DriveSafe: A Framework for Risk Detection and Safety Suggestions in Driving Scenarios

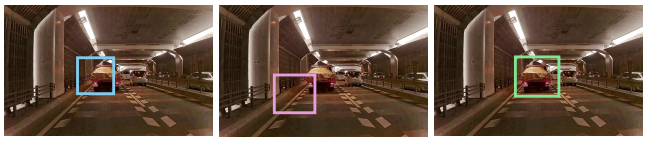
Sainithin Artham, Shankar Gangisetty, Avijit Dasgupta, C. V. Jawahar

***Abstract*— Comprehensive situational awareness is essential for autonomous vehicles operating in safety-critical environments, as it enables the identification and mitigation of potential risks. Although recent Multimodal Large Language Models (MLLMs) have shown promise on general vision–language tasks, our findings indicate that zero-shot MLLMs still underperform compared to domain-specific methods in fine-grained, spatially grounded risk assessment. To address this gap, we propose DriveSafe, a framework for risk-aware scene understanding that leverages structured natural language descriptions. Specifically, our method first generates spatially grounded captions enriched with multimodal context—including motion, spatial, and depth cues. These captions are then used for downstream risk assessment, explicitly identifying hazardous objects, their locations, and the unsafe behaviors they imply, followed by actionable safety suggestions. To further improve performance, we employ caption–risk pairings to fine-tune a lightweight adapter module, efficiently injecting domain-specific knowledge into the base LLM. By conditioning risk assessment on explicit language-based scene representations, DriveSafe achieves significant gains over both zero-shot MLLMs and prior domain-specific baselines. Exhaustive experiments on the DRAMA benchmark demonstrate state-of-the-art performance, while ablation studies validate the effectiveness of our key design choices. Project page: https://cvit.iiit.ac.in/research/projects/cvit-projects/drivesafe.**

I. Introduction

Risk assessment and safety prediction are central to many safety-critical domains such as aviation [1], healthcare [2], and robotics [3], where anticipating hazards is essential for preventing catastrophic failures. Road transportation presents a similarly critical challenge: traffic accidents remain one of the leading causes of mortality worldwide, with an estimated 1.2–1.35 million deaths annually and tens of millions of serious injuries [4]. In the United States alone, motor vehicle crashes claimed around $43,000$ lives in 2023, corresponding to a fatality rate of about 12.2 deaths per $100,000$ people [5]. These sobering statistics highlight the urgent need for rigorous risk assessment and proactive safety interventions across all forms of driving. With the growing adoption of autonomous vehicles, this demand becomes even more pressing, as self-driving systems must not only perceive and plan effectively but also reason under uncertainty, anticipate hazards, and provide reliable safety suggestions in complex, real-world environments.

Existing works have approached this challenge from multiple perspectives: attention-based forecasting models such as RAIN [6] emphasize risk-aware trajectory prediction by highlighting salient agents, reinforcement learning frameworks with latent state inference [7] aim to capture hidden risk cues for decision-making, and interaction graph models [8] are designed to represent spatio-temporal dependencies among on-road agents. While these approaches advance risk prediction, they do not address the problem of generating *actionable safety suggestions*, which are crucial for practical deployment.



Q: Which object presents a potential risk to the ego-vehicle's trajectory?

	Existing Models	MLLMs	DriveSafe (Ours)
Risk (Yes/No)	Yes	Yes	Yes
Risky Object Localization	Partial	✗	✓
Risk-related Keyword	Truck slowing	Truck slowing	✓ Yellow truck slowing down
Safety Suggestion	✗	✗ Follow the vehicle ahead	✓ Slow down

Fig. 1: Previous works in driving scenarios [9], [10], [11] primarily address risk perception but fall short of offering actionable safety guidance. Similarly, general-purpose MLLMs [12], [13], [14] are still unreliable in this regard. In contrast, our approach, *DriveSafe*, integrates risk assessment with clear, human-understandable safety suggestions.

The curation of large-scale datasets such as DRAMA [9] and Rank2Tell [15] has further accelerated research in this direction. While a few works [11], [16] have utilized these datasets, their primary focus has not been on advancing risk perception or prediction. As no existing dataset explicitly addresses safety suggestions, we extend the DRAMA [9] dataset by explicitly associating critical objects with risk-related behavioral keywords (see Section IV-A for details).

On the other hand, MLLMs have recently shown strong performance in general video understanding tasks such as visual question answering [17] and image captioning [18]. General-purpose models like Qwen2.5-VL [12], Video-LLaMA3 [14], and LLaVA-Next [13] demonstrate impressive capabilities in generating natural language descriptions of visual scenes. However, these captions are typically generic and descriptive in nature, focusing on objects, activities, and context, rather than identifying potential hazards or issuing safety-related guidance. As a result, current MLLMs fall short in producing risk-aware or safety-critical captions that can inform proactive interventions in driving scenarios.

IIIT-Hyderabad, India sainithin.artham@gmail.com, {shankar.gangisetty@ihub-data., avijit.dasgupta@research., jawahar@}iiit.ac.in

Bridging this gap requires models that explicitly align visual understanding with safety reasoning, moving beyond surface-level descriptions toward actionable safety suggestions.

In this work, we take an alternative approach to risk assessment by avoiding direct fine-tuning of a MLLM. We argue that generic vision encoders as in MLLMs, while effective for broad video understanding tasks, are insufficient for safety-critical driving scenarios where subtle motion cues, spatial context, and depth perception play a decisive role. To address this, our method, *DriveSafe*, leverages multimodal contextual signals—including optical flow, road and lane segmentation, and depth maps—together with video descriptions to generate driving-specific scene captions. These captions are then provided to a large language model (LLM), which is prompted to produce risk assessment outputs. Crucially, unlike prior approaches that focus solely on risk detection, DriveSafe explicitly generates actionable safety suggestions that are grounded in the detected hazards (e.g., "*Slow down*") (refer Fig. 1).

