LEARNING VEHICLE-CENTRIC RISK REPRESENTATION FOR DRIVING HAZARD IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Identifying hazards that threaten the ego vehicle is a critical capability for intelligent driving systems: it enables timely warnings to drivers and the screening of safety-critical samples from massive driving data. In this paper, we study *driving hazard identification*: given an ego-view image, determine whether the scene contains hazards relevant to the ego vehicle. This task resists supervised recognition, since hazardous events are rare, diverse, and unpredictable, arising from open-ended combinations of traffic participants, spatial layouts, and behaviors. General-purpose VLMs can query hazards directly with their open-world reasoning; yet without an understanding of which factors actually threaten the ego vehicle, their judgments are easily misled by safety-irrelevant content. We propose RISKREP, which addresses both limitations with a learned, vehicle-centric risk representation. Instead of asking a VLM to judge danger from raw pixels, RISKREP first constructs a *vehicle-centric graph* (VCG): each traffic participant becomes a node carrying its category attribute, while its edge to the ego vehicle encodes the relative position, heading, and motion. RISKREP then compresses this graph into a compact representation through shared encoding and attention-based aggregation, and trains a risk scoring function on top of it with scene-level hazard supervision. To locate hazard sources, we use the learned scoring function to search for a risk-relevant sub-VCG via counterfactual node removal, and render it as structured hazard cues that guide a frozen VLM toward the final hazard identification. Because RISKREP captures general ego-relative relations rather than specific hazards, our method can generalize to hazard combinations unseen in training. We build a hazard identification benchmark from DrivingDojo and BDD100K, and comprehensive experiments show that RISKREP consistently improves accuracy and F1-scores across VLMs of various scales.

1 INTRODUCTION

Identifying traffic hazards that could threaten the safety of the ego-vehicle is a fundamental capability of intelligent driving systems. On one hand, accurate hazard identification enables timely risk warnings for drivers or autonomous driving systems; on the other, it facilitates the filtering of safety-critical scenarios from vast amounts of naturalistic driving data, providing high-value samples for system testing and model training. Unlike traditional visual recognition tasks, driving hazards are not typically defined merely by the presence of a specific type of traffic participant. The same vehicle, pedestrian, or cyclist can represent either a safe or hazardous situation depending on their location, motion state, and the surrounding traffic configuration. Consequently, hazard identification relies more on the traffic participants' positions relative to the ego-vehicle, their motion trends, and the contextual relationships formed by multiple participants, rather than on mere visual appearance or object categories (Fang et al., 2019; Li & Chan, 2023; Malla et al., 2023). This characteristic results in significant diversity and openness regarding driving hazards: potential hazards can arise from open-ended combinations of different traffic objects, spatial layouts, and behavioral patterns. Existing literature has already linked risk to the relative positions, motion states, and overall traffic configurations of traffic participants.

In recent years, vision-language models (VLMs) have demonstrated powerful capabilities in understanding scene semantics, making the use of natural language prompts to assess driving scene risks a simple yet highly attractive approach. However, the internal representations of general-purpose

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

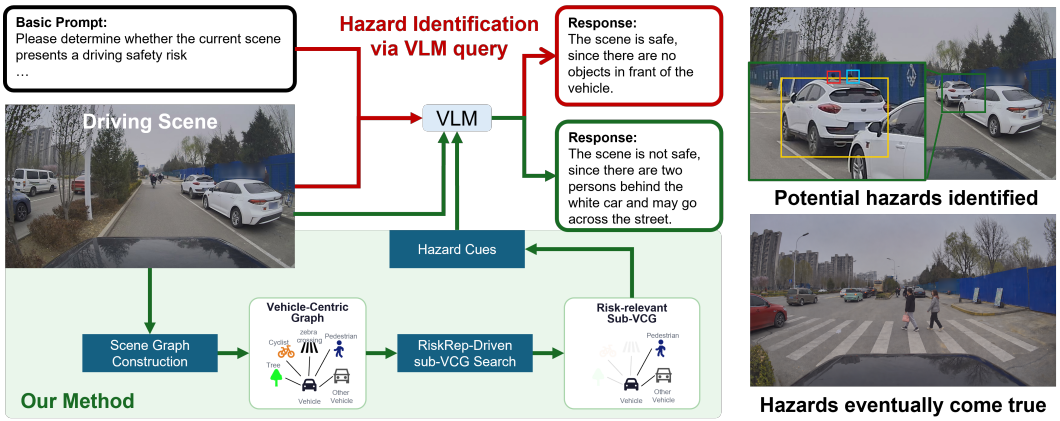


Figure 1: Our method vs. direct VLM query. RISKREP first organizes the driving scene into a vehicle-centric graph (VCG), from which a compact risk representation is learned. It then searches the risk-relevant subgraph and converts it into hazard cues that direct the VLM to the two pedestrians obscured by the white vehicle missed by the baseline: the source of the hazard in this example.

