





Article

Using Multimodal Large Language Models (MLLMs) for Automated Detection of Traffic Safety-Critical Events

Mohammad Abu Tami ¹, Huthaifa I. Ashqar ^{2,3,*}, Mohammed Elhenawy ^{4,5}, Sebastien Glaser ⁴ and Andry Rakotonirainy ⁴

¹ Natural, Engineering and Technology Sciences Department, Arab American University, Jenin P.O Box 240, Palestine; m.abutami@student.aaup.edu

² Civil Engineering Department, Arab American University, Jenin P.O Box 240, Palestine

³ Artificial Intelligence Program, Fu Foundation School of Engineering and Applied Science, Columbia University, New York, NY 10027, USA

⁴ CARRS-Q, Queensland University of Technology, Kelvin Grove, QLD 4059, Australia; mohammed.elhenawy@qut.edu.au (M.E.); sebastien.glaser@qut.edu.au (S.G.); r.andry@qut.edu.au (A.R.)

⁵ Centre for Data Science, Queensland University of Technology, Kelvin Grove, QLD 4059, Australia

* Correspondence: huthaifa.ashqar@aaup.edu

Abstract: Traditional approaches to safety event analysis in autonomous systems have relied on complex machine and deep learning models and extensive datasets for high accuracy and reliability. However, the emerge of multimodal large language models (MLLMs) offers a novel approach by integrating textual, visual, and audio modalities. Our framework leverages the logical and visual reasoning power of MLLMs, directing their output through object-level question–answer (QA) prompts to ensure accurate, reliable, and actionable insights for investigating safety-critical event detection and analysis. By incorporating models like Gemini-Pro-Vision 1.5, we aim to automate safety-critical event detection and analysis along with mitigating common issues such as hallucinations in MLLM outputs. The results demonstrate the framework’s potential in different in-context learning (ICT) settings such as zero-shot and few-shot learning methods. Furthermore, we investigate other settings such as self-ensemble learning and a varying number of frames. The results show that a few-shot learning model consistently outperformed other learning models, achieving the highest overall accuracy of about 79%. The comparative analysis with previous studies on visual reasoning revealed that previous models showed moderate performance in driving safety tasks, while our proposed model significantly outperformed them. To the best of our knowledge, our proposed MLLM model stands out as the first of its kind, capable of handling multiple tasks for each safety-critical event. It can identify risky scenarios, classify diverse scenes, determine car directions, categorize agents, and recommend the appropriate actions, setting a new standard in safety-critical event management. This study shows the significance of MLLMs in advancing the analysis of naturalistic driving videos to improve safety-critical event detection and understanding the interactions in complex environments.

Keywords: multimodal large language models (MLLMs); safety-critical events; in-context learning (ICL); self-ensemble learning; object-level question–answers (QAs)



Citation: Abu Tami, M.; Ashqar, H.I.; Elhenawy, M.; Glaser, S.; Rakotonirainy, A. Using Multimodal Large Language Models (MLLMs) for Automated Detection of Traffic Safety-Critical Events. *Vehicles* **2024**, *6*, 1571–1590. <https://doi.org/10.3390/vehicles6030074>

Academic Editor: Elzbieta Macioszek

Received: 5 August 2024

Revised: 30 August 2024

Accepted: 1 September 2024

Published: 2 September 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The advent and development of autonomous driving technologies have marked a significant transformation in the automotive industry, reshaping how vehicles interact with their environment and with each other. The concept of autonomous driving has evolved over several decades, beginning with basic cruise control systems in the 1950s [1,2] and progressing to the sophisticated connected and automated vehicles (CAVs) of today [3,4]. The integration of advanced sensors, machine learning algorithms, and communication technologies has enabled vehicles to perform complex tasks such as navigation, obstacle avoidance, and decision-making with minimal human intervention [5].

As the capabilities of autonomous vehicles have expanded, so too have the challenges associated with ensuring their safety and reliability. Traffic safety has always been a critical concern in the development of autonomous vehicles, as these systems must be able to respond to a wide range of dynamic and unpredictable situations on the road. This is particularly important in the context of safety-critical events, such as sudden changes in traffic patterns, unexpected obstacles, and potential collisions. Additionally, traffic safety is influenced not only by the type of vehicle—whether conventional or autonomous—but also by the volume of traffic, as higher traffic volumes increase the probability of unsafe conditions [6]. The ability of autonomous vehicles to detect, analyze, and respond to these events in real time is crucial for preventing accidents and ensuring the safety of all road users [7].

The recent advancement breakthrough in large language models (LLMs) has revealed the potential usage in the complex challenging environment of analyzing driving videos. Many researchers have investigated the potential of utilizing LLMs in analyzing driving videos through textual representations [8–11]. With the advancement of MLLMs, a new merger has been reached with the power reasoning of LLMs in the different modalities of text, image, and audio [12,13].

Critical-safety event analysis is considered one of the complex and critical environments that could benefit from the new MLLM breakthrough. While full critical-safety event detection might still be a far reach, MLLM could advance the understanding of the dynamic variety of road transportation through providing textual analysis of the visual representation of the environment and the different agents in it, then using this analysis to provide a direct, concise, and actionable early warning to the ego-driver in the case of any potential hazards.

This capability of MLLMs to synthesize information across multiple modalities—such as visual cues from driving videos, environmental sounds, and contextual data—opens new avenues for enhancing driver assistance systems. By integrating textual analysis with real-time video and audio inputs, MLLMs can facilitate more accurate and context-aware interpretations of driving scenarios, which are crucial for preventing accidents and improving road safety. The ability to generate natural language descriptions or warnings based on complex visual and auditory inputs allows for a more intuitive interface between the technology and the driver, which has the ability to reduce cognitive load and improve reaction times in critical situations.

Moreover, the adaptability of MLLMs to learn from diverse data sources, including different driving environments and conditions, enhances their robustness and reliability. As MLLMs continue to evolve, their role in the domain of autonomous driving and driver assistance is expected to expand, offering more sophisticated solutions for anticipating and mitigating safety risks. The integration of these models into real-world applications could mark a significant step forward in achieving safer and more efficient transportation systems. In this context, the development of MLLMs presents an exciting opportunity to revolutionize the field of critical-safety event detection in driving. As the technology matures, the potential for creating systems that can not only detect but also predict, prevent, and recommend about critical-safety events becomes increasingly feasible, moving us closer to a future where road transportation is not only smarter but also significantly safer.

2. Related Works

Before the era of MLLM, researchers in safety event analysis relied on developing a complex machine learning model from the ground up, utilizing thousands of annotated datasets to achieve high accuracy and reliability. For instance, the authors in [14] proposed a supervised encoder–decoder model where a pre-trained ResNet-101 was used as encoder to extract the visual and flow features of 17k distinct ego-car dash cam scenarios, and a neural image caption generation structure [15] as a decoder to predict the caption of the street frames while attaining the extracted features from the encoder part.

The study by Zhenjie et al. LLM4Drive [16] reviews the integration of large language models (LLMs) in autonomous driving systems, highlighting their potential to enhance decision-making, perception, and interaction through advanced reasoning and contextual understanding. The survey categorizes current research into planning, perception, question answering, and generation, addressing the challenges of transparency, scalability, and real-world application. It underscores the need for robust datasets and interpretable models to build trust and improve system reliability in autonomous driving.

Cui et al. explores the integration of LLMs and vision foundation models (VFM) in enhancing autonomous driving systems [17]. The work covers the historical evolution from early sensor-based approaches to advanced deep learning techniques that improve perception, planning, and decision-making. The paper also reviews existing multimodal tools and datasets like KITTI [18] and nuScenes [19]. A study by Chen et al. [20] proposed a pre-training method that aligns numeric vector modalities with LLM (GPT3.5) representations, improving the system's ability to interpret driving scenarios, answer questions, and make decisions. Furthermore, the study titled "DriveMLM" [21] introduces an LLM-based autonomous driving (AD) framework that aligns multi-modal LLMs with behavioral planning states, enabling closed-loop autonomous driving in realistic simulators. It bridges the gap between language decisions and vehicle control commands by standardizing decision states according to the off-the-shelf motion planning module. On another hand, the "Drive As you Speak" paper [22] presents an approach to enabling human-like interaction with large language models in autonomous vehicles. It leverages LLMs to understand and respond to human commands, demonstrating the potential of LLMs in creating more intuitive and user-friendly autonomous driving experiences. Moreover, AccidentGPT [23] introduces a multi-modal model for traffic accident analysis, which was capable of reconstructing crash processes and providing comprehensive reports.

Recent advancements also explored the integration of sensor data and real-time processing using LLMs to enhance autonomous driving capabilities. A study by Zhang et al. [24] examined the integration of LLMs with LiDAR and radar data to improve object detection and tracking. Similarly, a study by Singh et al. [25] highlighted the use of LLMs in predicting pedestrian behavior by analyzing both visual signals and contextual information, which increased the reliability of autonomous systems in urban settings. Moreover, another study by Lopez et al. [26] focused on utilizing LLMs to interpret driver motions and voice commands, which might facilitate a more natural interaction between the driver and the vehicle. Furthermore, a study by Kim et al. [27] explored the analysis of live video feeds from dashboard cameras, which enabled the early detection of potential hazards such as sudden lane changes or road obstacles. This approach allowed for timely warnings and interventions, which has the potential to enhance the safety of autonomous driving systems.

A study by Hussien et al. [28] also highlighted the potential of integrating LLMs with knowledge graphs and retrieval-augmented generation (RAG)-based explainable frameworks to provide explainable predictions of road user behaviors, which is crucial for developing safe automated driving systems. This study underscored the importance of explainability in the potential deployment of LLMs in critical environments like autonomous driving.

