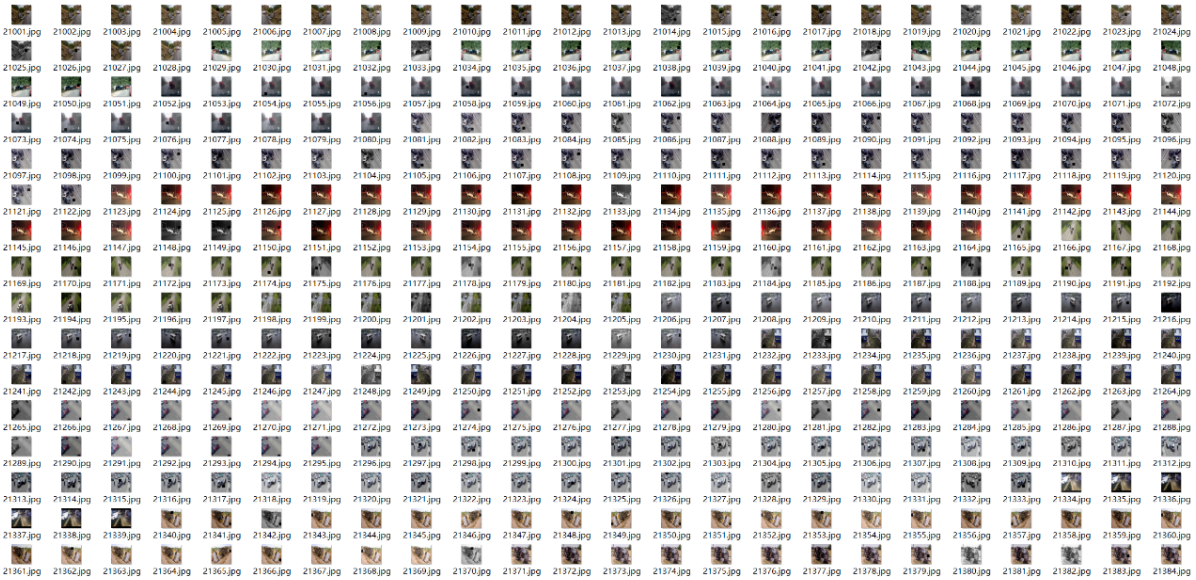

Figure 4. Partial dataset preview, illustrating traffic accident samples across multiple scenes and scales.

4.2 Ablation Experiment Results

Ablation experiments are conducted on YOLOv8, YOLO11, and YOLO26. For each architecture, Baseline, RSPD, ECA, and RSPD-ECA variants are trained on the training set and evaluated on the

validation set, forming 12 experiments. Table 2 reports the results.

Table 2. YOLO ablation comparison after corrected naming.

Architecture	Variant	Precision	Recall	mAP50	mAP50-95	Gain
YOLOv8	Baseline	0.5624	0.6600	0.5633	0.4438	--
YOLOv8	RSPD	0.5340	0.5775	0.5843	0.4626	+0.0188
YOLOv8	ECA	0.5710	0.6439	0.5969	0.4670	+0.0232
YOLOv8	RSPD-ECA	0.5425	0.7071	0.5975	0.4754	+0.0316
YOLO11	Baseline	0.5144	0.6741	0.5814	0.4521	--
YOLO11	RSPD	0.5420	0.6574	0.6050	0.4750	+0.0229
YOLO11	ECA	0.5436	0.6637	0.6127	0.4822	+0.0301
YOLO11	RSPD-ECA	0.5640	0.6587	0.6077	0.4834	+0.0313
YOLO26	Baseline	0.5649	0.6404	0.6082	0.4699	--
YOLO26	RSPD	0.5767	0.6135	0.5927	0.4737	+0.0038
YOLO26	ECA	0.5753	0.6356	0.6108	0.4814	+0.0114
YOLO26	RSPD-ECA	0.5903	0.6296	0.6077	0.4833	+0.0134

Table 2 shows consistent mAP50-95 gains from RSPD-ECA. YOLOv8 improves from 0.4438 to 0.4754, YOLO11 from 0.4521 to 0.4834, and YOLO26 from 0.4699 to 0.4833. The gains indicate that detail-preserving downsampling and channel recalibration transfer across YOLO backbones.