VLMs are not specifically designed for modeling driving risks, since driving scenes involve processing vast amounts of visual content(e.g. road structures, the surrounding environment, vehicles, pedestrians, and traffic infrastructure), not only a fraction of this information truly determines the safety of the current scene, typically involving just a few traffic participants that share critical spatial or motion relationships with the ego-vehicle. Furthermore, existing research indicates that VLMs still face limitations regarding fine-grained visual grounding and the understanding of spatial relationships. (Wen et al., 2024; Tian et al., 2025; Chen et al., 2024)

To investigate this gap, we first conducted a direct assessment of driving risk using general-purpose VLMs of varying scales. Experimental results reveal a pervasive and significant precision–recall imbalance across the models. Simply increasing model scale does not consistently solve this issue, indicating that superior general-purpose visual-semantic capabilities do not automatically yield more reliable risk representations. These results indicate that, although general-purpose VLMs possess extensive visual-semantic knowledge, their representations of general scenes do not naturally give rise to a representation space geared toward assessing hazards for the ego-vehicle. In other words, the key of the matter lies not merely in whether the model *understands* the image, but in whether it can identify which traffic factors are truly relevant to the risk faced by the ego-vehicle.

To address this issue, we propose RISKREP, a vehicle-centric risk representation learning framework designed for hazard identification in driving scenes. Unlike approaches that require VLMs to directly assess hazards from raw images, RISKREP first organizes driving scenes into a **Vehicle-Centric Graph** (VCG), where traffic elements serve as nodes while explicitly preserving their spatial and motion relationships with the ego-vehicle. Building on this, we encode VCGs of varying scales into compact, vehicle-centric risk representations through shared node encoding and attention-based aggregation, learning their risk-relevant structures under scene-level hazard supervision. After that, RISKREP employs the risk-scoring function to evaluate the contribution of various traffic nodes to the scene's risk response and searches the complete VCG for a risk-relevant sub-VCG, thereby retaining potential hazard structures that warrant further inspection. Finally, this risk-relevant sub-VCG is converted into structured hazard cues to guide the general-purpose VLM in focusing on potential hazard factors relevant to the ego-vehicle, while the VLM still performs the final hazard identification by integrating these cues with the visual scene.

We conducted a systematic evaluation using a driving risk benchmark constructed from the DrivingDojo, BDD100K datasets, featuring scene-disjoint data and balanced categories. Experiments demonstrate that directly employing general-purpose VLMs, supervised finetuning, and structured baselines always suffer from varying degrees of precision-recall imbalance. In contrast, RISKREP achieves the best performance in terms of both accuracy and F1-score across VLM backbones of all three scales.

In summary, this paper makes the following contributions:

- We study vehicle-centric risk representation learning for driving hazard identification, where hazards arise from open-ended combinations of ego-relevant traffic participants, spatial configurations, and motion states;
- We propose RISKREP, which organizes driving scenes as Vehicle-Centric Graphs, learns compact hazard-oriented representations, and searches for risk-relevant sub-VCGs to provide structured hazard cues for VLM reasoning;
- We conducted systematic experiments across multiple driving datasets and VLMs of varying scales; the results demonstrate that RISKREP achieves superior overall hazard identification performance compared to direct VLM inference, supervised fine-tuning, and existing structured methods.

2 RELATED WORK

Driving risk representation and risk reasoning. In recent years, research on driving risk has gradually shifted from accident prediction toward more fine-grained risk identification and reasoning. RiskBench systematically evaluates risk identification methods across aspects such as risk detection, localization, and decision support (Kung et al., 2024); HiLM-D further jointly models risk object localization, intent understanding, and safety recommendations (Ding et al., 2025). DRAMA-X performs compositional risk reasoning using scene graphs and intent inference (Godbole et al., 2025), while DriveSafe transforms spatial, motion, and depth information into structured scene descriptions for risk representation and safety recommendation generation (Artham et al., 2026). Recent work has also begun leveraging multimodal large models for threat-oriented reasoning and driving risk prediction (Wang et al., 2026). INSIGHT(Chen et al., 2026) leverages vision-language models with multimodal semantic-visual fusion and supervised fine-tuning to improve context-aware hazard detection and spatial hazard localization in challenging driving scenarios. Unlike methods that primarily perform direct risk identification or reasoning, RISKREP focuses on risk representation learning. We first compress the set of traffic participants into a compact, permutation-invariant risk representation, and then employ risk factor analysis to extract key risk participants and their combinations for subsequent risk representation and VLM reasoning.