In addition to LLMs, contributions have been made in the domain of cooperative control of CAVs. The study by Liang et al. [29] explores a multi-agent system (MAS) architecture designed to facilitate the cooperative control of CAVs. This hierarchical architecture enables vehicles to collaborate effectively in complex traffic environments, sharing information and making collective decisions that enhance safety and efficiency. The study emphasizes the importance of cooperation among autonomous vehicles, particularly in scenarios where rapid decision-making and coordination are critical to preventing accidents and ensuring smooth traffic flow.

Despite the promising developments in using MLLMs for autonomous driving and intelligent transportation systems, a significant gap remains in the application of these models for safety-critical event analysis. Existing studies focused on enhancing autonomous driving capabilities through improved perception and decision-making processes without

specifically addressing the unique challenges posed by safety-critical situations. This gap shows the need for a specialized approach that leverages the multimodal capabilities of LLMs to directly address the complexity of safety-critical events in driving scenarios and provide more explainable information and recommendations, which is very important for taking the right safety countermeasures. Current methods mostly rely on complex machine and deep learning models and extensive annotated datasets, which are not always feasible or scalable in real-world applications. There is a need for an easy-to-implement, scalable, and explainable framework that can automate the extraction of visual representations from raw video feeds and utilize object-level question–answer (QA) prompts to guide MLLMs in generating actionable insights for hazard detection and response.

This study aims to bridge this gap by introducing an MLLM framework specifically designed for the analysis and interpretation of safety-critical events. By integrating the different modalities of texts and images, our framework seeks to provide a more holistic and scalable view of dynamic driving environments. Furthermore, our approach emphasizes the automation of extracting visual representation from the raw video and feeding it to an MLLM with the creation of object-level QA prompts to guide the MLLM’s analysis, which focuses on generating actionable insights for safety-critical event detection and response. This study introduces a novel application of MLLMs in a domain where precision and reliable decision-making are essential, marking a significant step forward in the development of safer driving.

3. Preliminary

This section introduces the dataset utilized for evaluating the proposed framework and the Gemini model, which form the foundation for the experimental work in this study.

3.1. Dataset

Creating a dataset from driving videos that integrates language for visual understanding is a challenging task. This process is a resource-extensive task that requires trained human annotators for optimal accuracy and reliability. In addition, the variety and complexity of driving scenarios require a dataset rich in visual scenes. The dataset needs to cover a variety ranging from simple driving directions to complex situations involving pedestrians, other vehicles, and road signs.

Many researchers have either enhanced existing datasets with textual information [30–33] or developed new ones from scratch [14,34]. Notable among these are the DRAMA datasets [14]. DRAMA focuses on driving hazards and related objects, featuring video and object-level inquiries. This dataset supports visual captioning with free-form language descriptions and accommodates both closed and open-ended questions, making it essential for assessing various visual captioning skills in driving contexts. In addition, the vast variety found in DRAMA scenarios makes it a uniquely comprehensive resource for investigating and evaluating MLLM models on complex driving situations.

Considering these factors, this study selected the DRAMA dataset to utilize the ground truth label to report this paper’s experimental results. DRAMA’s detailed focus on hazard detection and its comprehensive framework for handling natural language queries make it exceptionally suitable for pushing forward research in safety-critical event analysis.

The DRAMA dataset includes multiple levels of human-labeled question–answer (QA) pairs. These include base questions regarding whether a risk exists in the scene, with Yes/No answers; scene classification into urban road, intersection, and narrow lane categories; questions about the direction of the ego-car, with options such as straight, right, and left; questions about potential hazard-causing agents in the scene, such as vehicles, pedestrians, cyclists, and infrastructure; and finally, questions about recommended actions for the ego-car driver based on the scene analysis, with eight possible actions, including stop, slow down, be aware, follow the vehicle ahead, carefully maneuver, start moving, accelerate, and yield. Figure 1 illustrates the distribution of each QA used in the studies,

where 300 distinct videos ranging from 2 to 5 s were employed to examine the effectiveness of MLLMs in detecting traffic safety-critical events.

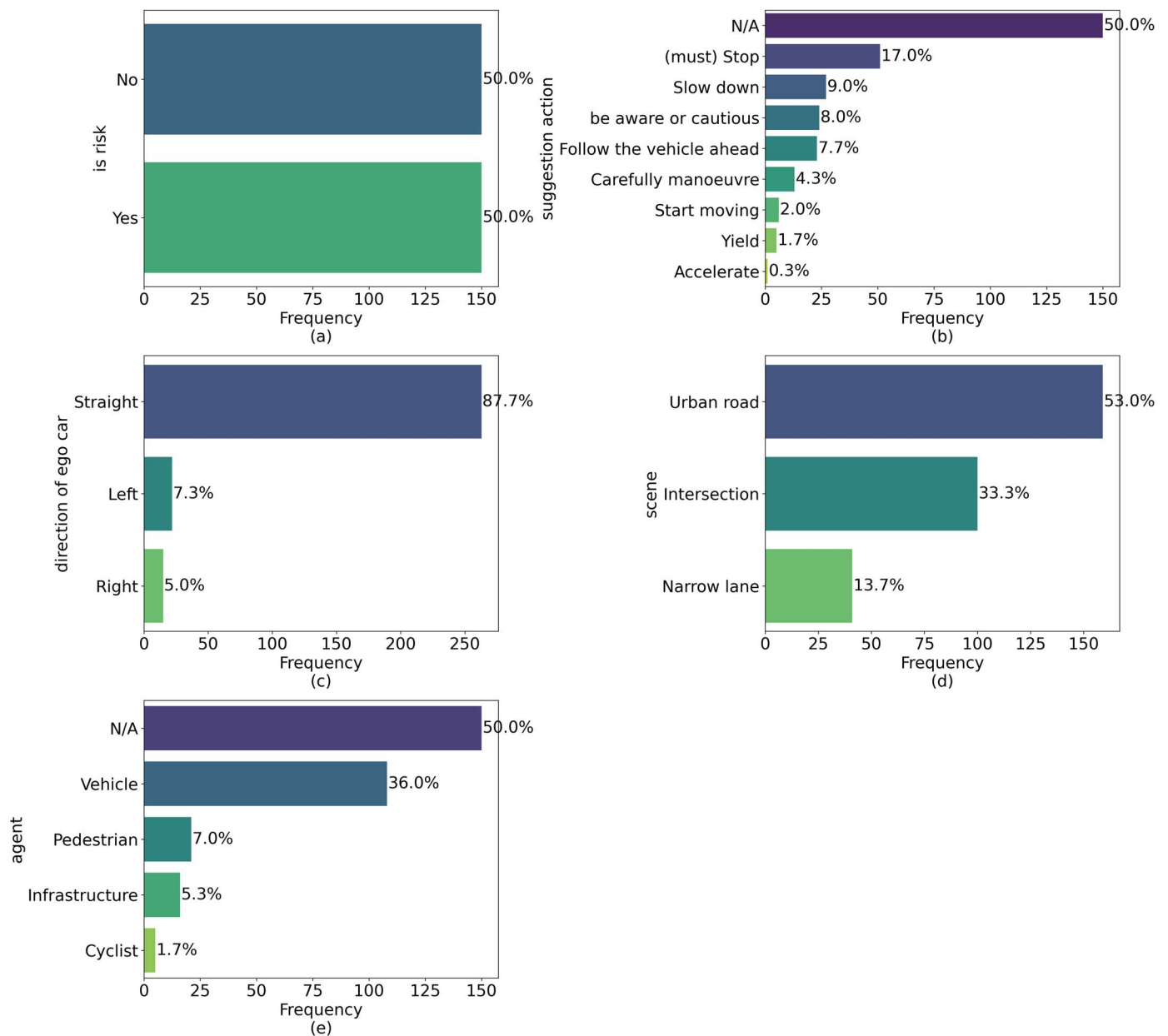


Figure 1. Distribution of QA categories in the DRAMA dataset for traffic safety-critical event detection including (a) is risk, (b) suggested action, (c) direction of ego car, (d) scene description, and (e) agent type.

3.2. Gemini MLLM

The framework for detecting safety-critical events from driving videos in this study utilizes the Gemini-Pro-Vision 1.5 MLLM [35]. This model was chosen for its advanced capabilities in logical and visual reasoning, particularly in identifying potential hazards across diverse traffic scenarios.

Gemini 1.5 is designed to process and integrate information across multiple modalities—text, images, and video—making it highly effective for tasks that require a deep understanding of both visual and textual data. This is crucial for driving scenarios where the model needs to interpret video frames and respond to natural language prompts simultaneously. One of the most striking features of the Gemini 1.5 model is its ability to handle a context window of up to

1 million tokens, which is significantly larger than most other models. This allows Gemini to process large amounts of data in a single pass.

4. Methodology

We conducted multiple experiments to investigate the capability and logical and visual reasoning power of MLLMs in identifying potential hazards across diverse traffic scenarios. To guide our investigation, we formulated the following research questions (RQs):

- RQ1: How effective are MLLMs at identifying traffic hazards using in-context learning (ICL) with zero-shot and few-shot learning approaches?
- RQ2: Does the number of frames used impact the accuracy of hazard detection in traffic scenarios?
- RQ3: What is the impact of self-ensembling techniques on the reliability and robustness of MLLMs in detecting critical traffic safety events?

The employed methods range from ICL with zero-shot and few-shot learning to varying the number of frames, utilizing textual context alongside visual frames, and implementing self-ensembling techniques. The subsequent sections present the proposed framework and its operational flow for detecting critical traffic safety events, followed by an overview of the different methodologies employed and the implemented prompt design.