In summary, our key contributions are as follows:

1) We propose *DriveSafe*, a novel pipeline for driving risk assessment that avoids direct fine-tuning of MLLMs. Instead, it integrates multimodal contextual cues (optical flow, lane/road segmentation, depth) with video descriptions to produce driving-specific captions, which are then processed by an LLM for risk prediction and safety suggestion.
2) We extend the DRAMA dataset by associating risky behaviors with critical objects, introducing safety-suggestion annotations. To ensure scalability and reproducibility, we annotate the training set automatically with an LLM while reserving manual annotations for the test set.
3) Extensive experiments demonstrate that DriveSafe outperforms prior baselines in both risk assessment and safety suggestion quality, while maintaining strong generalization across diverse driving scenarios.

II. Related Works

A. Risk Reasoning and Safety in Driving.

In the driving domain, numerous works have explored identifying risks that directly influence decision-making and overall safety [16], [11], [19], [20]. Goal-oriented importance estimation in on-road videos [21] and deep spatio-temporal importance prediction [22] focus on highlighting objects that affect vehicle navigation, while interaction-graph approaches [23] explicitly model agent–agent relations to infer critical elements for future behavior. Advancing toward risk-aware reasoning, causal inference methods have been introduced to identify objects influencing driver actions, such as those that lead to stops [24], and adaptive situational awareness systems [25] aim to prioritize scene elements in complex urban contexts. More recently, efforts such as joint risk localization and captioning [9] attempt to both detect hazardous objects and generate textual explanations of their impact on the ego-vehicle, while others explore risk assessment based on driving behaviors that violate safety commonsense [26].

System: You are an expert driving scene summarizer. Analyze multimodal driving data and produce a concise, geometry-aware summary.

User: Given the multimodal inputs for a driving video:
Spatial context ($\mathcal{S}_t$): {example}
Motion dynamics ($\mathcal{M}_t$): {example}
Depth context ($\mathcal{D}_t$): {example}
Video description (d_v): {example} Generate a structured natural language summary that includes:

1) Ego vehicle maneuvers (*lane changes*, *turns*, stability)
2) Behavior of surrounding agents (*approaching*, *overtaking*, *falling behind*)
3) Lane position and road context
4) Bounding boxes, depth, and motion cues for key objects

The output should be a clear paragraph suitable for downstream reasoning tasks.

Fig. 2: Prompt template P used to guide the LLM in generating structured, spatially grounded summaries of driving scenes.

B. MLLMs in Autonomous Driving.

MLLMs have recently garnered significant interest for their ability to analyze non-textual modalities, such as images and point clouds, through language-based reasoning [27], [28], [29], [30]. Leveraging their flexibility, MLLMs have been applied to various risk-related driving tasks. For instance, HiLM-D [31] utilizes hybrid-resolution perception to jointly detect safety-critical objects and predict their intentions, thereby improving risk localization. Likewise, MLLM-SUL [32] introduces a dual-branch visual encoding framework integrated with LLaMA-based reasoning to perform semantic risk inference and hazardous agent localization, building upon DRAMA-ROLISP [31]. Extending beyond single-vehicle contexts, V2V-LLM [33] advances risk reasoning into cooperative multi-vehicle environments via shared perceptual fusion, enabling collaborative assessment across multiple agents.

However, despite these advancements, no existing work has explicitly addressed safety as an integral objective, leaving a crucial gap in bridging risk understanding with actionable safety guidance. Motivated by this gap, our work seeks to integrate both risk reasoning and safety assessment within the MLLM paradigm.

III. DriveSafe Framework

Given a video sequence v our framework predicts a tuple: $(\hat{r}, C_r, b_r, s_r) = \mathcal{F}(v)$, where $\hat{r}$ denotes *yes/no* for the risk, C_r is the generated risk caption, b_r is the bounding box of the visually grounded risky object, and s_r is the associated safety

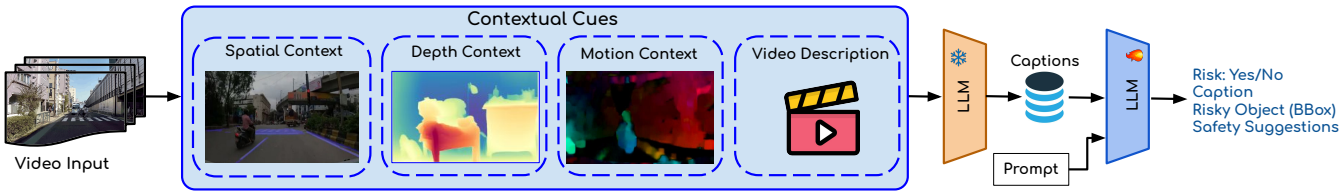


Fig. 3: Our proposed DriveSafe framework for the caption generation and safety suggestion task in driving. We first derive contextual cues to guide caption generation, and then use the resulting captions for risk assessment and safety suggestion.

suggestion for the risky behavior. The proposed framework is structured in two stages:

- **Caption Generation** (Section III-A): Our framework generates scene-level captions enriched with multimodal contextual cues and geometric awareness of the driving environment.
- **Risk Assessment and Safety Suggestion** (Section III-B): The generated captions are processed with an LLM to infer a binary risk classification $\hat{r}$, a risk caption C_r, risk-related keywords $\hat{K}$, and bounding boxes of risky objects $\hat{b}$. Safety suggestions are then derived from the predicted keywords $\hat{K}$. While the framework can operate in a zero-shot setting, we additionally fine-tune the LLM using lightweight adapters to enhance both performance and robustness.