VLM for autonomous driving. VLMs have been widely used for scene understanding and reasoning in autonomous driving. DriveLM organizes perception, prediction, and planning through Graph VQA(Sima et al., 2024), while Reason2Drive and Dolphins focus on chain-based reasoning and grounded multimodal reasoning, respectively. Going further, Talk2BEV introduces BEV scene structures to enhance spatial grounding and risk reasoning(Nie et al., 2024; Ma et al., 2024). Unlike directly training or enhancing a full driving VLM, RISKREP first learns an vehicle-centric risk representation and only converts the key risk elements and objects identified through attribution into structured hazard cues. So, in this work, the VLM acts as a downstream consumer of risk representation rather than the risk representation itself(Choudhary et al., 2024).

3 PROBLEM FORMULATION AND CHALLENGES

3.1 PROBLEM DEFINITION

Given an image S of a driving scene from the ego-vehicle's perspective, this paper investigates the task of *driving hazard identification*, which aims to determine whether there are potential hazards in the current scene that could affect the ego-vehicle's normal operation. We use binary labels $y \in \{0, 1\}$ to label the hazard existence

Unlike typical visual recognition tasks, driving hazards are not independently determined by the category or visual appearance of a specific traffic participant; rather, they depend on the participant's position and motion state relative to the ego-vehicle, as well as the scene configuration formed by multiple traffic participants. Consequently, traffic objects of the same category may correspond to either hazardous or non-hazardous states depending on their spatial and motion relationships. Therefore, we define the scene risk representation as

$$z \in \mathbb{R}^d, \tag{1}$$

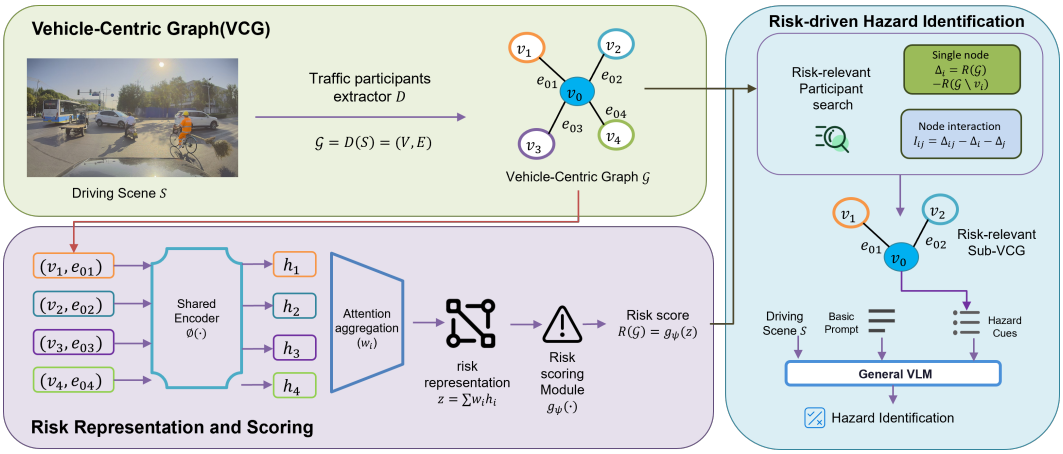


Figure 2: Overview of RISKREP

where z is a compact representation of the scene, which can preserve information relevant to the driving hazards. An ideal risk representation should highlight the traffic factors and their configurations that determine the risk level of the current scene, while suppressing visual information irrelevant to hazard assessment. Ultimately, driving hazard identification can be formulated as

$$\hat{y} = h(z) \tag{2}$$

where $h(\cdot)$ represents a prediction function that performs scene-level hazard identification. This paper aims to learn a risk representation z capable of effectively characterizing hazard factors relevant to the ego-vehicle, thereby providing a structured basis for driving hazard reasoning by general-purpose VLMs.

3.2 CHALLENGE

	F1-score(%)	Accuracy(%)	Precision(%)	Recall(%)
Qwen3.5-9B	59.48	56.45	46.67	81.99
Qwen3.6-35B-A3B	66.81	73.63	65.58	68.08
Qwen3.8-27B	68.09	67.90	55.64	87.74
Qwen3.8-Flash(125B-A6B)	69.99	69.57	54.70	97.02
GLM-5.3-Flash(320B-A18B)	66.94	63.28	51.57	95.36

Table 1: Hazard identification performance of general-purpose VLMs under direct inference. All models are evaluated using their default reasoning settings.

We first evaluate direct VLM inference on DrivingDojo and BDD100K. As shown in Table 1, general-purpose VLMs exhibit a consistent precision–recall imbalance, with several models achieving high recall at the cost of substantial false positives. Increasing model scale does not consistently mitigate this issue, suggesting that general visual-semantic capability does not directly translate into ego-centric hazard representations.

4 METHOD

4.1 OVERVIEW

Figure 2 shows the overview of RISKREP.