4.1. Framework

The framework illustrated in Figure 2 is designed for detecting safety-critical events from driving video extracted from car dash cams, utilizing a multi-stage QA approach with an MLLM, specifically, Gemini-pro-vision 1.5. The process initiates with frame extraction, where the system automatically collects video frames from the ego-vehicle's camera at regular intervals (i.e., every second). These frames are subjected to the hazard detection phase, where the model assesses the scene for potential dangers.

Upon identifying a hazard, the framework employs a tripartite categorization strategy to probe the nature of the threat further, using "What," "Which," and "Where" queries to reveal the object-level details. In the "What" phase, the MLLM classifies the entities detected by the camera, differentiating among agents like pedestrians, vehicles, or infrastructure elements. The "Which" stage involves the MLLM identifying specific features and attributes of these agents, such as pedestrian appearance, vehicle make and model, or infrastructure type, providing vital contextual insights for decision-making.

The final "Where" phase tasks the MLLM with determining the spatial location and distance of the hazard agents from the ego-car, including their position on the road, proximity to the vehicle, and movement direction. This spatial information is critical for the ego-car system to make a safer navigation decision. We tested the model across different dimensions to evaluate model performance in various tasks for each safety-critical event, including identifying risky scenarios, classifying different scenes, determining car direction, classifying agents, and suggesting correct actions.

The framework addresses traffic safety-critical events through a thorough analysis of interactions and road environments in three folds. First, the framework recognizes and evaluates scenarios where the interaction between the ego-vehicle and other road users (i.e., vehicles, pedestrians, and cyclists) or infrastructure may result in safety-critical events. These events include sudden stops, lane changes, or crossing pedestrians that could lead to hazardous situations if not managed correctly. Second, the model identifies and localizes risks within the driving environment, determining the exact location and potential impact of hazardous objects. It assesses the relative position of these objects, such as vehicles cutting into the lane or pedestrians crossing unexpectedly, which are crucial for proactive hazard mitigation. Third, the framework adapts to various road types, such as wide roads, intersections, and narrow streets, each presenting unique challenges. For example, intersections are flagged as particularly high risk due to the convergence of multiple traffic flows, necessitating precise detection and decision-making capabilities from the system.

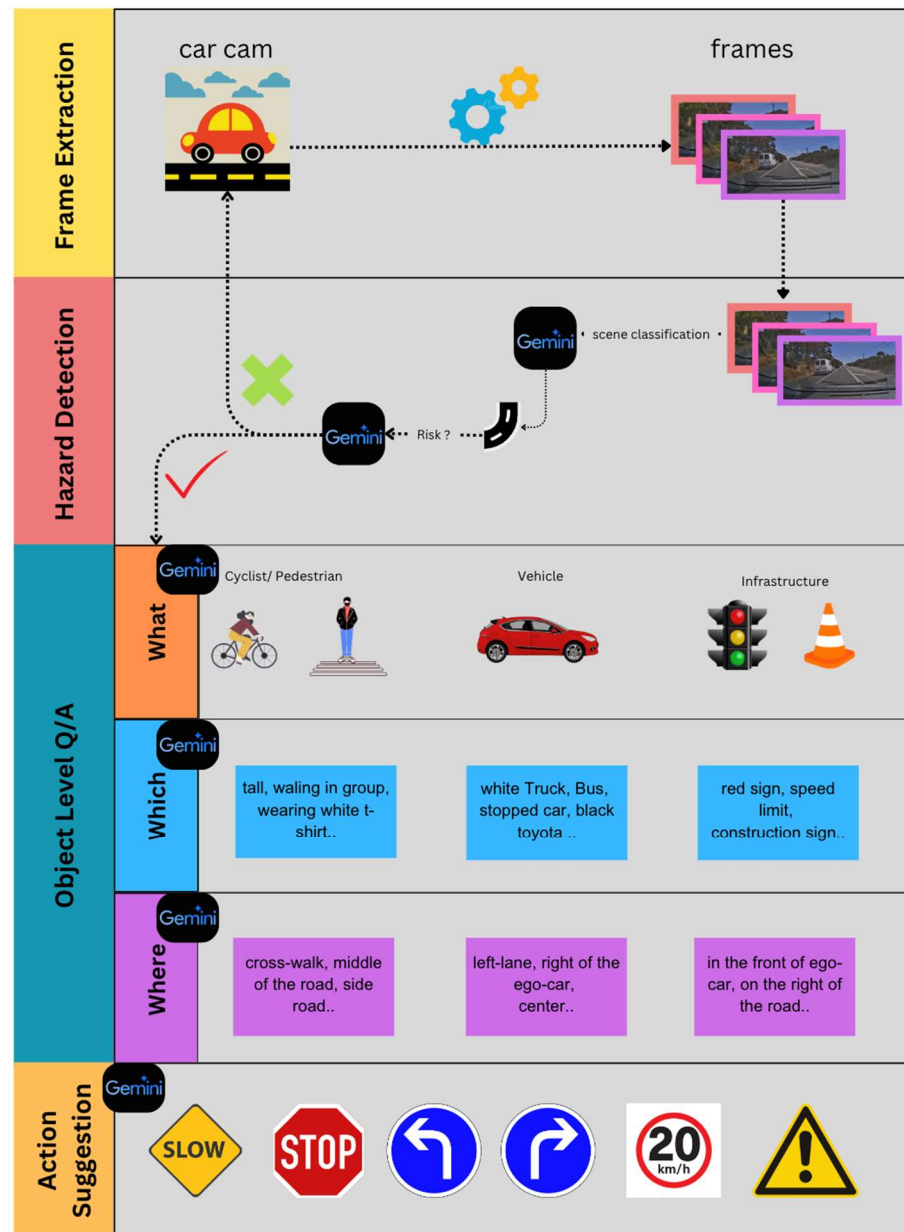


Figure 2. Automated multi-stage hazard detection framework for safety-critical events using MLLMs.

4.2. Analysis Methods

We incorporated different methods to enhance the detection of safety-critical events. These methods were experimented with to enable the system to focus on the most relevant information, thereby optimizing processing speed and accuracy in detecting safety-critical events.

One approach employed was the sliding window frame capture technique, which systematically captures subsets of video frames by defining a window that slides over the video timeline. This window captures a specific range of frames from t_i to t_{i+n} , where t_i represents the initial frame in the window, and n is the number of frames included in each window. The mathematical representation of this method can be expressed in Equation (1). This method allows for the dynamic adjustment of the window size based on the specific requirements of the analysis, which allows the framework to balance data completeness and processing efficiency.

$$Window(t_i) = \{Frame(t_i), Frame(t_{i+1}), \dots, Frame(t_{i+n-1})\} \tag{1}$$

In-context learning (ICL) was also integrated into the framework to enhance the predictive capabilities of MLLMs by providing them with relevant examples during the inference process. This technique is particularly effective in scenarios where annotated data are rare or the model has to adapt to new situations while in progress. In our framework, we explored two primary settings of in-context learning: zero-shot and few-shot learning. Zero-shot learning allows the model to make predictions about safety-critical events without having seen any prior examples specific to the task. The model relies completely on its base knowledge and reasoning capabilities. This method might be beneficial for its ability to generalize across diverse and unforeseen scenarios. The zero-shot learning process can be mathematically described in Equation (2).

$$Prediction = MLLM(Prompt, Frames) \quad (2)$$

where *Frames* is a sequence of specific frames and *Prompt* is a general question or instruction provided to the MLLM to guide the analysis.

On the other hand, few-shot learning allows the model to be exposed to a small number of annotated examples relevant to the task before making predictions. This kind of learning enables the model to quickly adapt and improve its accuracy by leveraging specific patterns and features observed in the examples seen. The few-shot learning process is presented in Equation (3).

$$Prediction = MLLM(Prompt, Examples, Frames) \quad (3)$$

where *Examples* is the annotated observations used to fine-tune the model's reasoning.

Label-augmented learning (LAL) was another method employed, providing context for the MLLM about how the data were originally annotated. This method helps the MLLM to understand the labeling scheme and the specific characteristics that were considered during the annotation process. By incorporating this context, the model can align its outputs more closely with the annotated data, thereby improving accuracy and consistency.

Image-augmented learning (IAL) involved applying various image augmentation techniques to the images before they were fed into the MLLM for safety-critical event detection QA. These augmented images aim to direct the MLLM to different areas within the language distribution it relies on for generating responses, as illustrated in Figure 3.

By introducing augmented images in the prompt, the MLLM can start at various points within the data distribution, which influences the diversity of local sampling results.

Subsequently, the outcomes from different model sampling processes are aggregated using a top-k voting mechanism to determine the outcome response. This approach aims to aid the model in producing textual responses that more accurately represent the scene under query.

Self-ensemble learning is another strategy used to boost the performance of our framework. It involves generating multiple predictions from the MLLM using slightly different contexts or parameters, such as model temperature, and then combining these predictions to obtain a more robust and accurate result. This approach reduces the likelihood of errors and increases the reliability of hazard detection. The self-ensemble process can be described mathematically Equation (4).

$$Prediction = Top - k \left(\{ MLLM(Prompt, Frames) \}_{k=1}^N \right) \quad (4)$$

where N is the number of individual predictions. The $top - k$ voting mechanism selects the k most frequent predictions among the N generated predictions, which enhances the overall performance by focusing on the most consistently identified outcomes.