ECA alone produces stable gains, while RSPD mainly changes the precision-recall balance and improves localization quality. Their combination achieves the best mAP50-95 in all three architectures. The smaller YOLO26 gain is consistent with a stronger baseline and partial performance saturation. Figure 5 visualizes the gains.

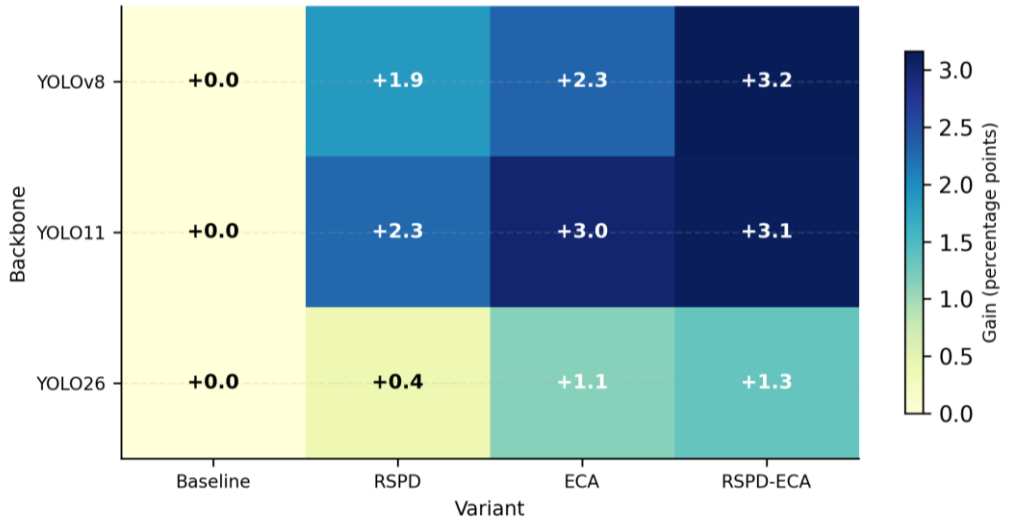

Figure 5. Absolute mAP50-95 gains of 12 YOLO combinations relative to their corresponding baselines under the corrected protocol.

4.3 Traffic Accident Video Understanding Results

Figure 6 presents selected test results, showing detection of accident-related targets across different scales and scenes.