Given a vehicle-centric driving scene S, we first represent it as a Vehicle-Centric Graph (VCG)

$$\mathcal{G} = (V, E) \tag{3}$$

where the node set $V = \{v_0, v_1, ..., v_n\}$ comprises the ego vehicle node v_0 and nodes v_i representing traffic participants in the scene, and each node records attributes such as the category, dimensions,

and motion state of the corresponding traffic element. The edge set $E = \{e_{0i} | i = 1, ..., n\}$ connecting only the ego-vehicle v_0 with other traffic participants v_i, describes their position, orientation, and motion relationship relative to the ego vehicle.

Building upon the VCG, RISKREP learns a vehicle-centric risk representation for driving hazards. We employ shared encoding for the traffic nodes within the VCG and their relative relationships with the ego vehicle, subsequently compressing the VCG with various numbers of nodes into a fixed-dimensional representation z via attention-based aggregation. Finally, we train a risk scoring function using scene-level hazard supervision, enabling us to assign a risk score to a given VCG.

Based on the learned risk-scoring function, we further perform node analysis to identify nodes most relevant to the scenario risk from the complete VCG and construct a risk-relevant sub-VCG $\mathcal{G}_r \subseteq \mathcal{G}$. Finally, RISKREP converts risk-relevant traffic nodes within the sub-VCG, the ego vehicle's relative state, and critical inter-node risk interactions into structured hazard cues, along with the original scene, and feeds them into the general-purpose VLM, which performs the final driving hazard identification.

4.2 VEHICLE-CENTRIC GRAPHS (VCGS)

After defining the VCG in the 4.1, this subsection further explains how to construct the corresponding graph structure from the input driving scenario S. We first employ a traffic participants extractor D to identify relevant participants within the scene from the raw image. For each detected element v_i, the extractor further estimates its category, dimensions, orientation, location, and motion state, thereby transforming the raw visual scene into a set of explicit, structured traffic information.

Then, each traffic participants is represented as a node feature composed of two types of information:

$$v_i = [a_i, m_i] \tag{4}$$

where a_i captures the appearance attributes of the traffic element, such as its semantic category and physical dimensions, while m_i describes its motion attributes, including heading and motion state. Its spatial relation to the ego vehicle is represented separately by the corresponding vehicle-centric edge.

4.3 VEHICLE-CENTRIC RISK REPRESENTATION AND SCORING

4.3.1 VCG-DRIVEN RISK ENCODING

Given a constructed VCG, we further encode the variable number of traffic participants within it into a fixed-dimensional scene representation. Since the number of traffic elements varies across driving scenarios and the ordering of nodes lacks inherent semantic meaning, the encoding process must accommodate inputs of **variable scale** while ensuring **permutation invariance** regarding node order.

To address this requirement, we adopt the fundamental idea of Deep Sets (Zaheer et al., 2017) by independently encoding various traffic nodes using a shared function and generating a scene-level representation through an aggregation operation that is invariant to node ordering. Specifically, for each traffic node and its corresponding vehicle-centric relation, we employ a shared encoder $\phi(\cdot)$ to map it into a unified latent space:

$$h_i = \phi(v_i, e_{0i}) \tag{5}$$

where h_i denotes the latent representation of the i-th traffic participant under the vehicle-centric context, and all nodes share the same encoding parameters. This mechanism enables different traffic elements to be represented within a unified feature space, while ensuring that the encoding structure itself remains unaffected by the number of nodes or their permutation order.

However, compared to other image tasks, the importance of various traffic elements for scene-level hazard identification often differs in driving scenarios: numerous background traffic elements may provide only limited information, whereas a small number of objects which bearing critical spatial or motion relationships to the ego-vehicle may play a dominant role in risk assessment. Therefore, while preserving permutation invariance, we further employ **attention-based aggregation**, enabling the model to adaptively adjust the contribution of different nodes to the final representation based

on the current scene. For each node representation h_i, we first compute its weight w_i using a shared attention scorer $\alpha(\cdot)$:

$$w_i = \frac{\exp(\alpha(h_i))}{\sum_{j=1}^{n} \exp(\alpha(h_j))} \tag{6}$$

Then, the entire VCG is aggregated into a fixed-dimensional scene risk representation z as:

$$z = \sum_{i=1}^{n} w_i h_i \tag{7}$$

Since the attention scorer is shared across all nodes, and both normalization and weighted summation operate on the set of nodes without depending on their ordering, the aggregation process preserves permutation invariance. Moreover, compared to fixed symmetric pooling, attention-based aggregation offers a learnable node-weighting mechanism, allowing different traffic elements to contribute differently to the representation z based on their importance within the current scenario. The resulting z constitutes a vehicle-centric representation for RISKREP and serves as input for the subsequent risk scoring.

4.3.2 RISK ENCODING TRAINING

This representation z preserves vehicle-centric information regarding distinct traffic elements from the VCG; however, relying solely on the encoding structure does not guarantee that the representation effectively captures driving hazards. To align the representation space with scene-level hazard semantics, we further introduce scene-level supervision signals to jointly train the VCG encoder and the risk scoring module.