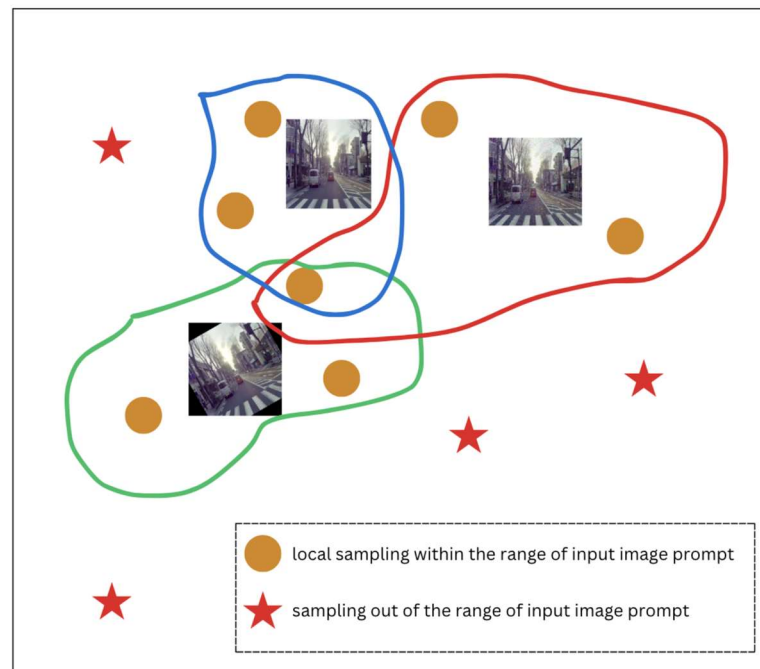


Figure 3. Conceptual 2-D diagram of augmented image prompting. The key idea of using different augmentation for the same scene under investigation is to direct the model to different places in the language distribution, which could help the model with generating more textual representation of the scene when generating a response through local sampling. The different colored areas showed the an example of how image augmentation can be done.

4.3. Prompt Design

The design of the prompt is pivotal in guiding the MLLM to accurately evaluate and respond to safety-critical events in driving scenarios. This prompt was designed to ensure a structured and systematic analysis of the input frames, thereby minimizing the risk of hallucination and enhancing the reliability of the MLLM's outputs, as seen in Figure 4. The structure of the prompt is intended to break down the evaluation process into clear, logical steps, which helps the MLLM to focus on specific aspects of the scene sequentially.

The prompt design benefits the MLLM by ensuring a structured analysis that breaks down the evaluation into discrete steps, allowing the model to focus on one aspect at a time and reducing cognitive overload. By using predefined categories and limited response options, it controls the model's output, minimizing the risk of hallucinations. Each step builds on the previous one, providing a holistic and context-aware understanding of the scene, which improves decision-making. The final step of suggesting actionable recommendations ensures that the analysis is not just descriptive but also prescriptive, offering clear and practical guidance for safe driving.

In summary, the prompt design is tailored to enhance the MLLM's ability to accurately and reliably detect safety-critical events from driving video frames. Its structured approach, combined with controlled output options, significantly mitigates the risk of hallucination, ensuring that the model's responses are both relevant and actionable.



Figure 4. Example of textual prompt with two-frame scene with the corresponding response from Gemini.

5. Results

The work presented in this paper investigated the potential of leveraging the capabilities of MLLMs in analyzing safety-critical event scenarios using multi-modal data integration and dynamic contextual data reduction for guiding the model's output. The prediction illustrated in Figure 5 showcases the proficiency of Gemini-Pro-Vision 1.5 in zero-shot learning scenarios.

To understand the effectiveness of the proposed framework, a series of experiments was carried out utilizing Gemini-Pro-Vision 1.5. The results as shown in Table 1 are analyzed across various frames and in-context learning settings, including zero-shot and few-shot learning, as well as additional strategies like self-ensemble learning and image-augmented learning. We tested the model across different dimensions to evaluate model performance in various tasks for each safety-critical event, including identifying risky scenarios, classifying different scenes, determining car direction, classifying agents, and suggesting correct actions.

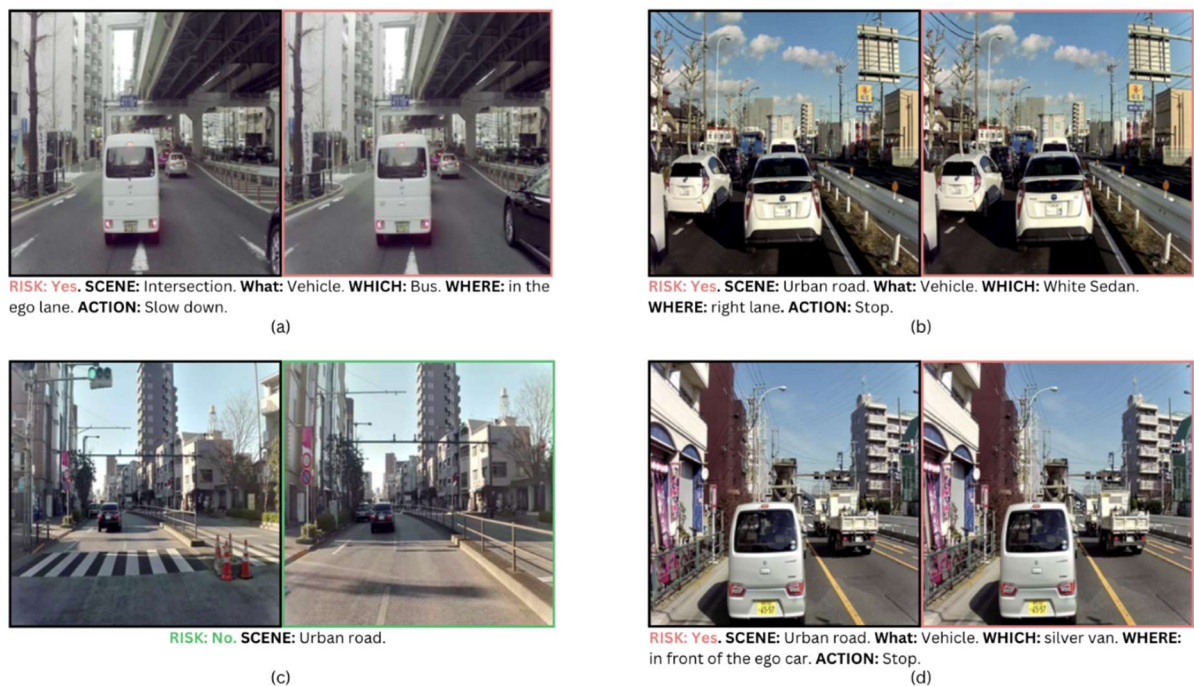


Figure 5. Output from Gemini-Pro-Vision 1.5 analysis with sliding window ($n = 2$). Gemini predicted (a), (b), and (d) as critical-safety events, while (c) is not.

Table 1. Comparative performance analysis of QA frameworks across different methods.

Method	Frame 1/Candidates ²	Is Risk %	Scene %	Direction of Car %	Agent %	Suggestion Action %	Overall %
Zero-shot	1	68	64	87	56	37	62.4
Zero-shot	2	51	66	89	46	39	58.2
Zero-shot	3	52	66	86	48	38	58
Zero-shot	4	47	72	86	44	35	56.8
Few-shot	1	72	73	87	57	39	65.6
Few-shot	3	76	76	87	59	40	67.6
Few-shot	5	76	76	87	59	40	67.6
Few-shot	7	75	80	89	63	45	70.4
Few-shot	10	79	81	90	64	44	71.6
Self-ensemble	3	69	65	87	56	38	63
Self-ensemble	5	71	67	89	55	39	64.2
Self-ensemble	7	70	67	88	55	39	63.8
Self-ensemble	9	66	66	88	54	38	62.4
LAL	-	68	66	83	48	34	59.8
IAL	-	67	60	80	55	33	59

¹ For zero- and few-shot learning. ² For self-ensemble learning.

5.1. Zero-Shot Learning Results

Zero-shot learning demonstrated a variable performance profile across different metrics and frame counts, as seen in Figure 6. Initially, a single frame yielded an overall accuracy of 62.4%, with notable performance in detecting the direction of the car (87%) and scene classification (64%). However, as the number of frames increased, overall performance slightly decreased, reaching 56.8% with four frames. The decrease in performance with additional frames suggests potential trade-offs between the depth of context provided and the model’s ability to generalize without prior task-specific examples. The impact of frame count on metrics like agent classification and suggested actions also reflects the model’s challenge in maintaining accuracy across varying contexts.

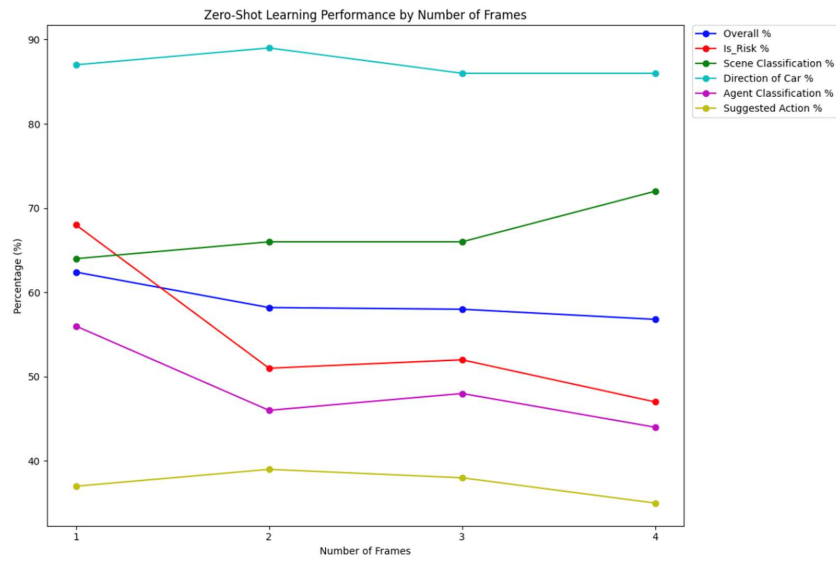


Figure 6. Zero-shot learning performance across different numbers of frames.