A. Caption Generation

Given a video sequence $v = \{f_1, f_2, \ldots, f_T\}$, where each frame $f_t \in \mathbb{R}^{H\times W\times 3}$ and T denotes the number of frames with spatial resolution (H, W), we extract three types of contextual information. The spatial context $\mathcal{S}_t$ is obtained using road and lane segmentation masks that delineate boundaries and drivable areas. The motion context $\mathcal{M}_t$ is derived from optical flow, highlighting regions of relative movement corresponding to surrounding agents such as vehicles, cyclists, and pedestrians. Finally, the depth context $\mathcal{D}_t$ is computed at the object level, ensuring that each detected object in the surrounding (e.g., vehicles, pedestrians, traffic lights, and barriers) is associated with a depth estimate that reflects its spatial proximity to the ego-vehicle.

While $\mathcal{S}_t$, $\mathcal{M}_t$, and $\mathcal{D}_t$ capture frame-level spatial, motion, and depth cues, we further enrich the representation with video-level semantics. Specifically, a high-level description d_v is generated using a MLLM, summarizing the global scene context, such as "*a car is slowing down in front*" or "*a pedestrian is crossing from the right*". This global description complements the fine-grained frame-level cues by providing a holistic understanding of driving scenario.

The complete multimodal representation of a video is defined as:

$$X_v = \{\{\mathcal{S}_t, \mathcal{M}_t, \mathcal{D}_t\}_{t=1}^{T}, d_v\}. \tag{1}$$

where frame-level spatial, motion, and depth contexts are complemented by the video-level semantic description. To leverage this representation, we construct a structured prompt $P(X_v)$ (illustrated in Fig. 2) that integrates both frame-level and global information into a unified input for the LLM F_θ. The model then fuses these modalities to produce a geometry-aware description of the video:

$$C_v = F_\theta(P(X_v)), \tag{2}$$

where C_v denotes the generated caption for the sequence.

System: You are a risk-aware driving scene analyst. Given a caption C_v with object tags and bounding boxes, analyze risks, explain them, and suggest safety-aware actions.

Inputs: Caption (C_v)
Outputs (per object):
1) Risk label $\hat{r} \in \{Yes, No\}$
2) Risk caption (C_r)
3) Risk-related keywords ($\hat{K}$)
4) Object Localization - Bounding box ($\hat{b}$)

Reasoning should be accurate, structured, and safety-focused.

Example Input C_v**:** "A cyclist [bbox: 612, 350, 720, 480] is crossing from the left; a red car [bbox: 1000, 400, 1200, 550] is stopped in the ego lane."
Example Output:
1) $\hat{r}$ =Yes; C_r: cyclist crossing may intersect ego path; $\hat{K}$ ={Cyclist, Crossing}; $\hat{b}$ =[612, 350, 720, 480].
2) $\hat{r}$ =Yes; C_r: red car stopped in ego lane blocks motion; $\hat{K}$ ={Stopped vehicle}; $\hat{b}$ =[1000, 400, 1200, 550].

Fig. 4: Risk-aware prompt template consistent with notation. The LLM F_θ maps C_v to $(\hat{r}, C_r, \hat{K}, \hat{b})$.

B. Risk Assessment and Safety Suggestion

Zero-Shot: Given the geometric-aware caption C_v, we prompt the LLM F_θ using a structured template as illustrated in Fig. 4 to produce risk-aware outputs. Specifically, the model generates:

- *Risk label:* a binary indicator of whether the scene involves risk,
- *Refined risk caption:* a descriptive explanation specifying the risky object and its spatial location,
- *Risk-associated keywords:* salient terms that capture the risky behavior or objects involved, and
- *Bounding box:* the localized coordinates of the identified risky object.

Formally, this process is defined as:

$$(\hat{r}, \hat{C}_r, \hat{K}, \hat{b}) = f_\theta(C_v), \tag{3}$$

where $\hat{r} \in \{Yes, No\}$ denotes the binary risk classification, $\hat{C}_r$ is the generated risk caption, $\hat{K}$ is the set of extracted

Safety Suggestions	Risk-related Keywords
(Must) Stop	Pedestrian crossing (19); Stopped vehicle (860); Crosswalk (105); Traffic light red (751); Traffic light yellow (5); Traffic congestion (877)
Be aware / cautious (object may affect future but no direct influence)	Cyclist nearby (8); Pedestrian nearby (12); Traffic signal (19); Traffic sign (62); Leading vehicle (151)
Slow down	Slowing (277); Pedestrian ahead (4); Heavy traffic (159); Cut-in (8); Cyclist (12)
Carefully maneuver (around important object)	Parked vehicle (28); Traffic cones (9)
Follow the vehicle ahead	Vehicle in front (234); Following traffic (44); Same lane (228); Near the intersection (77)
Yield	Merging traffic (7); Vulnerable Road User (VRU) (37); Right of way (123); Oncoming traffic (109); At the crosswalk (86);
Start moving	Traffic cleared (18); Vehicle ahead moved (36); Traffic light green (10)
NA	Irrelevant (5); Background (5); No decision (5)

TABLE I: Mapping between safety suggestions and their corresponding risk-related keywords in the DRAMA [9] dataset. Instances of keywords are shown in parentheses.

risk-related keywords, and $\hat{b} = (\hat{x}_{\min}, \hat{y}_{\min}, \hat{x}_{\max}, \hat{y}_{\max})$ represents the predicted bounding box of the risky object.