We construct a risk-scoring module $g_\psi(\cdot)$ on the representation z to generate the hazard prediction distribution:

$$\hat{p} = \mathrm{softmax}\left(g_\psi(z)\right) \tag{8}$$

where $\hat{\mathbf{p}} = [\hat{p}_0, \hat{p}_1]$ represent the predicted probabilities for the non-hazardous and hazardous classes, respectively. Given the scene-level hazard label $y \in \{0, 1\}$, the entire representation learning process is optimized using cross-entropy loss:

$$\mathcal{L}_{\mathrm{risk}} = \mathbb{E}_{(\mathcal{G},y)\sim\mathcal{B}}\left[-\log \hat{p}_y\right] \tag{9}$$

where $\mathcal{B}$ denotes the training batch. During training, risk supervision is backpropagated via g_ψ to the preceding shared node encoder and attention scorer, enabling the joint adjustment of node representations h_i and their aggregation weights w_i based on scene-level danger assessments.

After training, z is no longer merely a general compression of the traffic information in the VCG; instead, it forms a vehicle-centric risk representation geared towards driving hazard identification. Meanwhile, we use the probability corresponding to the *hazardous* category in the predicted distribution as the continuous risk score for the complete VCG:

$$R(\mathcal{G}) = \hat{p}_1 = \left[\mathrm{softmax}\left(g_\psi(z)\right)\right]_1 . \tag{10}$$

Noted that, although $R(\mathcal{G})$ could be converted into a binary prediction via thresholding, this paper retains it as a continuous risk response function rather than using it as the final determination of scene risk. Directly thresholding $R(\mathcal{G})$ would compress fine-grained variations in risk response and cause the final judgment to rely excessively on the structured representation itself. In contrast, the continuous form captures response changes resulting from structural perturbations, thereby providing a unified scoring basis for subsequent node contribution analysis and the search for risk-related substructures, while preserving the scope for a downstream VLM to make the final determination by integrating the original visual scene.

4.4 RISKREP-DRIVEN VISION-LANGUAGE HAZARD IDENTIFICATION

4.4.1 RISKREP-DRIVEN SUB-VCG SEARCHING

Following risk representation training, the scoring function $R(\mathcal{G})$ is capable of measuring the full VCG's response to the *hazard* category. However, the scene-level risk score alone does not reveal

which traffic elements might be the source of the hazard. To further pinpoint potential risk factors, we perform a leave-one-out risk contribution analysis on the nodes of the VCG based on $R(\mathcal{G})$ and, on this basis, search for a risk-relevant sub-VCG within the complete VCG.

For any traffic node v_i, we remove the node from the current VCG and recalculate the risk score of the remaining graph structure. The risk contribution of node v_i is defined as

$$\Delta_i = R(\mathcal{G}) - R(\mathcal{G} \setminus v_i) \tag{11}$$

A large positive value of Δ_i indicates a significant reduction in the scene risk score upon the removal of v_i; thus, this node contributes substantially to the current risk representation.

In addition to the impact of individual nodes, certain potential risks may also be associated with the simultaneous presence of multiple traffic elements. To characterize this combined effect, we further conduct a **pairwise removal analysis**. For nodes v_i and v_j, the change in risk resulting from their simultaneous removal is defined as:

$$\Delta_{i,j} = R(\mathcal{G}) - R(\mathcal{G} \setminus \{v_i, v_j\}) \tag{12}$$

and further calculate its pairwise risk interaction:

$$I_{ij} = \Delta_{ij} - \Delta_i - \Delta_j \tag{13}$$

When $I_{ij} > 0$, the risk impact resulting from the co-occurrence of the two nodes exceeds the sum of their individual contributions, indicating a risk-amplifying interaction between them.

After obtaining the node-level risk contributions and node-interaction pairs, we select the top K nodes with the highest contributions from all nodes with positive contributions and retain their original vehicle-centric edges with the ego vehicle, thereby constructing the risk-relevant sub-VCG:

$$V_r = \mathrm{TopK}_{i:\Delta_i>0}\left(\Delta_i\right), \qquad \mathcal{G}_r = (\{v_0\} \cup V_r, E_r) \tag{14}$$

Throughout this process, $\mathcal{G}_r$ maintains the same vehicle-centric topology as the original VCG, retaining only those traffic nodes that contribute significantly to the current risk score.

4.4.2 VLM-DRIVEN HAZARD INFERENCE

Through risk-relevant sub-VCG extraction, we have identified the nodes most relevant to the scene's risk response; however, general-purpose VLMs cannot directly utilize graph structures or continuous risk contribution scores for reasoning. Therefore, we further convert this information into structured hazard cues to guide the VLM in focusing its attention on traffic factors more likely to impact the ego-vehicle's safety.