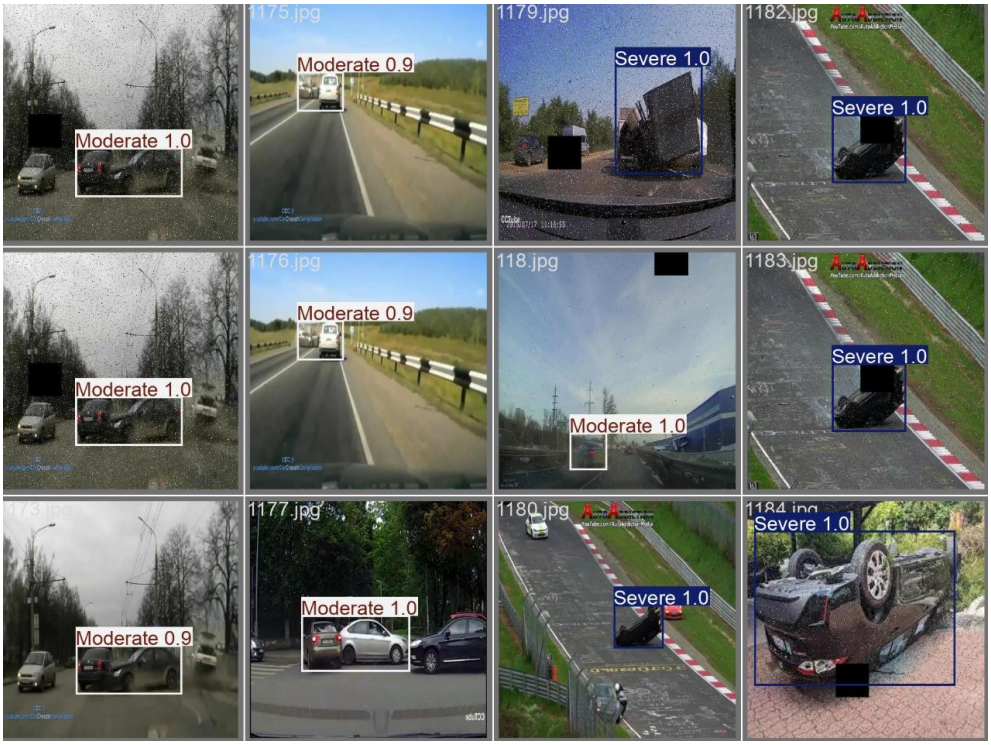

Figure 6. Detection visualization results on selected test samples.

For video understanding, YOLO11-RSPD-ECA is used as the real-time front end and connected to Qwen3.5-397B-A17B. The system completes frame sampling, detection, key-frame saving, report generation, and record archiving for all 101 videos. Each 5-10 min sample is analyzed in about 2 min. Figure 7 shows the interface and response-analysis example.

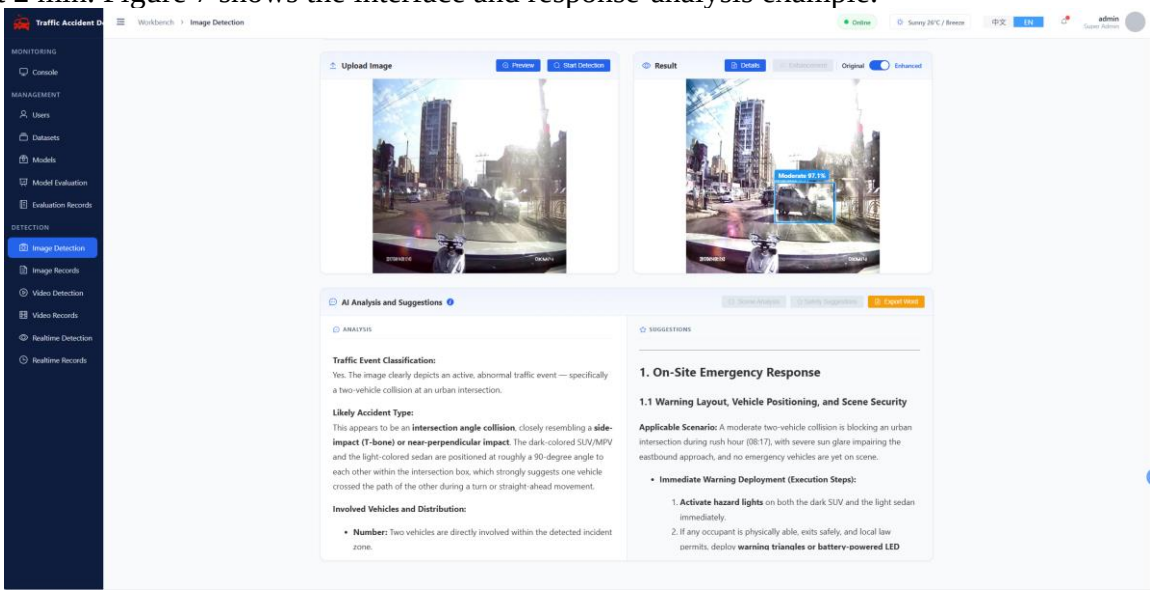

Figure 7. Example traffic accident detection and response-measure analysis combining the vision-language model with YOLO.

Table 3 shows that Correct and Basically correct reports total 86, giving an acceptable-report rate of 85.15%; the average expert correctness score is 0.833. The 12 Incorrect or Unparseable cases mainly arise from insufficient video evidence, ambiguous severity boundaries, or low detection confidence leading to over-specific generated judgments.