To derive *safety suggestions*, the extracted keywords $\hat{K}$ are mapped to corresponding driving action categories using a predefined rule set (see Table I). The final safety suggestion s_r is then obtained through keyword matching with ground-truth categories

$$s_r = g(\hat{K}), \tag{4}$$

where $g(\cdot)$ denotes the mapping function from risk keywords to safety suggestions.

Adapter Finetuning: In the fine-tuning stage, we adapt the LLM F_θ using lightweight parameter-efficient adapters trained on a dataset of caption–instruction–response triplets. Given an input caption c_v and instruction q, the model is supervised to generate a structured response $\hat{a} = (\hat{r}, \hat{C}_r, \hat{K}, \hat{\mathbf{b}})$. Formally, the output is defined as:

$$\hat{a} = F_\theta(C_v, q), \tag{5}$$

where $\hat{a}$ represents the predicted risk reasoning sequence.

The training objective minimizing the negative log-likelihood of the ground-truth structured response:

$$\mathcal{L} = -\sum_{t=1}^{T} \log p_\theta(a_t \mid C_v, q, a_{<t}), \tag{6}$$

where a_t is the t^{th} token of the response and $a_{<t}$ are the previously generated tokens.

This adapter-based fine-tuning explicitly aligns video captions with risk classification $\hat{r}$, descriptive risk captions C_r, risk-related keywords $\hat{K}$, and bounding boxes $\hat{\mathbf{b}}$. Safety suggestions are then derived directly from the generated risk keywords $\hat{K}$ via string matching to predefined categories. Model performance is evaluated using accuracy and F_1 scores over nine safety suggestion classes (see Table I), thereby providing a direct measure of both correctness and reliability.

IV. Experiments

A. Dataset

We use DRAMA dataset [9] which consists of 17K interactive driving scenario videos with rich annotations, including driving risk-related captions, nine categories of safety suggestions, and bounding boxes for important objects to enable visual grounding. To evaluate safety suggestions in the context of risky behaviors, we re-organize the DRAMA dataset by explicitly linking safety suggestions with the corresponding risk-related keywords derived from risky behaviors described in the captions, as illustrated in Table I. The test set of DRAMA dataset is manually reviewed by annotators, and mapped the nine predefined safety suggestions one-to-one with risk-related keywords by carefully observing videos and their captions. This re-annotation enables a more reliable safety evaluation. The distribution of safety-critical cues and their frequency of occurrence are shown in Fig. 5. At the dataset level, the five most frequent risk-related keywords were *traffic congestion*, *stopped vehicle*, *traffic light red*, *slowing*, and *vehicle in front*.

Method	B1↑	B4↑	M↑	R↑	C↑	S↑	CLAIR	Mean IoU	Acc@ 0.5
LCP [9]	73.9	54.7	39.1	70.0	**3.7**	56.0	–	61.4	68.4
VTS [16]	75.3	55.8	40.7	74.7	2.8	58.0	–	66.8	74.4
LLaVA-v1.5† [18]	75.8	56.1	41.0	78.0	2.9	58.4	–	–	–
Efficient HoP [11]	76.0	56.2	41.3	78.5	2.7	58.8	–	–	–
HoP [11]	**76.2**	56.3	41.7	79.8	2.87	59.1	–	–	–
Qwen2.5-VL [12]	27.72	4.94	18.72	26.15	0.08	16.30	24.33	0.0	0.0
LLaVA-NeXT [13]	28.17	4.30	18.87	26.60	0.06	16.76	23.22	**83.6**	13.1
VideoLLaMA3 [14]	26.73	3.24	21.36	23.70	0.07	11.91	24.55	3.9	4.3
DriveSafe-Zeroshot	30.65	10.55	33.92	35.70	0.12	22.11	30.47	82.4	0.9
DriveSafe-Finetuned	64.47	**60.38**	**64.78**	**80.85**	3.27	**64.91**	**58.93**	59.8	**74.8**

TABLE II: Performance comparison of **Caption Generation** and **Risky Object Grounding** across Existing Methods, General VLMs, and DriveSafe on the DRAMA [9] dataset.

Model	Accuracy	F_1 (Weighted)
LLaVA-Next [13]	15.83	20.95
VideoLLaMA 3 [14]	13.00	19.55
Qwen2.5 VL [12]	18.88	23.19
DriveSafe - Zeroshot	23.49	24.80
DriveSafe - Finetuned	**52.85**	**37.15**

TABLE III: Performance comparison of **Safety Suggestion** prediction across General-VLMs and DriveSafe on the DRAMA [9].

B. Experimental Settings

Models. We compare our framework with several video-based MLLMs. We consider risk assessment models such

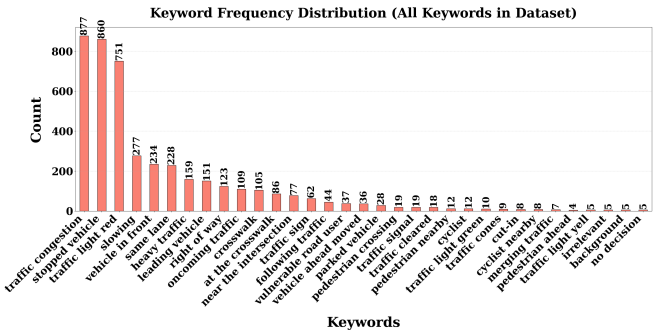


Fig. 5: Distribution of driving decision categories in the curated test set. Each bar corresponds to a safety suggestion category, showing the aggregated frequency of its representative keywords.

as LCP [9], VTS [10], LLaVA-v1.5 [18] and HoP [11]. In addition, we also compare with several open source MLLMs such as Qwen2.5 [12], LLaVA-NeXT [13], and VideoLLaMA 3 [14] to assess their performance on risk prediction, risky object localization and safety suggestion prediction.

