2025-01-8213 Published 01 Apr 2025

Avoiding the Crash: A Vision-Language Model Evaluation of Critical Traffic Scenarios

David Fernandez, Pedram MohajerAnsari, Amir Salarpour, and Mert D. Pesé Clemson University

Citation: Fernandez, D., MohajerAnsari, P., Salarpour, A., and Pesé, M.D., "Avoiding the Crash: A Vision-Language Model Evaluation of Critical Traffic Scenarios," SAE Technical Paper 2025-01-8213, 2025, doi:10.4271/2025-01-8213.

Received: 22 Oct 2024 Revised: 24 Dec 2024 Accepted: 20 Jan 2025

Abstract

Autonomous Vehicles (AVs) have transformed transportation by reducing human error and enhancing traffic efficiency, driven by deep neural network (DNN) models that power image classification and object detection. However, to maintain optimal performance, these models require periodic re-training; failure to do so can result in malfunctions that may lead to accidents. Recently, Vision-Language Models (VLMs), such as LLaVA-7B and MoE-LLaVA, have emerged as powerful alternatives, capable of correlating visual and textual data with a high degree of accuracy. These models' robustness and ability to generalize across diverse environments make them especially suited to analyzing complex driving scenarios like crashes. To evaluate the decision-making capabilities of these models across common crash scenarios, a set of real-world crash incident videos was collected. By decomposing these videos into frame-by-frame images, we task the VLMs to determine the appropriate driving action at each frame: accelerate, brake, turn left, turn right, or maintain the current course. For each frame, three sets of outputs are analyzed: the actual action executed in the video, the action a human driver would likely take to avoid a crash, and the action the VLM predicts as optimal to avoid a crash. To measure and compare the effectiveness of the VLMs, we introduce a metric called Crash Prevention Efficiency (CPE) which evaluates the model's performance in detecting crash scenarios and taking appropriate actions to avoid them. CPE assesses how well a VLM can respond to potential crashes by analyzing both the timing of the detection and the proximity to a predefined point in the crash sequence. Our findings reveal that VLMs demonstrate a high level of consistency in decision-making, with LLaVA-7B and MoE-LLaVA models identifying potential crash scenarios 1.13 to 1.33 seconds earlier than humans, respectively. This highlights their potential role in autonomous driving systems (ADS), supporting both real-time decision-making for human drivers and fully autonomous operations.[1]

[1] The first two authors contributed equally and are ordered alphabetically.

1. Introduction

In recent years, advancements in DNN models have transformed transportation using AVs by improving tasks such as image classification and object detection [58, 25]. To achieve autonomous driving capabilities, AVs rely on modular systems known as automated driving systems (ADS). These systems integrate the perception, planning, and actuation modules to facilitate independent navigation [46, 4]. The perception module relies on sensors, such as cameras and LiDAR, to capture RGB video and 3D point clouds, using DNN models to detect, track, and predict the movements of nearby objects [49, 65, 34, 35, 38].

Despite their effectiveness, DNN models require periodic retraining to maintain reliability and adapt to changing environments [6]. As the operational conditions of AVs evolve and new data become available, the models may become less accurate in detecting and responding to new or unforeseen scenarios, risking data-distribution shifts and performance decline without regular updates [20, 12]. Such degradation can lead to critical system failures, potentially resulting in safety risks, including accidents or loss of vehicle control. While retraining is essential for maintaining AV performance and safety, it presents challenges, such as high computational costs, the need for large and diverse datasets, and the rigorous validation required to ensure model reliability [5]. Additionally, DNN models, even with retraining, may struggle with unpredictable situations outside their training scope, leading to potential failures in real-world scenarios [27].

In vehicles with partial automation (SAE Levels 1 to 3), drivers remain responsible for most driving tasks, as features such as adaptive cruise control or lane-keeping assist provide only partial support, requiring them to monitor their surroundings and respond to unexpected situations [36]. When an emergency arises, such as a

sudden obstacle, loss of control, or a potential crash, drivers often experience stress, which can cloud judgment and lead to impulsive reactions, increasing the risk of collisions [30]. Statistics reveal that human error contributes to more than 94% of road accidents, with common causes including distraction, misjudgment, and delayed reactions [26]. This highlights the risks inherent in human-driven vehicles, especially when quick and accurate decision-making is required to avoid crashes.

Recently, Large Language Models (LLMs) have gained significant attention for their ability to emulate human-like intelligence [53], sparking increased interest in LLMs. By combining LLMs' advanced reasoning with visual data, VLMs enhance performance in tasks such as text-to-image alignment and image classification [61]. Furthermore, studies show that LLMs can be applied to a variety of robotics tasks, from logical and geometrical reasoning to complex operations including aerial navigation and controlling embodied agents [56, 44]. The strengths of VLMs lie in their multimodal learning, allowing them to process visual and textual information simultaneously [31]; their ability to generalize across different environments, supported by training on diverse datasets [43]; their improved contextual understanding for interpreting visual scenes using textual data [66]; and their capacity for comprehensive feature extraction by integrating features from both visual and textual inputs [8].