Table 3. Expert review statistics for video reports.

Expert review label	Count	Proportion
Correct	61	60.40%
Basically correct	25	24.75%
Partially correct	3	2.97%
Unparseable	3	2.97%
Incorrect	9	8.91%
Total	101	100.00%

5. Conclusion

This paper proposes YOLO-RSPD-ECA and a detection-guided vision-language assessment framework for traffic accident video understanding. RSPDConv reduces downsampling detail loss, ECA performs lightweight channel recalibration, and Qwen3.5-397B-A17B turns detection summaries and key frames into structured reports. Experiments on 33,462 images and 101 videos show consistent mAP50-95 improvements across YOLOv8, YOLO11, and YOLO26; the video reports reach an average expert correctness score of 0.833 and an acceptable-report rate of 85.15%. The framework is plug-and-play, lightweight, and traceable, making it suitable for traffic monitoring platforms with limited computing resources. Remaining limitations include the lack of calibration against external accident records, possible missed short collision moments under uniform sampling, and insufficient trajectory or temporal-causality modeling. Future work will incorporate tracking, temporal modeling, and dispatch-feedback calibration.

References

[1] World Health Organization. Global status report on road safety 2023. Geneva: World Health Organization, 2023. https://www.who.int/publications/i/item/9789240086517.

[2] Wang K, Li Z. Global, regional, and national burdens of road injuries from 1990 to 2021: Findings from the 2021 Global Burden of Disease Study. Injury, 2025, 56: 112221.

[3] Yao Y, Wang X, Xu M, Pu Z, Wang Y, Atkins E, Crandall D J. DoTA: Unsupervised detection of traffic anomaly in driving videos. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(1): 444-459.

[4] Pawar K, Jalem R S, Tiwari V. Deep learning based detection and localization of road accidents from traffic surveillance videos. ICT Express, 2022, 8(3): 379-387.

[5] Kumar P P, Kant K. TU-DAT: A computer vision dataset on road traffic anomalies. Sensors, 2025, 25(11): 3259.

[6] Liu H, Hu X, Sun G, Zhang W, Zhan J, Fang H, Li Y, Ma W. Intelligent traffic accident detection system in complex dynamic scenarios based on the dual-stream spatiotemporal-fusion model. Engineering Applications of Artificial Intelligence, 2025, 162(Part E): 112675.

[7] Tan M, Pang R, Le Q V. EfficientDet: Scalable and Efficient Object Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020: 10781-10790.

[8] Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-End Object Detection with Transformers. In: Proceedings of the European Conference on Computer Vision (ECCV), 2020: 213-229.

[9] Wang C Y, Bochkovskiy A, Liao H Y M. Scaled-YOLOv4: Scaling Cross Stage Partial Network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 13029-13038.

[10] Ge Z, Liu S, Wang F, Li Z, Sun J. YOLOX: Exceeding YOLO Series in 2021. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2021: 988-999.

[11] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023: 7464-7475.

[12] Wang A, Chen H, Liu L, Chen K, Lin Z, Han J, Ding G. YOLOv10: Real-Time End-to-End Object Detection. In: Advances in Neural Information Processing Systems (NeurIPS), 2024, 37.

[13] Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning Transferable Visual Models from Natural Language Supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML), 2021: 8748-8763.

[14] Li J, Li D, Savarese S, Hoi S. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In: Proceedings of the 40th International Conference on Machine Learning (ICML), 2023: 19730-19742.
[15] Liu H, Li C, Wu Q, Lee Y J. Visual Instruction Tuning. In: Advances in Neural Information Processing Systems (NeurIPS), 2023, 36: 34892-34916.
[16] Abootorabi M M, Zobeiri A, Dehghani M, et al. Ask in Any Modality: A Comprehensive Survey on Multimodal Retrieval-Augmented Generation. In: Findings of the Association for Computational Linguistics: ACL 2025, 2025: 16776-16809.
[17] Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020: 11534-11542.
[18] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 770-778.