Metrics. We evaluate caption generation using standard metrics, including BLEU-1 (B1) [34], BLEU-4 (B4) [34], METEOR (M) [35], ROUGE-L (R) [36], CIDEr (C) [37], and SPICE (S) [38]. In addition, we incorporate CLAIR [39], a recently proposed LLM-based evaluation metric designed to better capture semantic quality beyond n-gram overlap. For localization, we report Intersection-over-Union (IoU) as the evaluation metric.

Safety suggestions are generated using the ground-truth rule map that links risk-related keywords to suggestion categories (Section III-A). Following DRAMA [9], we consider eight safety suggestion classes (excluding NA). Accuracy and F_1 score are reported as the primary metrics for evaluating suggestion quality.

Implementation Details. For spatial context, we use HybridNets [40], motion cues are obtained by computing dense optical flow using the Farnebäck algorithm from OpenCV, and depth cues are estimated using DepthAnything-v2 [41]. For fine-tuning, we adopt the LLaMA-Adapter framework [42]. Training is performed with a batch size of 4, learning rate of 2×10^{-5}, and weight decay of 0.01 for 5 epochs on NVIDIA A6000 (40GB) GPU. To improve efficiency, we apply gradient checkpointing and mixed precision, along with a linear warmup schedule over the first 10% of training steps. For zero-shot inference, we employ LLaMA-3.1-8B [43], which directly conditions on captions without additional finetuning.

C. Quantitative results

Caption Generation. Table II presents the quantitative comparison across captioning metrics (BLEU [34], METEOR [35], ROUGE [36], CIDEr [37], SPICE [38] and CLAIR [39]). DriveSafe-Finetuned consistently outperforms its zero-shot counterpart across all metrics, with BLEU-4 improving from 10.55 to 60.38, METEOR from 33.92 to 64.78, and ROUGE from 35.70 to 80.85. These gains highlight the effectiveness of our caption-to-risk assessment adapter-based instruction tuning in capturing safety-critical driving narratives.

Compared to specialized driving models such as HoP [11] (BLEU-4: 56.3, METEOR: 41.7) and LLaVA-v1.5 [18] (BLEU-4: 56.1, METEOR: 41.0), DriveSafe-Finetuned achieves superior performance despite relying only on lightweight adapter tuning rather than full-scale video pretraining. In contrast, general-purpose VLMs show severe degradation on driving scenarios, with Qwen2.5-VL [12], LLaVA-NeXT [13], and VideoLLaMA-3 [14] achieving BLEU-4 scores below 30 and METEOR scores below 22. This clearly underscores the challenges of directly adapting general VLMs to safety-critical domains, which our domain-specific finetuning strategy effectively bridges this gap.

Risky Object Grounding. For object grounding, we evaluate performance using MeanIoU and Acc@0.5. General-purpose VLMs perform poorly in this setting, with Qwen2.5-VL [12] and VideoLLaMA-3 [14] nearly fail, while LLaVA-NeXT [13] achieves only limited accuracy (MeanIoU 83.6, Acc@0.5 13.1). In contrast, DriveSafe-Finetuned shows strong localization ability, reaching an Acc@0.5 of 74.8, far surpassing all other models. This demonstrates its effectiveness in linking descriptive captions with precise spatial grounding, an ability that is essential for downstream safety reasoning.

To further analyze grounding reliability, we evaluated accuracy and IoU across multiple thresholds. In the zero-shot setting, accuracy remains near-zero (e.g., 3.4% at 0.1, dropping to 1.0% at 0.4), while IoU appears deceptively high (rising from 35.8% at 0.1 to 80.2% at 0.4). This discrepancy arises from a few rare correct matches that inflate IoU scores, despite the model failing to capture most risky agents. In contrast, the finetuned model achieves both high accuracy and strong IoU jointly, i.e., 91.0% / 90.0% at 0.1, 83.0% / 81.0% at 0.3, and 79.0% / 66.0% at 0.4.

Overall, these results highlight that IoU alone is an unreliable measure for safety-critical grounding, as zero-shot models can achieve seemingly high IoU while missing most risks. The consistent combination of high accuracy and IoU achieved by DriveSafe-Finetuned underscores the necessity of task-specific adaptation for reliable safety reasoning.

Safety Suggestion. Table III presents the accuracy and weighted F_1 scores for the safety suggestion task. General-purpose VLMs such as LLaVA-Next, VideoLLaMA-3, and Qwen2.5-VL perform poorly, with accuracies in the range of 13–19% and weighted F_1 scores below 24%. These results indicate that open-domain VLMs lack the fine-grained risk understanding necessary for safety-critical driving scenarios. In contrast, our proposed DriveSafe model shows clear improvements. In the zeroshot setting, DriveSafe already surpasses all general VLM baselines, achieving 23.49% accuracy and a 24.80 weighted F_1, demonstrating the benefit of structured multimodal prompting even without finetuning. After finetuning, DriveSafe achieves 52.85% accuracy and 37.15 weighted F_1, more than doubling its zeroshot performance and substantially outperforming all baselines.

Fig. 6: Qualitative comparison of DriveSafe-ZeroShot, DriveSafe-Finetuned, and Qwen2.5-VL [12] on three driving scenarios from the DRAMA dataset [9]. Risky object grounding is shown with bounding boxes so is respective models with text highlighting, while generated captions and safety suggestions are marked as correct (green) or incorrect (red).

On an NVIDIA A6000 GPU, LLaMA-Adapter 3.1 (8B) achieves a per-token latency of approximately 7–11 ms (∼90–140 tokens/s), demonstrating suitability for near real-time deployment.