Structured hazard cues preserve the categories, relative positions, motion states, and relationships with the ego vehicle of traffic elements within the sub-VCG, thereby describing—in a compact form—potential hazard factors in the current scene that warrant further attention. It should be noted that these cues are derived solely from the filtered scene structure and do not incorporate the scene-level risk score $R(\mathcal{G})$.

Finally, we feed the raw driving scene S, the fixed hazard identification prompt P_{hazard}, and the structured hazard cues c_r into the general-purpose VLM:

$$\hat{y} = \mathrm{VLM}(S, P_{hazard} \oplus c_r) \tag{15}$$

where, $\oplus$ denotes the combination of hazard cues with base prompts, and $\hat{y}$ represents the final scene-level hazard identification result. In this framework, RISKREP is responsible for filtering and organizing potential hazard-related structures from the complete VCG, while the VLM integrates these cues with the raw visual information to perform the final hazard inference.

5 EVALUATION

5.1 EXPERIMENT SETTINGS

Datasets We utilized two widely used traffic scenario datasets: DrivingDojo(Wang et al., 2024) and BDD100K(Yu et al., 2020). And we randomly sampling a total of 20,000 scenarios for our

Table 2: Performance comparison of RISKREP and other methods

		F1-score(%)	Accuracy(%)	Precision(%)	Recall(%)
Qwen3.5-9B	Origin	59.48	56.45	46.67	81.99
	Finetune	69.28	66.32	53.75	97.45
	Drivesafe	67.78	75.93	70.88	64.94
	SGG-Intent	73.21	65.61	61.94	89.49
	RISKREP	74.64	78.53	69.16	81.08
Qwen3.6-35B-A3B	Origin	66.81	73.63	65.58	68.08
	Finetune	80.85	83.57	74.08	88.99
	Drivesafe	64.39	75.79	75.47	56.14
	SGG-Intent	69.17	74.04	61.94	78.30
	RISKREP	83.79	87.15	82.49	85.12
Qwen3.8-27B	Origin	68.09	67.90	55.64	87.74
	Finetune	74.62	74.05	60.30	97.84
	Drivesafe	65.94	74.84	69.79	62.50
	SGG-Intent	63.76	72.53	62.60	64.96
	RISKREP	85.99	88.11	79.51	93.61

experiments. To ensure diversity, no more than one scenario was extracted from any single video across the original datasets. Scenarios involving a collision or near-collision within the next three seconds were classified as risk, while those that did not were classified as safe. We randomly split the training and test sets in an 8:2 ratio.

Model Selection We conducted experiments using three VLMs of varying sizes and generations: Qwen3.5-9B(Qwen Team, 2026a), Qwen3.6-35B-A3B(Qwen Team, 2026b) and Qwen3.8-27B(Qwen Team, 2026c). To facilitate subsequent fine-tuning and integration with other methods, all models were deployed locally (in FP8 format), thus large-scale models(exceed 50B) in table 1 are exclude in this experiments. In addition, we utilize Qwen3.5-2B as the lightweight model to extract NPCs in section 4.2. In fact, for this simple extraction task, a 2B size model is sufficient for our method; we have also demonstrated this through ablation studies in section 5.3.

Comparison Methods To evaluate the effectiveness of RISKREP, we compared it with solutions for various similar tasks, as shown below.

Origin VLM Using general-purpose, open-source VLMs is the simplest and most straightforward approach, we simply put the image and prompts to the origin VLM to get the results.(The details of prompt see Appendix)

Finetune Once a dataset is available, fine-tuning is the most intuitive method for improving a model's performance on a specific task. It must be noted, however, that fine-tuning takes the most computation power of all the methods.

DriveSafe DriveSafe(Artham et al., 2026) converts driving scenes into structured natural-language descriptions containing spatial, motion, and depth information, and subsequently performs risk representation based on these explicit scene semantics.

SGG-Intent SGG-Intent(Godbole et al., 2025) constructs a scene graph from the input image to explicitly represent traffic participants and their contextual relations, and then employs an LLM to perform compositional reasoning for intent prediction, risk representation, and action suggestion.

Note that the open source code for DriveSafe and SGG-Intent were conducted on different datasets; we replicate them with minimal changes and run them on the same dataset of RISKREP.

5.2 RESULTS

Table 2 reports the performance comparison between RISKREP and four baseline methods across three VLM models of varying scales. Overall, RISKREP achieves the best Accuracy and F1-score across all model scales, demonstrating superior comprehensive performance and scalability. Notably, RISKREP achieves the best results in the entire table on the Qwen3.6-35B-A3B model, with an Accuracy of 87.15% and an F1-score of 83.79%; these figures represent improvements of 13.52%

and 16.98% over the original model, and 3.58% and 2.94% over the strongest baseline (Finetune). It also performs exceptionally well on the Qwen3.8-27B model, achieving an Accuracy of 88.11% and the highest F1-score in the table at 85.99%.