5.2. Few-Shot Learning Results

Few-shot learning demonstrated a clear trend of improvement with an increasing number of shots. The performance improved progressively from 1-shot to 10-shot scenarios, with the highest overall percentage (71.6%) achieved with 10 shots. This improvement, as seen in Figure 7, was evident across all metrics, particularly in scene classification and direction of cars, where the highest values were observed with 10-shot learning. The consistency in performance metrics with five-shot and seven-shot suggests that a moderate number of examples already offers substantial benefits.

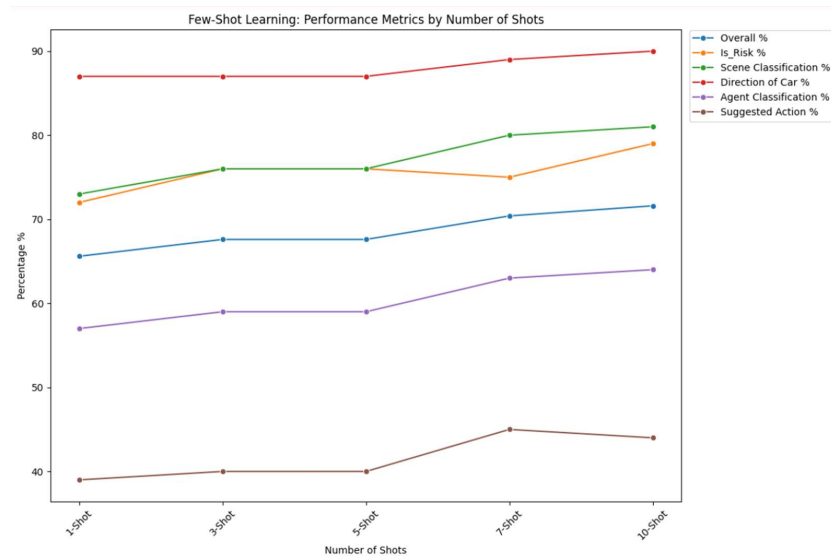


Figure 7. A few-shot learning performance across different numbers of examples.

When comparing zero-shot methods (including one frame, two frames, three frames, and four frames) to few-shot methods, it is evident that few-shot learning consistently outperforms zero-shot learning, as seen in Figure 8. The bar plots highlight that, with zero-shot methods achieving lower percentages across all metrics. For instance, the “is_risk %” metric showed a significant improvement from 68% in the one-frame zero-shot method to 79% in the 10-shot method. Similarly, “scene classification %” saw an increase from 64% with 1 frame to 81% with 10 shots. The comparison underscores the robustness of

few-shot learning in improving model performance across various metrics, showcasing its superiority over zero-shot learning approaches.

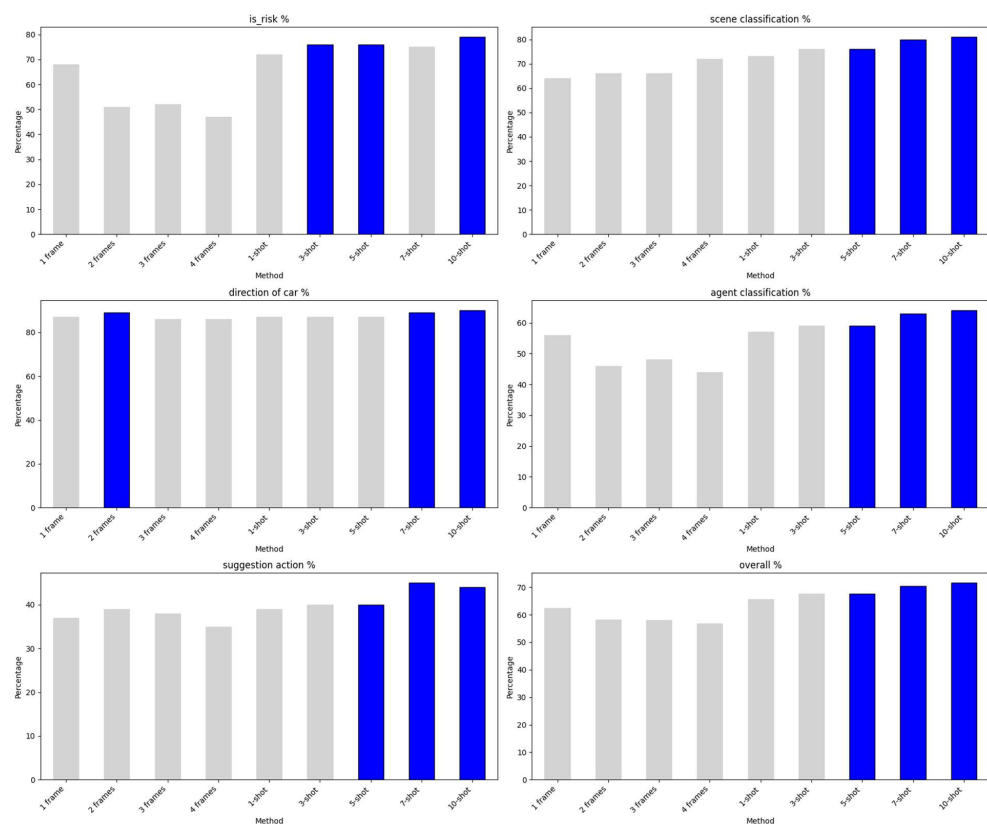


Figure 8. Comparison of zero-shot and few-shot methods across various metrics (top 3 highlighted).

5.3. Self-Ensemble Learning Results

Self-ensemble learning provided a relatively stable performance, with slight improvements as the number of candidates increased. The five-candidate configuration yielded the highest overall percentage (64.2%), as illustrated in Figure 9, showing that aggregating predictions from multiple candidates helped enhance performance. Although the improvements in metrics such as *is_risk* and *scene classification* were not drastic, the approach demonstrated increased reliability in hazard detection.

When comparing zero-shot (1-frame) methods to self-ensemble methods (3, 5, 7, 9 candidates), as in Figure 10, it is clear that self-ensemble methods generally offered better performance. The bar plots show that self-ensemble methods frequently surpassed the zero-shot (1-frame) approach. For instance, the “*scene classification %*” and “*direction of car %*” metrics showed noticeable improvements with self-ensemble methods. The three-candidate and five-candidate configurations consistently performed well across these metrics.

The use of self-ensemble methods enhanced the overall metric, which meant a more balanced and robust model performance. The highest overall in self-ensemble methods (64.2% with 5 candidates) still outperformed the zero-shot (1-frame) approach (62.4%). This trend is consistent across other metrics, such as “*agent classification %*” and “*suggestion action %*,” where the self-ensemble methods exhibited a slight edge.

While the improvements in individual metrics like “*is_risk %*” and “*scene classification %*” were modest, the aggregated gains across all metrics suggest that self-ensemble learning provides a more reliable and effective approach than the zero-shot (1-frame) method. This highlights the value of leveraging multiple candidate predictions to improve the robustness and accuracy of the model’s performance across diverse evaluation criteria.

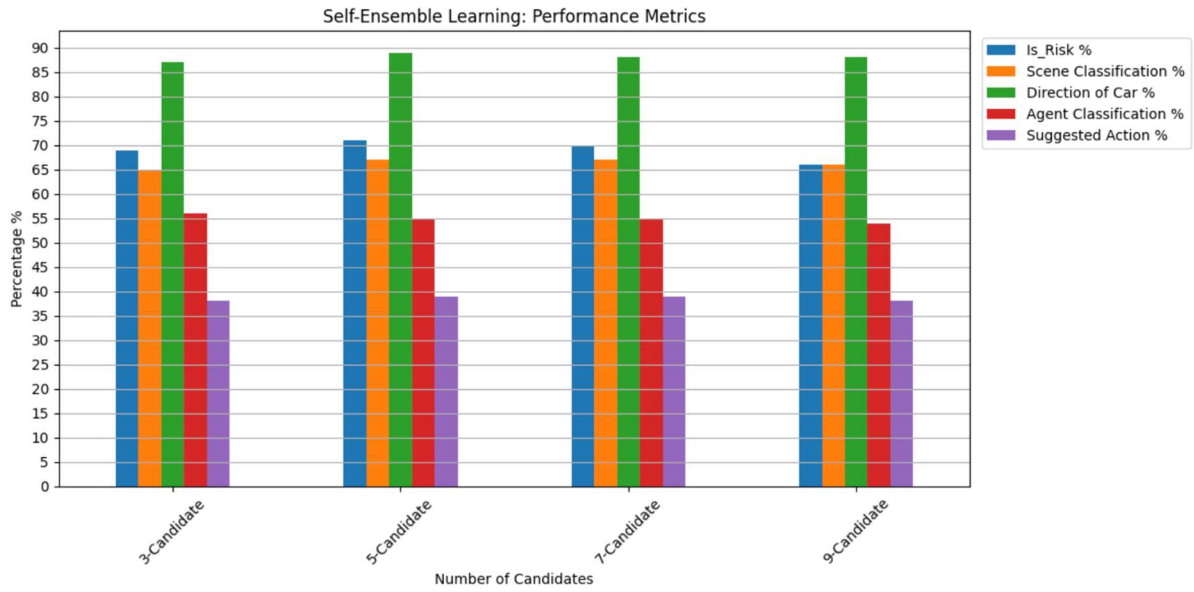


Figure 9. Self-ensemble learning across different number of candidates with top-k voting.