D. Qualitative Analysis

As shown in Fig. 6, we observe in the first column that Qwen2.5-VL [12] misidentifies risks, for example describing "*a pedestrian is crossing*" instead of a cyclist, shifting attention away from the true hazard. DriveSafe-ZeroShot also adds false context, e.g., "*a cyclist is passing by the pedestrian crossing*". In contrast, DriveSafe-Finetuned provides grounded interpretations such as "*the truck is driving at reduced speed on a city street*" with the correct suggestion "*Slow down*". In the second column, DriveSafe-ZeroShot exaggerates motion risk (e.g., "*a car moving at a different velocity in front of the truck*"), whereas DriveSafe-Finetuned correctly identifies the "*black sedan stopped in the ego lane*". In the third column, Qwen2.5-VL misses the critical ego-lane sedan, while DriveSafe-ZeroShot gives only a vague suggestion. DriveSafe-Finetuned instead highlights the collision risk and provides the precise recommendation "*Be aware / cautious*". These examples show that fine-tuning reduces hallucinations and exaggeration while improving focus on safety-critical evidence.

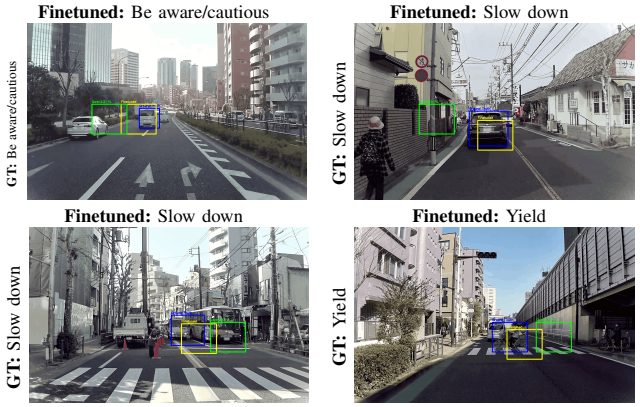


Fig. 7: Qualitative comparison of **Safety Suggestions** across Ground-truth, Qwen2.5-VL and DriveSafe-Finetuned in different challenging driving scenarios.

Fig. 7 highlights two common error types in Qwen2.5-VL and how DriveSafe-Finetuned overcomes them. The first is missing critical agents, where Qwen2.5-VL overlooks cyclists or pedestrians and generates irrelevant advice (e.g., "*follow the vehicle ahead*"), while DriveSafe provides accurate suggestions such as "*Slow down*" or "*Yield*", consistent with the ground-truth. The second is overcautious responses, where Qwen2.5-VL issues premature guidance (e.g., "*slow*

down" when the leading vehicle is distant). DriveSafe instead produces balanced, context-aware advice (e.g., "*Be aware / cautious*"). These cases show that DriveSafe reduces ambiguity and delivers grounded, norm-consistent safety suggestions.

E. Ablation Experiments

Effects of Model Selection for Pseudo-labeling. We investigate the impact of different backbone models for generating pseudo-labels. For VLMs, pseudo-labels are derived using both the video and its ground-truth caption as input, while for LLMs, only the caption is provided. Table IV reports performance on the safety suggestion prediction task. Among the compared models, LLaMA 3.1 [43] achieves the best F_1 score (55.9), outperforming all baselines. In contrast, DriveSafe-Finetuned attains a higher accuracy (52.85) but a substantially lower F_1 (37.15), suggesting a tendency toward majority-class predictions.

Pseudo-labeling Model	Accuracy	F_1
LLaVA-NeXT [13]	33.5	31.6
Qwen2.5-VL [12]	34.4	30.1
DeepSeek [44]	44.2	42.8
LLaMA-3.1-8B [43]	47.6	**55.9**
DriveSafe-Finetuned	**52.85**	37.15

TABLE IV: Comparison of different models used for pseudo-label assignment, evaluated on safety suggestion prediction.

Effects of Contextual Cues. Table V presents a component-level ablation study on the DRAMA [9] dataset, where we progressively introduce three contextual cues. The baseline system, without any contextual cues, yields relatively low scores. Incorporating either spatial context $\mathcal{S}_t$ with depth $\mathcal{D}_t$, or motion $\mathcal{M}_t$ with depth $\mathcal{D}_t$, results in moderate gains METEOR [35] (+21.4%) and (+27.7%) respectively, highlighting the individual contributions of spatial and motion context. Combining both spatial $\mathcal{S}_t$ and motion $\mathcal{M}_t$ leads to larger gains (METEOR [35]: +41.03%, CLAIR [39]: 8.88%). Finally, the full model with all three contexts achieves the best results (METEOR [35]: 33.92, CLAIR [39]: 30.47), marking an 81.2% improvement on METEOR and 43.7% on CLAIR over the baseline.

$\mathcal{M}_t$	$\mathcal{S}_t$	$\mathcal{D}_t$	METEOR	CLAIR
✗	✗	✗	18.72	21.19
✗	✓	✓	22.73 (↑21.42%)	24.46 (↑15.45%)
✓	✗	✓	23.90 (↑27.67%)	21.78 (↑2.81%)
✓	✓	✗	26.40 (↑41.03%)	23.06 (↑8.88%)
✓	✓	✓	**33.92** (↑81.20%)	**30.47** (↑43.75%)

TABLE V: Component-level ablation on DRAMA [9] dataset. Relative improvements are computed w.r.t. the baseline (first row).