A closer examination of the failure modes of the baseline methods reveals a common imbalance between precision and recall: fine-tuning pushes recall to extremely high levels (e.g., 97.84% for Qwen3.8-27B) but causes a significant drop in precision (to only 60.30%), demonstrating a clear tendency toward "over-triggering"—where the model classifies a large number of samples as risky, thereby sacrificing control over false positives. Conversely, Drivesafe achieves higher precision but lower recall on larger models (only 56.14% for the 35B model), resulting in a high rate of missed risky samples. This one-sided bias drags down the F1 scores for both baseline categories. SGG-Intent exhibits middling and unstable performance across different model scales. In contrast, RISKREP achieves a balance between precision and recall (e.g., 79.51% / 93.61% for the 27B model), validating that its risk detection mechanism effectively manages both missed detections and false positives.

Furthermore, the advantages of RISKREP remain consistent across different model scales: ranging from the smaller 9B model to the larger 35B model, the F1-score improvements over the best baseline are $+1.43\%$, $+2.94\%$, and $+11.37\%$, respectively. The particularly large gain on Qwen3.8-27B suggests that the proposed risk representation can substantially improve hazard identification even when the underlying VLM already exhibits high recall.

5.3 ABLATION STUDY: EFFECTIVENESS OF RISK-RELEVANT SUB-VCG EXTRACTION

Due to space constraints, the ablation study on the Effectiveness of Attention Pooling Strategy in Appendix A.2, and the ablation study on the lightweight VLM used for NPC extraction is provided in Appendix A.3.

Table 3: Final VLM-based hazard identification performance under different VCG selection strategies. For each scene, the random sub-VCG contains the same number of nodes as the corresponding risk-relevant sub-VCG.

VCG Selection Strategy	F1-score (%)	Accuracy (%)	Precision (%)	Recall (%)
Origin VLM	66.81	73.63	65.58	68.08
Full VCG	80.90	84.63	78.52	83.42
Random sub-VCG	73.65	81.57	83.21	66.07
Risk-relevant sub-VCG	83.79	87.15	82.49	85.12

As shown in Table 3, using the full VCG improves performance over the origin VLM, but does not achieve the best results. This suggests that simply providing all traffic participants introduces redundant or weakly relevant information, which may dilute the truly hazardous cues and make it harder for the VLM to focus on the participants most responsible for the scene risk. In contrast, the risk-relevant sub-VCG achieves the highest F1-score of 83.79%, indicating that explicitly filtering the VCG to retain risk-relevant elements provides more focused guidance for hazard identification.

6 CONCLUSION

This paper investigates the problem of vehicle-centric driving hazard identification and proposes RISKREP, a vehicle-centric risk representation learning framework. The method models driving scenes as vehicle-centric graphs, learns risk representations through shared encoding, attention aggregation, and scene-level supervision, and selects risk-relevant sub-VCGs based on node contributions. These subgraphs are converted into structured hazard cues that guide the VLM to make the final judgment. Experimental results demonstrate that RISKREP achieves superior accuracy and F1 scores across VLMs of various scales, validating the effectiveness of vehicle-centric risk representation and risk-relevant structure selection in enhancing the driving hazard identification capabilities of VLMs.

REFERENCES

Sainithin Artham, Shankar Gangisetty, Avijit Dasgupta, and CV Jawahar. Drivesafe: A framework for risk detection and safety suggestions in driving scenarios. *arXiv preprint arXiv:2605.16892*, 2026.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14455–14465, June 2024.

Dianwei Chen, Zifan Zhang, Lei Cheng, Yuchen Liu, and Xianfeng Terry Yang. Insight: Enhancing autonomous driving safety through vision-language models on context-aware hazard detection and reasoning. In *2026 IEEE Intelligent Vehicles Symposium Workshops (IV Workshops)*, pp. 2190–2197. IEEE, 2026.

Tushar Choudhary, Vikrant Dewangan, Shivam Chandhok, Shubham Priyadarshan, Anushka Jain, Arun K Singh, Siddharth Srivastava, Krishna Murthy Jatavallabhula, and K Madhava Krishna. Talk2bev: Language-enhanced bird's-eye view maps for autonomous driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 16345–16352. IEEE, 2024.

Xinpeng Ding, Jianhua Han, Hang Xu, Wei Zhang, and Xiaomeng Li. Hilm-d: Enhancing mllms with multi-scale high-resolution details for autonomous driving. *International Journal of Computer Vision*, 133(8):5379–5395, 2025.

Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, He Wang, and Sen Li. Dada-2000: Can driving accident be predicted by driver attention*f* analyzed by a benchmark. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pp. 4303–4309. IEEE, 2019.

Mihir Godbole, Xiangbo Gao, and Zhengzhong Tu. Drama-x: A fine-grained intent prediction and risk reasoning benchmark for driving. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 826–831. IEEE, 2025.