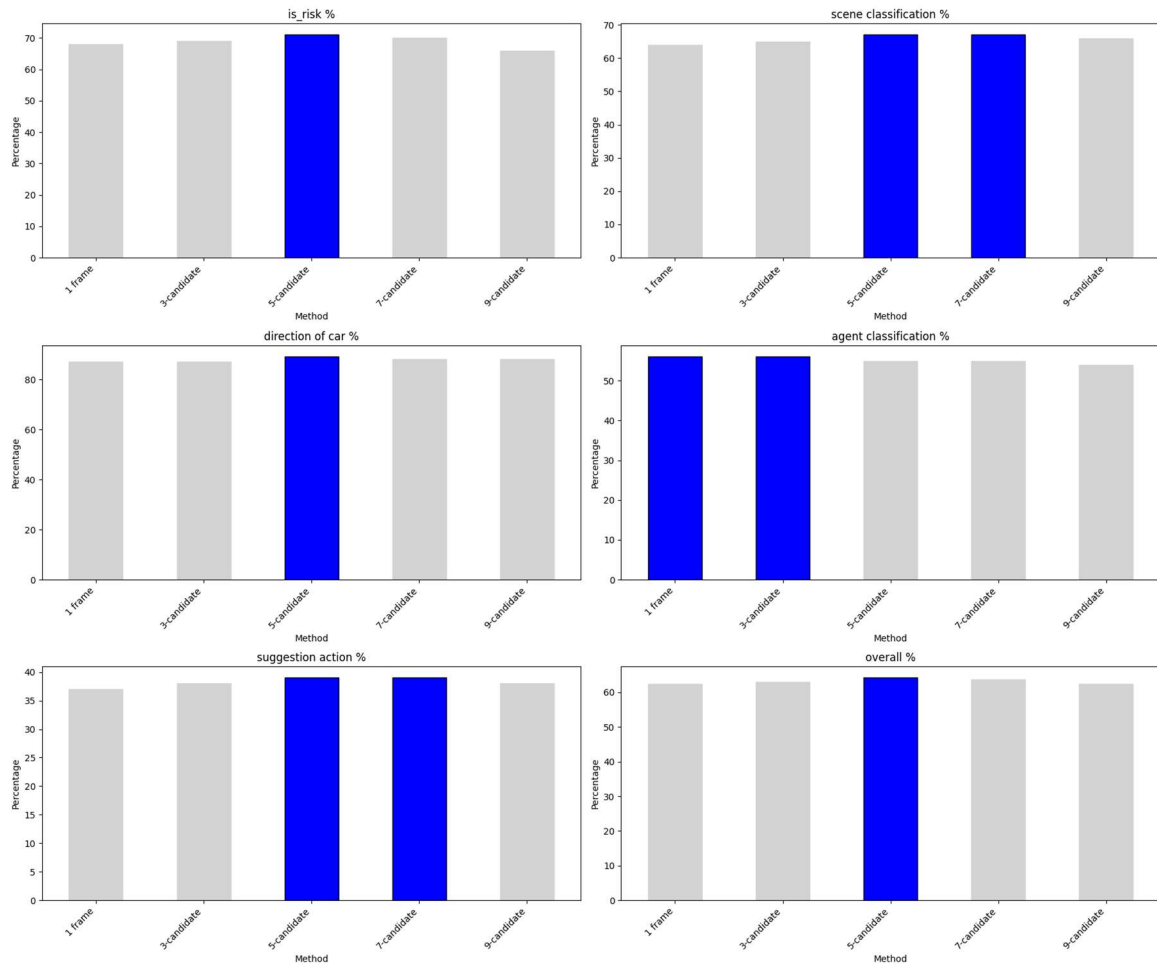


Figure 10. Comparison of zero-shot (1-frame) and self-ensemble methods across various metrics (top bar highlighted).

5.4. Image-Augmented Learning Results

Image-augmented learning with the top-k method resulted in lower overall performance compared to other methods, as seen in Figure 11, with an overall percentage of

59.0%. The image augmentation approach appeared to have a mixed impact, providing a moderate enhancement in some metrics but falling short in overall accuracy and suggested action classification. This suggests that while image augmentation introduces variability, its effect on overall performance needs further refinement and evaluation.

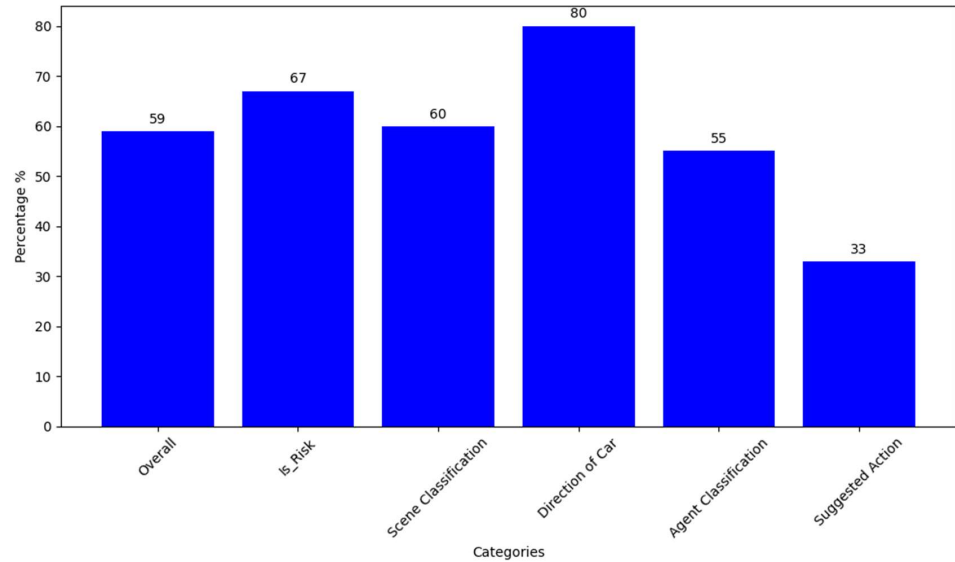


Figure 11. Image-augmented learning performance with top-k voting.

When comparing the zero-shot (1-frame) method to the image-augmented method across different metrics, as in Figure 12, it becomes evident that each approach has its strengths and weaknesses. The zero-shot (1-frame) method achieved a higher “is_risk %” (68%) compared to the image-augmented method (67%). Similarly, in “scene classification %,” the zero-shot method performed better (64%) than the image-augmented method (60%).

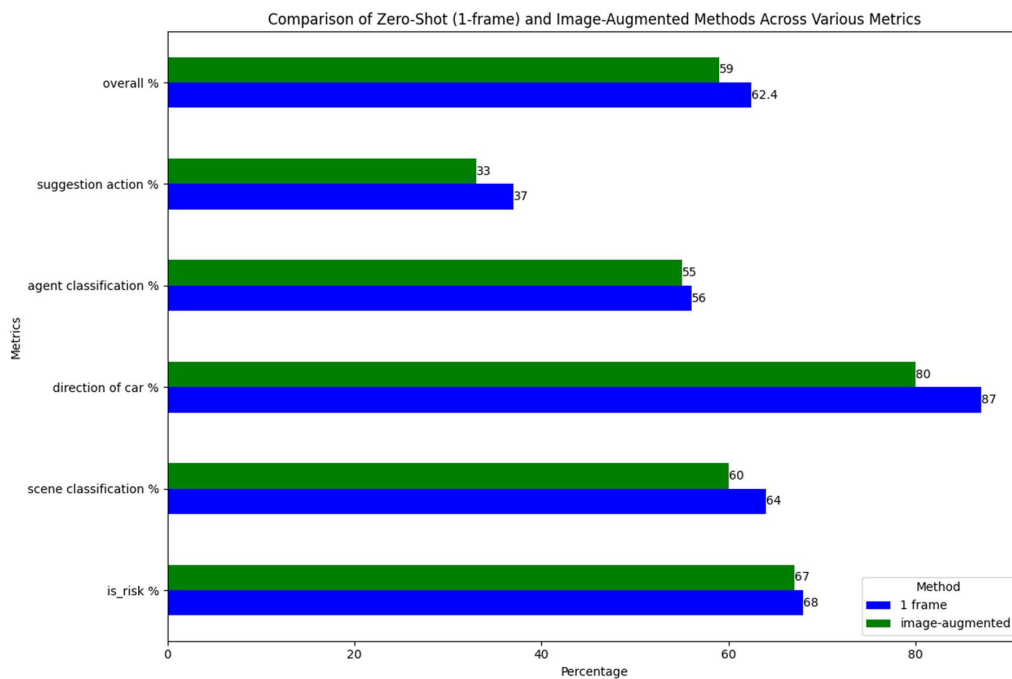


Figure 12. Comparison of zero-shot (1-frame) and image-augmented methods across various metrics.

Few-shot learning, as shown in Figure 13, consistently outperformed other methodologies across most metrics, with a notable improvement in overall performance as the

number of shots increased. Zero-shot learning, while useful, showed decreased performance with an increasing number of frames, indicating that it may benefit from being combined with other methods for optimal results. Self-ensemble learning provided a modest increase in performance and stability, particularly in is_risk and scene classification metrics. Image-augmented learning, although innovative, showed less effectiveness compared to the other methods, suggesting that further exploration and refinement of augmentation techniques are necessary. These results highlight the potential of MLLMs in automated traffic safety event detection and offer insights into optimizing their use for various safety-critical scenarios.