We conducted an experiment using ten crash videos to evaluate VLMs' ability to handle challenging driving scenarios. The videos included instances where human drivers made poor decisions under pressure and cases where AVs failed to detect obstacles, road signs and other vehicles due to rare, unforeseen conditions or data-distribution shifts. The goal was to assess how well VLMs could determine the appropriate driving actions in these situations. These included five cases involving Tesla vehicles at SAE Level 2 automation, where AVs were at fault, as illustrated in Figure 1, and five cases involving non-automated vehicles, where human drivers were unable to prevent crashes. In the experiment, VLMs determined the appropriate driving action for each video frame, choosing from five options: acceleration, braking, turning left, turning right, or maintaining the current course. We evaluated the driving decisions from three perspectives: First, by examining the actions actually taken in the video; second, by considering what a human driver could have done to avoid the crash; and third, by analyzing the actions suggested by the VLM, which suggested the best course of action to avoid a collision.

To evaluate the effectiveness of VLMs in predicting optimal driving actions, we defined the CPE metric, which measures how well the model can detect potential crashes and respond appropriately in terms of timing and proximity to the crash point. The CPE metric emphasizes the timing and effectiveness of responses at critical moments. Our results revealed that VLMs consistently outperformed both human drivers and existing AV systems in crash scenarios, predicting actions that could more reliably and efficiently prevent accidents. This comparative analysis demonstrates that VLMs surpass both current human-driven and automated systems in critical driving situations, providing a more effective approach to accident prevention.

This paper makes the following contributions:

- The study compares the effectiveness of VLMs (LLaVA-7B [32] and MoE-LLaVA [29]) and human drivers in crash scenarios, showing that VLMs prove better performance by making more timely decisions that could have prevented accidents. This analysis provides evidence of VLMs' superior capability in high-pressure driving situations.
- The research evaluates VLMs against current AV systems in terms of crash-avoidance decision-making, finding that VLMs offer safety and reliability by more precisely detecting and reacting to potential collisions. The results indicate that VLMs have the potential to surpass existing ADS in accident prevention.
- We introduce the *CPE* metric, which measures how effectively VLMs, human drivers, and AV systems detect crash scenarios and take actions based on the timing of detection and the proximity to the crash location. This metric allows for a direct comparison of performance across different scenarios.

2. Related Work

Recent studies have highlighted the potential of LLMs to enhance various aspects of AVs, particularly perception, motion control, and motion planning. In terms of perception systems, LLMs leverage external APIs to access real-time information sources, such as traffic reports, significantly improving the vehicle's understanding of its environment [15, 16]. Aldeen et al. [3] explored the application of Large Multimodal Models (LMMs) to bolster AV cybersecurity. For motion control, Song et al. [48] proposed HUDSON, an LLM-based driving agent that enhances decision making during perception attacks by identifying inconsistencies in real-time data, thereby improving attack detection and avoidance compared to legacy systems.