Chi-Hsi Kung, Chieh-Chi Yang, Pang-Yuan Pao, Shu-Wei Lu, Pin-Lun Chen, Hsin-Cheng Lu, and Yi-Ting Chen. Riskbench: A scenario-based benchmark for risk identification. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14800–14807. IEEE, 2024.

Chengxi Li and Stanley H. Chan. Droid: Driver-centric risk object identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. Dolphins: Multimodal language model for driving. In *European Conference on Computer Vision*, pp. 403–420. Springer, 2024.

Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. Drama: Joint risk localization and captioning in driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1043–1052, 2023.

Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. In *European Conference on Computer Vision*, pp. 292–308. Springer, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL `https://qwen.ai/blog?id=qwen3.5`.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026b. URL `https://qwen.ai/blog?id=qwen3.6-35b-a3b`.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026c. URL `https://qwen.ai/blog?id=qwen3.8`.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, pp. 256–274. Springer, 2024.

Kexin Tian, Jingrui Mao, Yunlong Zhang, Jiwan Jiang, Yang Zhou, and Zhengzhong Tu. Nuscenes-spatialqa: A spatial understanding and reasoning benchmark for vision-language models in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 4626–4635, October 2025.

Junhua Wang, Wenhao Zhang, Ting Fu, and Qiangqiang Shangguan. Drpvlm: A generative multimodal large language model for real-time driving risk prediction. *Accident Analysis & Prevention*, 231:108505, 2026.

Yuqi Wang, Ke Cheng, Jiawei He, Qitai Wang, Hengchen Dai, Yuntao Chen, Fei Xia, and Zhaoxiang Zhang. Drivingdojo dataset: Advancing interactive and knowledge-enriched driving world model. In *NeurIPS*, 2024.

Licheng Wen, Xuemeng Yang, Daocheng Fu, Xiaofeng Wang, Pinlong Cai, Xin Li, Tao Ma, Yingxuan Li, Linran Xu, Dengke Shang, et al. On the road with gpt-4v (ision): Explorations of utilizing visual-language model as autonomous driving agent. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

A Appendix

A.1 Prompt for VLMs

For all VLM query, we use the basic prompt as follow:

You are an autonomous driving scenario safety assessment assistant.
Based on the input single in-vehicle view image, determine whether the current scene presents a potential driving safety risk.
Definition of risk:
1. Pedestrians, vehicles, cyclists, animals, obstacles, or other traffic participants may enter or affect the vehicle's path.
2. There is a risk of collision, crossing, cutting in, rear-end collision, obstruction, driving against traffic, road construction, or traffic signal conflict.

A.2 Ablation Experiment: The Effectiveness of Attention Pooling Strategy

Table 4: Scene-level hazard classification performance of the risk-scoring head under different permutation-invariant aggregation strategies.

	F1-score(%)	acc(%)	precision(%)	recall(%)
Avg Pooling	72.67	79.14	70.11	75.43
Max Pooling	73.85	70.75	73.77	73.93
Sum Pooling	74.26	79.75	72.63	75.96
Attention Pooling	76.38	81.34	75.39	77.39

To isolate the effect of the aggregation strategy on risk representation learning, we replace attention pooling with average, max, and sum pooling while keeping the node encoder, risk-scoring head, training data, and optimization settings unchanged. We report the scene-level hazard classification performance of the learned risk-scoring head $R(\mathcal{G})$ on the test set, before sub-VCG extraction and downstream VLM inference. As shown in Table 3, attention pooling achieves the best overall performance, suggesting that adaptively weighting traffic participants produces a more discriminative vehicle-centric risk representation than fixed permutation-invariant aggregation.

A.3 Ablation Experiment: The Effectiveness of the Lightweight VLM for Traffic participants extraction

Table 5: Final performance of Qwen3.6-35B-A3B guided by structured hazard cues extracted using different traffic participant extractors.

NPC Abstract Model	F1-score(%)	acc(%)	precision(%)	recall(%)
Qwen3.5-2B	83.79	87.15	82.49	85.12
Qwen3.5-4B	84.35	87.69	83.54	85.16
Qwen3.5-9B	81.20	82.66	70.28	96.12
Qwen3.6-35B-A3B	85.54	87.94	80.23	91.61
Qwen3.8-27B	81.63	84.17	74.47	90.32

To evaluate the impact of different NPC abstraction models, we employ models of varying sizes to complete the entire RiskRep pipeline, while fixing Qwen3.6-35B-A3B as the final VLM for hazard identification. As shown in Table 5, increasing the size of the NPC abstraction model does not consistently improve downstream performance. This suggests that the NPC abstraction stage does not strongly depend on model scale. Therefore, we adopt Qwen3.5-2B as the default abstraction model for its favorable trade-off between effectiveness and computational efficiency.