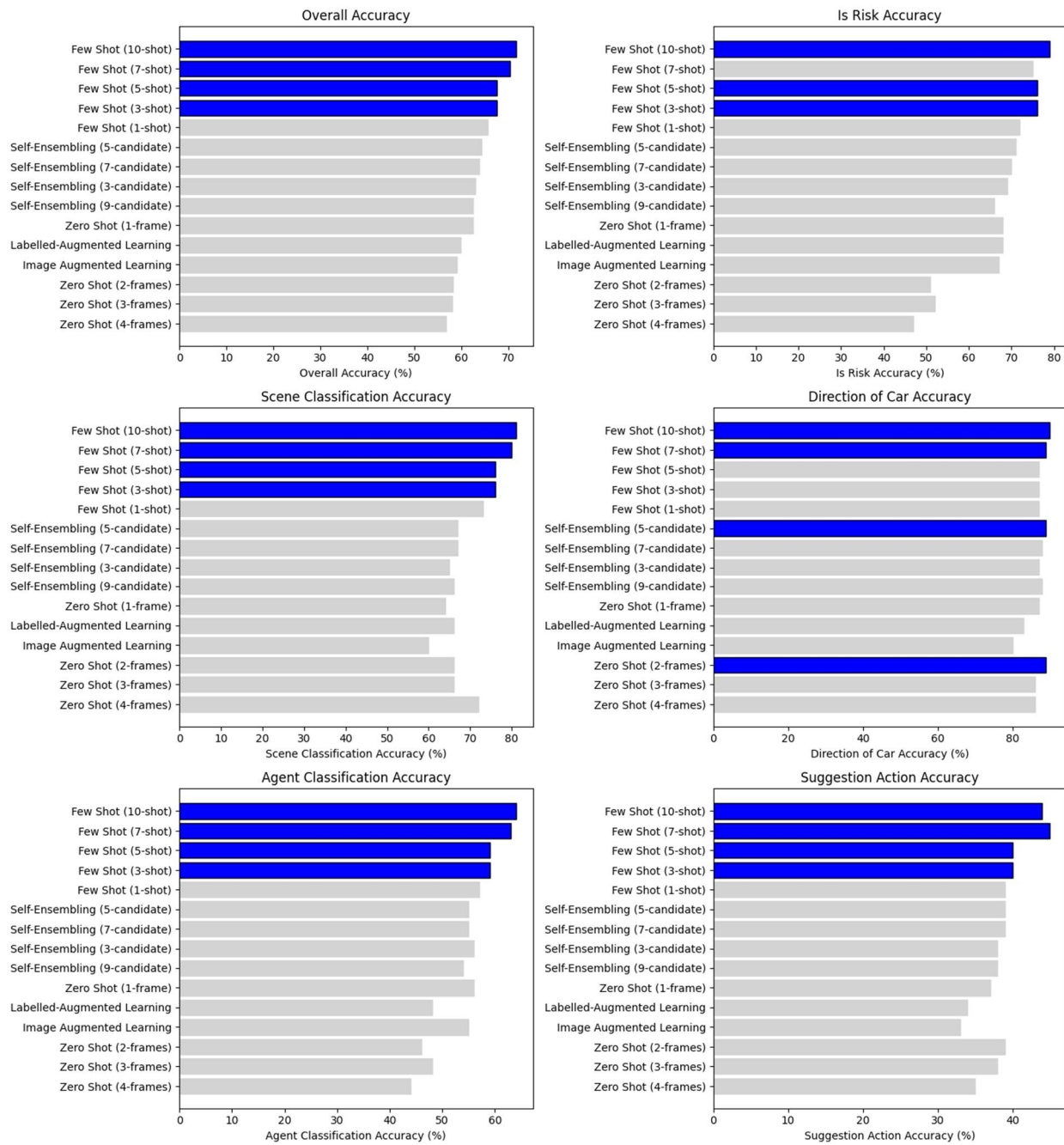


Figure 13. Overall performance comparison across different learning methods. The highlighted bars showed the highest accuracy from each category.

6. Discussion

Across all methods, few-shot learning stands out as the most effective approach for improving overall accuracy and performance in various metrics. The ability to leverage annotated examples allows for significant enhancements in scene classification, direction of the car, and agent classification. This aligns with the general observation that models benefit from specific, contextually relevant examples to improve their predictive capabilities.

Self-ensemble learning provides a robust alternative by stabilizing performance across different candidate predictions, showcasing its strength in minimizing errors and achieving consistent results. This approach is particularly useful in scenarios where model outputs can be uncertain or variable.

Zero-shot learning, while valuable for its generalization capabilities, shows limitations when handling varying frame contexts and specific hazard scenarios. The decrease in performance with additional frames indicates a need for more sophisticated methods to balance context depth with generalization.

Image-augmented learning, although effective in enhancing specific metrics, does not match the overall accuracy of few-shot or self-ensemble learning methods. This suggests that while image augmentation can improve certain aspects of model performance, it may not provide a comprehensive solution for all types of safety-critical event detection.

The results obtained from the proposed framework compared with other baselines that utilize visual-language QA for driving safety are presented in Table 2.

Table 2. Comparative performance analysis of QA frameworks across different baselines.

Method	Dataset	Accuracy
LLaVA-1.5 [36]	VQA-v2 [37]	38.5
Cube-LLM [38]	Talk2Car [30]	38.5
SimpleLLM4AD [39]	DriveLM-nuScenes [32]	66.5
Our Proposed Model	DRAMA	79

The comparative performance analysis in the table highlights the differences in how various visual-language QA frameworks perform in the context of driving safety tasks. Each method was tested on different datasets, and the results reveal significant variations in accuracy, reflecting the strengths and limitations of each approach. LLaVA-1.5 is a model that represents an advanced multimodal approach to integrating a vision encoder with an LLM fine-tuned using the VQA-v2 dataset. The model achieved a moderate accuracy of 38.5% in the driving safety context, suggesting that LLaVA-1.5 is not capable of handling all visual-language tasks well. Similarly, Cube-LLM, which was tested using the Talk2Car dataset, also achieved an accuracy of about 38.5%. The moderate performance of both models indicates that they might struggle with the real-time command interpretation in dynamic and mixed driving environments. In the case of SimpleLLM4AD, when tested on the DriveLM-nuScenes dataset, it achieved a significantly higher accuracy of about 66.5%. This suggests that SimpleLLM4AD is better optimized for driving-related tasks and is more able to understand the challenging scenarios that are closer to real-world driving conditions. However, our proposed model outperformed all the abovementioned methods, with an accuracy of about 79%. Our proposed model appears to be able to understand different driving scenarios and extract the contextual information necessary to excel in visual-language tasks related to driving safety. In addition, our MLLM model can also perform various tasks for each safety-critical event, including identifying risky scenarios, classifying different scenes, determining car direction, classifying agents, and suggesting correct actions, which, to the best of our knowledge, is the first model to do so. This performance shows the importance of domain-specific fine-tuning and training. This allows the model to better understand and respond to the unique challenges presented in autonomous driving.

7. Conclusions

The findings underscore the potential of MLLMs to advance the automated analysis of driving videos for traffic safety. The performance of different learning methods highlights the importance of choosing appropriate techniques based on specific detection requirements and available resources. Few-shot learning offers a promising avenue for improving hazard detection accuracy and adaptability in real-world scenarios. The few-shot model consistently outperformed other learning techniques, achieving the highest overall accuracy (about 67%), “is_risk” accuracy (78%), scene classification accuracy (65%), direction of car accuracy (82%), and agent classification accuracy (68%). This demonstrates its superior effectiveness across various tasks.

Future research should explore the integration of these methodologies to leverage their complementary strengths. Combining few-shot learning with self-ensemble or image-augmented techniques might provide a balanced approach that enhances overall performance while addressing the limitations observed in individual methods. Additionally, fine-tuning MLLMs on task-specific data is a crucial area for future investigation. Fine-tuning could enhance model performance by adapting the pre-trained models to the nuances of safety-critical event detection, thus improving accuracy and reliability. Further exploration into optimizing frame selection and processing strategies could help refine model accuracy and efficiency.

Moreover, we plan to incorporate RAG flow in future work. This approach would enable the model to dynamically retrieve and apply relevant information, such as implicit traffic rules, during inference. Incorporating RAG could further enhance the model’s capability in handling complex traffic safety scenarios, making it more robust in detecting and managing safety-critical events. This addition to the future work underscores our commitment to advancing the effectiveness of MLLMs in autonomous driving systems.

Although the DRAMA dataset was constructed in limited geographical locations, our proposed MLLM model was tested using different scenarios, which included a variety of road scenes, such as urban and rural roads, narrow lanes, and intersections. Our proposed framework has demonstrated promising results across these different road conditions, indicating its potential robustness to be scaled and generalized in other geographical locations. Validating the framework’s ability to detect safety-critical traffic events in diverse geographical contexts is indeed a crucial step, and we plan to incorporate this in our future research.

This study demonstrates the value of MLLMs in traffic safety applications and provides a foundation for further exploration and development of automated hazard detection systems. The insights gained from this research can guide the design of more effective and reliable safety-critical event detection frameworks in autonomous driving systems.