Backbone Comparison. Table VI presents an ablation study comparing different VLM–LLM backbones within DriveSafe. In the VLM-wise comparison (LLM fixed to LLaMA-3.1 [43]), Qwen2.5-VL [12] achieves approximately 40% higher METEOR and 30% higher CLAIR scores than LLaVA-NeXT [13], underscoring its stronger temporal grounding and video understanding. In the LLM-wise comparison (VLM fixed to Qwen2.5-VL), LLaMA-3.1 provides consistently strong alignment, while DeepSeek [44] yields a 55% lower METEOR but a 13% higher CLAIR score. Since CLAIR [39] relies on DeepSeek as its evaluator, this introduces a mild bias favoring its outputs. Overall, these results confirm that both VLM and LLM choices substantially influence DriveSafe's performance, and that the optimal configuration requires balancing fine-grained caption accuracy with semantic risk-awareness.

Model	Params	METEOR↑	CLAIR↑
VLM-wise comparison (LLaMA-3.1 [43] fixed)			
LLaVA-Next Video [13]	7B	23.64	23.22
Qwen2.5 VL [12]	7B	**33.92**	30.47
LLM-wise comparison (Qwen2.5-VL [12] fixed)			
DeepSeek [44]	7B	14.66	34.46
LLaMA-3.1 [43]	8B	33.92	30.47

TABLE VI: DriveSafe model performance with different VLMs and LLMs.

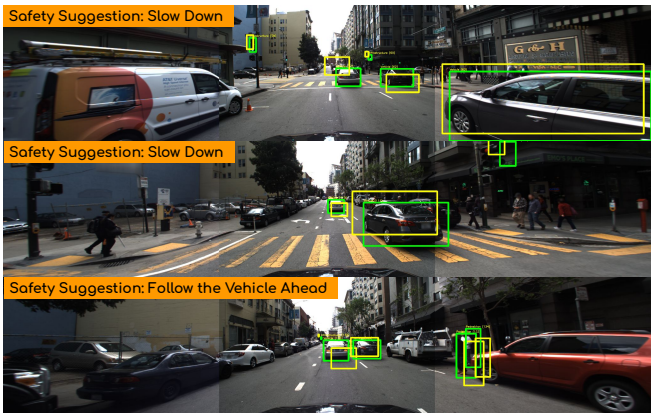


Fig. 8: Top-to-bottom sequence with DriveSafe-Finetuned and Ground-Truth predictions; safety suggestions appear top-left.

F. Application of DriveSafe

We evaluate our model's end-to-end ability to identify and localize multiple risky objects and generate safety suggestions in complex driving environments. For this evaluation, we use the Rank2Tell [15] dataset, which is ideal due to its multi-object annotations across entire video sequences. DriveSafe generates multimodal contextual cues (spatial, depth, motion), converts them into captions, and uses these for downstream risk identification. Figure 8 illustrates DriveSafe's performance across sequential frames. The model consistently localizes a potentially risky vehicle with bounding boxes and provides context-aware safety suggestions that evolve with the scene: it recommends "*Slow down*" as the car approaches the intersection (top and middle frames), and updates to "*Follow the vehicle ahead*" once the situation stabilizes (bottom frame).

V. Conclusions

This work introduced DriveSafe, a caption-based framework that enhances risk assessment in autonomous driving scenarios. Our evaluations demonstrate significant improvements over zero-shot MLLM baselines in risk assessment tasks, with fine-tuning substantially reducing hallucinations while improving the accurate identification and precise localization of risky driving behaviors. Beyond risk assessment, DriveSafe excels in safety applications by generating grounded safety recommendations that are directly linked to their underlying risk-inducing behaviors. This explicit connection between identified risks and corresponding safety suggestions addresses critical gaps in existing methods and provides the transparency necessary for building trust in autonomous systems. While the current risk-to-safety mapping is static, serving as a foundational first step to this novel framework. Future work will focus on learning dynamic mappings to improve scalability and adaptability, along with advanced environmental scaling, long-horizon temporal reasoning, and human-aligned explanation evaluation.

Acknowledgment. This project was supported by iHub-Data and Mobility at IIIT Hyderabad.

References

[1] I. Sikora, "Risk assessment, modelling and proactive safety management system in aviation: a literature review," in *Transportation Systems with International Participation*, 2015.

[2] Y. Voskanyan, I. Shikina, F. Kidalov, D. Davidov, and T. Abrosimova, "Risk management in the healthcare safety management system," *Journal of Digital Science*, 2021.

[3] F. Vicentini, M. Askarpour, M. G. Rossi, and D. Mandrioli, "Safety assessment of collaborative robotics through automated formal verification," *IEEE Transactions on Robotics*, 2019.

[4] World Health Organization, "Road traffic injuries fact sheet," 2024.

[5] Insurance Institute for Highway Safety, "Fatality statistics: State-by-state," 2023.

[6] J. Li, F. Yang, H. Ma, S. Malla, M. Tomizuka, and C. Choi, "Rain: Reinforced hybrid attention inference network for motion forecasting," in *ICCV*, 2021.

[7] X. Ma, J. Li, M. J. Kochenderfer, D. Isele, and K. Fujimura, "Reinforcement learning for autonomous driving with latent state inference and spatial-temporal relationships," in *ICRA*, 2021.

[8] Z. Zhang, A. Tawari, S. Martin, and D. Crandall, "Interaction graphs for object importance estimation in on-road driving videos," in *ICRA*, 2020.

[9] S. Malla, C. Choi, I. Dwivedi, J. H. Choi, and J. Li, "Drama: Joint risk localization and captioning in driving," in *WACV*, 2023.

[10] D. Bolya, C.-Y. Fu, X. Dai, P. Zhang, C. Feichtenhofer, and J. Hoffman, "Token merging: Your vit but faster," *arXiv*, 2022.