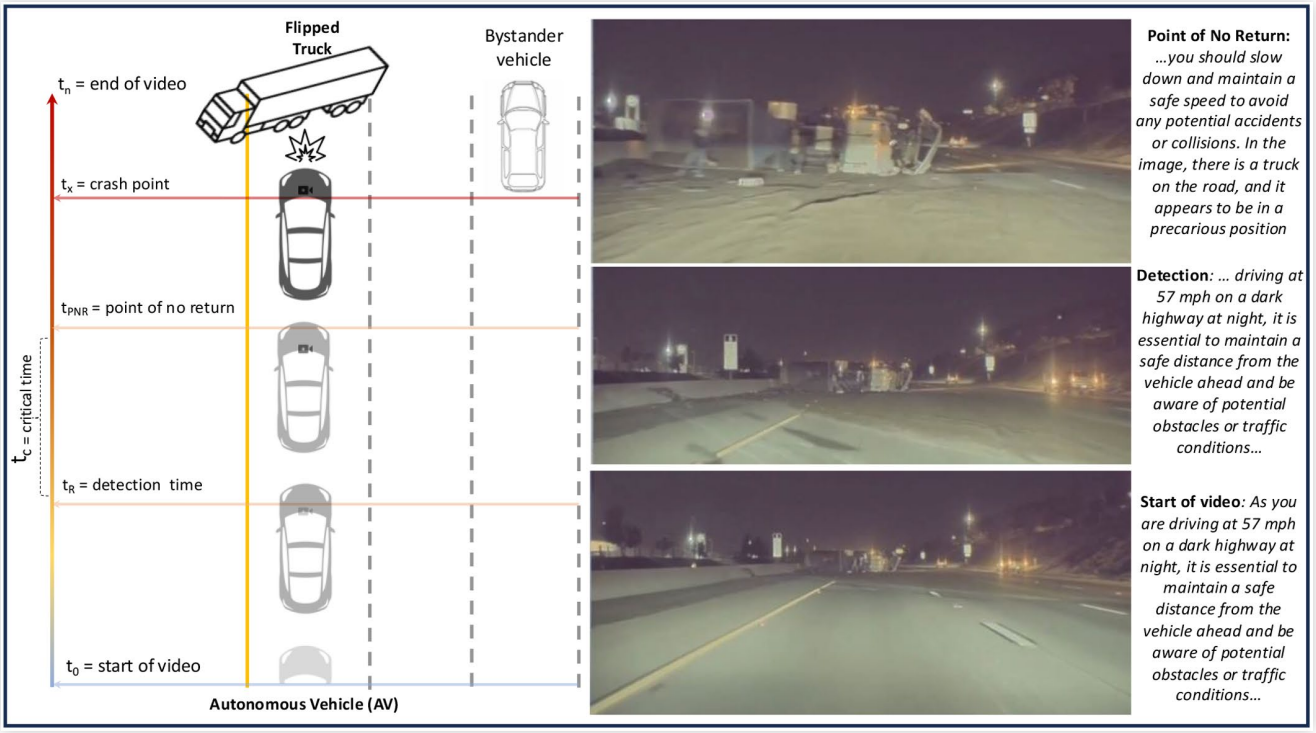
Additionally, LLMs improve transparency by offering detailed explanations for each step of the motion control process. Their capabilities enhance navigation by effectively analyzing real-time traffic data to pinpoint congested routes and recommend alternative paths, thereby optimizing navigation safety and efficiency [50]. In motion planning, LLMs utilize their advanced natural language understanding and reasoning abilities [33], facilitating user-friendly communication that allows passengers to express their intentions and preferences in natural languages.

3. Background

3.1. Advancements in Autonomous Driving System

The Society of Automotive Engineers (SAE) has categorized Autonomous Driving Systems (ADS) into five levels

FIGURE 1 AV Crash of Fail to Stop at Accident (AVC-FTSA) scenario with the MoE-LLaVA model description. In this scenario, the AV fails to recognize a flipped truck and nearby pedestrians blocking the road, resulting in a delayed reaction and causing a secondary collision.



ranging from Level 0 (driver assistance) to Level 5 (full autonomy) [19]. Levels 1 and 2 incorporate essential driver-assistance features, such as lane departure warnings and automatic lane centering (ALC), primarily based on camera-based systems. However, these levels' limited range of sensors may introduce safety hazards, particularly if the sensor data are compromised or malfunction. Level 3 represents a significant advancement because the vehicle can autonomously manage most driving tasks while requiring the driver to take control when necessary. Level 4 advances automation further, allowing vehicles to operate without human input in most situations. However, they are usually confined to specific geographic areas or conditions due to regulatory and infrastructure constraints. Ultimately, Level 5 signifies full autonomy, enabling vehicles to navigate any road or environment without human involvement.

3.2. Evolution of Multimodal Models

Between 1980 and 2020, computational models evolved from basic single-modality systems to sophisticated multimodal technologies [59]. In the 1980s and 1990s, statistical algorithms improved face and speech recognition, with developments such as Eigenfaces enhancing facial recognition accuracy [7, 45, 28]. By the 2000s, the focus shifted to human-computer interaction, exemplified by projects like AMI and CALO, which advanced multimedia data analysis and contributed to early virtual assistants including Siri [41, 54, 11]. From 2010 to 2020, the field saw groundbreaking progress in modality fusion, driven by advancements in deep learning. Notably, Ngiam et al. [37] introduced a multimodal deep learning algorithm in 2011, enhancing the processing of diverse data types such as images and text. Moreover, the advent of a neural image captioning algorithm with semantic attention in 2016 transformed image processing and description, enabling applications from automatic tagging to assistive technologies for the visually impaired [62].

Since 2020, advances in LLMs systems have transformed the AI landscape. Models such as GPT-3 [10], PaLM [14], LLaMA [52], and GPT-4 [1] have utilized extensive training on large datasets to enhance text generation and cognitive skills, including in-context learning (ICL) [10] and chain-of-thought (CoT) reasoning [57].

Concurrently, multimodal models have progressed, as exemplified by CLIP's [39] impact on image-text pair analysis and DALL-E 2's [40] ability to create images from text prompts. In 2023, Microsoft's BEiT-3 [9] and KOSMOS-1 [22] advanced the integration of sensory data and language, highlighting a shift towards AI systems that replicate human-like capabilities across various applications.