Author Contributions: Conceptualization, M.A.T., H.I.A. and M.E.; methodology, M.A.T., H.I.A. and M.E.; software, M.A.T.; formal analysis, M.A.T.; investigation, M.A.T.; resources, M.A.T.; data curation, M.A.T.; writing—original draft preparation, M.A.T.; writing—review and editing, H.I.A., M.E., S.G. and A.R.; visualization, M.A.T.; supervision, H.I.A. and M.E.; project administration, S.G. and A.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data used in this study can be requested for research work from this link: <https://usa.honda-ri.com/drama> (accessed on 22 February 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Beiker, S. History and status of automated driving in the united states. In *Road Vehicle Automation*; Springer: Cham, Switzerland, 2014; pp. 61–70.
2. Ashqar, H.I.; Alhadidi, T.I.; Elhenawy, M.; Jaradat, S. Factors affecting crash severity in Roundabouts: A comprehensive analysis in the Jordanian context. *Transp. Eng.* **2024**, *17*, 100261. [[CrossRef](#)]
3. Eskandarian, A.; Wu, C.; Sun, C. Research advances and challenges of autonomous and connected ground vehicles. *IEEE Trans. Intell. Transp. Syst.* **2019**, *22*, 683–711. [[CrossRef](#)]

4. Butt, F.A.; Chattha, J.N.; Ahmad, J.; Zia, M.U.; Rizwan, M.; Naqvi, I.H. On the integration of enabling wireless technologies and sensor fusion for next-generation connected and autonomous vehicles. *IEEE Access* **2022**, *10*, 14643–14668. [[CrossRef](#)]
5. Bathla, G.; Bhadane, K.; Singh, R.K.; Kumar, R.; Aluvalu, R.; Krishnamurthi, R.; Kumar, A.; Thakur, R.N.; Basheer, S. Autonomous vehicles and intelligent automation: Applications, challenges, and opportunities. *Mob. Inf. Syst.* **2022**, *2022*, 7632892. [[CrossRef](#)]
6. Macioszek, E. Analysis of the Volume of Passengers and Cargo in Rail and Road Transport in Poland in 2009–2019. *Sci. J. Silesian Univ. Technology. Ser. Transp.* **2021**, *113*, 133–143. [[CrossRef](#)]
7. Faisal, A.; Kamruzzaman, M.; Yigitcanlar, T.; Currie, G. Understanding autonomous vehicles. *J. Transp. Land Use* **2019**, *12*, 45–72.
8. Raiaan, M.A.K.; Mukta, M.S.H.; Fatema, K.; Fahad, N.M.; Sakib, S.; Mim, M.M.J.; Ahmad, J.; Ali, M.E.; Azam, S. A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. *IEEE Access* **2024**, *12*, 26839–26874. [[CrossRef](#)]
9. Bai, Y.; Wu, D.; Liu, Y.; Jia, F.; Mao, W.; Zhang, Z.; Zhao, Y.; Shen, J.; Wei, X.; Wang, T.; et al. Is a 3D-Tokenized LLM the Key to Reliable Autonomous Driving? *arXiv* **2024**, arXiv:2405.18361.
10. Cui, C.; Ma, Y.; Cao, X.; Ye, W.; Wang, Z. Receive, Reason, and React: Drive as You Say, With Large Language Models in Autonomous Vehicles. *IEEE Intell. Transp. Syst. Mag.* **2024**, *16*, 81–94. [[CrossRef](#)]
11. Prabhod, K.J. Advanced Techniques in Reinforcement Learning and Deep Learning for Autonomous Vehicle Navigation: Integrating Large Language Models for Real-Time Decision Making. *J. AI-Assist. Sci. Discov.* **2023**, *3*, 1–20.
12. Masri, S.; Ashqar, H.I.; Elhenawy, M. Leveraging Large Language Models (LLMs) for Traffic Management at Urban Intersections: The Case of Mixed Traffic Scenarios. *arXiv* **2024**, arXiv:2408.00948.
13. Zeng, A.; Attarian, M.; Ichter, B.; Choromanski, K.; Wong, A.; Welker, S.; Tombari, F.; Purohit, A.; Ryoo, M.; Sindhvani, V.; et al. Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language. *arXiv* **2022**, arXiv:2204.00598.
14. Malla, S.; Choi, C.; Dwivedi, I.; Choi, J.H.; Li, J. DRAMA: Joint Risk Localization and Captioning in Driving. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 1043–1052.
15. Ashqar, H.I.; Alhadidi, T.I.; Elhenawy, M.; Khanfar, N.O. The Use of Multimodal Large Language Models to Detect Objects from Thermal Images: Transportation Applications. *arXiv* **2024**, arXiv:2406.13898.
16. Elhenawy, M.; Abutahoun, A.; Alhadidi, T.I.; Jaber, A.; Ashqar, H.I.; Jaradat, S.; Abdelhay, A.; Glaser, S.; Rakotonirainy, A. Visual Reasoning and Multi-Agent Approach in Multimodal Large Language Models (MLLMs): Solving TSP and mTSP Combinatorial Challenges. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 1894–1920. [[CrossRef](#)]
17. Cui, C.; Ma, Y.; Cao, X.; Ye, W.; Zhou, Y.; Liang, K.; Chen, J.; Lu, J.; Yang, Z.; Liao, K.D.; et al. A survey on multimodal large language models for autonomous driving. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2024; pp. 958–979.
18. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The KITTI dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237. [[CrossRef](#)]
19. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuscenes: A multimodal dataset for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11621–11631.
20. Chen, L.; Sinavski, O.; Hünemann, J.; Karnsund, A.; Willmott, A.J.; Birch, D.; Maund, D.; Shotton, J. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. *arXiv* **2023**. [[CrossRef](#)]
21. Wang, W.; Xie, J.; Hu, C.; Zou, H.; Fan, J.; Tong, W.; Wen, Y.; Wu, S.; Deng, H.; Li, Z.; et al. Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving. *arXiv* **2023**. [[CrossRef](#)]
22. Cui, C.; Ma, Y.; Cao, X.; Ye, W.; Wang, Z. Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 902–909. [[CrossRef](#)]
23. Wang, L.; Ren, Y.; Jiang, H.; Cai, P.; Fu, D.; Wang, T.; Cui, Z.; Yu, H.; Wang, X.; Zhou, H.; et al. AccidentGPT: A V2X Environmental Perception Multi-modal Large Model for Accident Analysis and Prevention. In *2024 IEEE Intelligent Vehicles Symposium (IV)*; IEEE: New York, NY, USA, 2024; pp. 472–477.
24. Zhang, A.H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; Beijbom, O. Deep LiDAR-Radar-Visual Fusion for Object Detection in Urban Environments. *Remote Sens.* **2023**, *14*, 12697–12705.
25. Singh, M.; Lopez, E.; Antunes, M.; Gomez-Huélamo, C.; de la Peña, J.; and Bergasa, L.M. Towards LiDAR and RADAR Fusion for Object Detection and Multi-object Tracking in CARLA Simulator. *SpringerLink* **2023**, *14*, 710–715.
26. Lopez, E.; Singh, M.; Gomez-Huélamo, C.; de la Peña, J.; Antunes, M.; Bergasa, L.M. Real-time Object Detection Using LiDAR and Camera Fusion for Autonomous Driving. *Sci. Rep.* **2023**, *14*, 58443–58469.
27. Kim, S.; Julier, S.J.; Uhlmann, J.K. Smartmot: Exploiting the Fusion of HD Maps and Multi-object Tracking for Real-time Scene Understanding in Intelligent Vehicles Applications. *IEEE Intell. Veh. Symp.* **2023**, *14*, 710–715.
28. Hussien, M.M.; Melo, A.N.; Ballardini, A.L.; Maldonado, C.S.; Izquierdo, R.; Sotelo, M.Á. RAG-based Explainable Prediction of Road Users Behaviors for Automated Driving using Knowledge Graphs and Large Language Models. *arXiv* **2024**, arXiv:2405.00449.
29. Liang, J.; Li, Y.; Yin, G.; Xu, L.; Lu, Y.; Feng, J.; Shen, T.; Cai, G. A MAS-based hierarchical architecture for the cooperation control of connected and automated vehicles. *IEEE Trans. Veh. Technol.* **2022**, *72*, 1559–1573. [[CrossRef](#)]

30. Deruyttere, T.; Vandenhende, S.; Grujicic, D.; van Gool, L.; Moens, M.-F. Talk2Car: Taking Control of Your Self-Driving Car. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 2088–2098. [[CrossRef](#)]
31. Qian, T.; Chen, J.; Zhuo, L.; Jiao, Y.; Jiang, Y.G. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 4542–4550.
32. Sima, C.; Renz, K.; Chitta, K.; Chen, L.; Zhang, H.; Xie, C.; Luo, P.; Geiger, A.; Li, H. Drivelm: Driving with graph visual question answering. *arXiv* **2023**, arXiv:2312.14150.
33. Wu, D.; Han, W.; Wang, T.; Liu, Y.; Zhang, X.; Shen, J. Language prompt for autonomous driving. *arXiv* **2023**, arXiv:2309.04379.
34. Kim, J.; Rohrbach, A.; Darrell, T.; Canny, J.; Akata, Z. Textual explanations for self-driving vehicles. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 563–578.
35. Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A.M.; Hauth, A.; et al. Gemini: A family of highly capable multimodal models. *arXiv* **2023**, arXiv:2312.11805.
36. Liu, H.; Li, C.; Li, Y.; Lee, Y.J. Improved baselines with visual instruction tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 26296–26306.
37. Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; Parikh, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6904–6913.
38. Cho, J.H.; Ivanovic, B.; Cao, Y.; Schmerling, E.; Wang, Y.; Weng, X.; Li, B.; You, Y.; Krähenbühl, P.; Wang, Y.; et al. Language-Image Models with 3D Understanding. *arXiv* **2024**, arXiv:2405.03685.
39. Zheng, P.; Zhao, Y.; Gong, Z.; Zhu, H.; Wu, S. SimpleLLM4AD: An End-to-End Vision-Language Model with Graph Visual Question Answering for Autonomous Driving. *arXiv* **2024**, arXiv:2407.21293.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.