[11] H. Zhou, Z. Gao, M. Ye, Z. Chen, Q. Chen, T. Cao, and H. Qi, "Hints of prompt: Enhancing visual representation for multimodal llms in autonomous driving," *arXiv*, 2024.

[12] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, "Qwen2. 5-vl technical report," *arXiv*, 2025.

[13] F. Li, R. Zhang, H. Zhang, Y. Zhang, B. Li, W. Li, Z. Ma, and C. Li, "Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models," *arXiv*, 2024.

[14] B. Zhang, K. Li, Z. Cheng, Z. Hu, Y. Yuan, G. Chen, S. Leng, Y. Jiang, H. Zhang, X. Li *et al.*, "Videollama 3: Frontier multimodal foundation models for image and video understanding," *arXiv*, 2025.

[15] E. Sachdeva, N. Agarwal, S. Chundi, S. Roelofs, J. Li, M. Kochenderfer, C. Choi, and B. Dariush, "Rank2tell: A multimodal driving dataset for joint importance ranking and reasoning," in *WACV*, 2024.

[16] Y. Ma, A. Abdelraouf, R. Gupta, Z. Wang, and K. Han, "Video token sparsification for efficient multimodal llms in autonomous driving," *arXiv*, 2024.

[17] C. Parikh, D. Rawat, R. R. T., T. Ghosh, and R. K. Sarvadevabhatla, "Roadsocial: A diverse videoqa dataset and benchmark for road event understanding from social video narratives," *CVPR*, 2025.

[18] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *CVPR*, 2024.

[19] D. Wang, W. Fu, Q. Song, and J. Zhou, "Potential risk assessment for safe driving of autonomous vehicles under occluded vision," *Scientific Reports*, 2022.

[20] M. Aslantas, F. K. Gündogdu, and S. Moslem, "Evaluating the potential risks posed by autonomous vehicles by using a decomposed fuzzy multi-criteria decision-making model," *Transportation Engineering*, 2025.

[21] M. Gao, A. Tawari, and S. Martin, "Goal-oriented object importance estimation in on-road driving videos," in *ICRA*, 2019.

[22] E. Ohn-Bar and M. M. Trivedi, "Are all objects equal? deep spatio-temporal importance prediction in driving videos," *Pattern Recognition*, 2017.

[23] Z. Zhang, A. Tawari, S. Martin, and D. J. Crandall, "Interaction graphs for object importance estimation in on-road driving videos," *ICRA*, 2020.

[24] C. Li, S. H. Chan, and Y.-T. Chen, "Who make drivers stop? towards driver-centric risk assessment: Risk object identification via causal inference," *IROS*, 2020.

[25] T. Wu, E. Sachdeva, K. Akash, X. Wu, T. Misu, and J. Ortiz, "Toward an adaptive situational awareness support system for urban driving," *IV Symposium*, 2022.

[26] Z. Pang, Z. Chen, J. Lu, B. Sun, T. Gong, X. Feng, Y. Wang, S. Yang, and Y. Cao, "Risk assessment method for autonomous vehicles violating safety common sense based on driving behavior," *IEEE Access*, 2025.

[27] C. Sima, K. Renz, K. Chitta, L. Chen, H. Zhang, C. Xie, P. Luo, A. Geiger, and H. Li, "Drivelm: Driving with graph visual question answering," *arXiv*, 2023.

[28] Y. Zhou, L. Huang, Q. Bu, J. Zeng, T. Li, H. Qiu, H. Zhu, M. Guo, Y. Qiao, and H. Li, "Embodied understanding of driving scenarios," in *ECCV*, 2024.

[29] J. Mao, Y. Qian, J. Ye, H. Zhao, and Y. Wang, "Gpt-driver: Learning to drive with gpt," *arXiv*, 2023.

[30] X. Ding, J. Han, H. Xu, X. Liang, W. Zhang, and X. Li, "Holistic autonomous driving understanding by bird's-eye-view injected multi-modal large models," 2024.

[31] X. Ding, J. Han, H. Xu, W. Zhang, and X. Li, "Hilm-d: Enhancing mllms with multi-scale high-resolution details for autonomous driving," *IJCV*, 2025.

[32] J. Fan, J. Wu, J. Gao, J. Yu, Y. Wang, H. Chu, and B. Gao, "Mllm-sul: Multimodal large language model for semantic scene understanding and localization in traffic scenarios," *arXiv*, 2024.

[33] H.-k. Chiu, R. Hachiuma, C.-Y. Wang, S. F. Smith, Y.-C. F. Wang, and M.-H. Chen, "V2v-llm: Vehicle-to-vehicle cooperative autonomous driving with multi-modal large language models," *arXiv*, 2025.

[34] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002.

[35] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *ACL*, 2005.

[36] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *ACL*, 2004.

[37] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *CVPR*, 2015.

[38] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *ECCV*, 2016.

[39] D. Chan, S. Petryk, J. E. Gonzalez, T. Darrell, and J. Canny, "Clair: Evaluating image captions with large language models," *arXiv*, 2023.

[40] V. Dat, N. Bao, and P. Hung, "Hybridnets: End-to-end perception network," *Pattern Recognition and Image Analysis*, 2025.

[41] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *arXiv*, 2024.

[42] R. Zhang, J. Han, C. Liu, P. Gao, A. Zhou, X. Hu, S. Yan, P. Lu, H. Li, and Y. Qiao, "Llama-adapter: Efficient fine-tuning of language models with zero-init attention," in *ICLR*, 2024.

[43] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *arXiv*, 2024.

[44] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan *et al.*, "Deepseek-v3 technical report," *arXiv*, 2024.