3.3. DNNs and VLMs for Crash Prevention

DNNs have demonstrated remarkable success in various fields, including computer vision, natural language

processing, audio signal processing, and cross-modal applications [13, 60] However, VLMs have emerged as powerful tools for tasks involving visual data.

VLMs are advanced Neural Networks capable of understanding and processing visual data such as images and videos [23]. VLMs have demonstrated high performance in several computer vision tasks like object detection, segmentation, and image classifications [18]. The VLMs multi-modal capability allows them to interpret intricate driving environments, enhancing crash prevention systems.

VLMs are capable of generalizing without explicit training. This flexibility contrasts with DNNs which usually require retraining or finetuning for each specific task and often struggle with generalization across different domains. The integration of visual and textual information allows VLMs to generalize better across different domains compared to DNNs, which are trained on a single modality [64].

4. Limitations in Crash Scenario Decision-Making

In critical driving situations, the ability to make accurate decisions can be the difference between life or death. Human drivers often struggle under pressure due to cognitive limitations and distractions, while AVs face difficulties when reacting to unfamiliar scenarios. VLMs offer a promising alternative by leveraging their ability to process multi-modal data, potentially surpassing human and DNN-based decision-making systems. This section explores the limitations of human drivers and traditional AV approaches and presents three research questions.

4.1. RQ1: Can VLMs Overcome Human Driver Limitations?

Human drivers often face significant challenges when making decisions during high-stress car crash situations, as these moments demand immediate responses that can be heavily influenced by emotional and psychological factors [47]. Distractions, whether external, such as mobile devices, or internal, such as anxiety, can impair a driver's ability to respond appropriately, leading to delayed reactions or poor judgment, such as braking too late or executing an incorrect avoidance maneuver [30]. Research indicates that human error is a major contributor to road accidents, underscoring the limitations of decision-making under pressure [21].

In contrast, VLMs have the potential to enhance decision-making in emergencies, particularly when accurate responses are critical, as they can efficiently process vast amounts of data—unlike human drivers who may be affected by distractions and the stress of immediate decisions [63]. This capability may allow them to analyze traffic conditions, vehicle dynamics, and environmental factors without the limitations faced by humans. For instance, while human drivers might hesitate or misinterpret the positions and speeds of other vehicles, VLMs could evaluate these parameters, potentially predicting crash scenarios and determining the optimal course of action—whether to accelerate, brake, or change direction. This raises a critical question: Can VLMs provide a level of decision-making that consistently surpasses that of human drivers in preventing crashes?

4.2. RQ2:Can VLMs Overcome DNN-Based AV Systems Limitations?

AVs have advanced significantly in recent years, primarily because of the development of DNN models that underpin their decision-making capabilities [58]. However, these systems are not without limitations. A pressing challenge is the requirement for periodic retraining to maintain its effectiveness [6]. As the operational environments of AVs evolve, it is essential to retrain models with updated data to ensure reliable performance. This retraining process can be resource-intensive, requiring significant computational power and time, as well as access to high-quality and diverse datasets that accurately reflect current driving conditions [5].

Also, DNN models often struggle with scenarios they have not encountered during training, and this inability to generalize to unseen situations can lead to critical failures, particularly in edge cases that fall outside the model's training data [27]. For instance, a DNN-based AV may perform well in familiar urban settings but could falter when faced with unique circumstances, such as unexpected obstacles, rare traffic situations, or unusual road configurations. These limitations highlight a fundamental issue: Reliance on DNNs can create performance gaps in novel scenarios, increasing the risk of accidents.

By contrast, VLMs could offer a more robust solution for decision-making in unseen crash scenarios. Their multimodal capability to process visual and textual information enables better contextualization and understanding of unique driving situations, allowing for effective predictions and reactions even under unfamiliar conditions. Unlike DNNs, VLMs are trained on extensive datasets with diverse inputs, making them more adaptable to different environments and able to generalize well to novel situations [43, 31]. Their ability to incorporate contextual understanding from textual data further enhances the interpretation of complex visual scenes, leading to improved decision-making in rare or unusual circumstances [66]. This combination of comprehensive training and multimodal learning makes VLMs particularly suitable for crash scenarios, where accurate decisions are crucial to avoid collisions. This raises the question: *Can VLMs outperform traditional DNN-based models in improving decision-making and safety during unseen crash scenarios?*

4.3. RQ3: How Do Conventional Metrics Fall Short in Evaluating Crash Scenarios?

In the evaluation of AVs for classification tasks, several metrics, such as accuracy, precision, recall, and F1-score, are commonly used. While these metrics provide valuable insights into overall model performance, they often fall short in crash prevention scenarios due to their inability to account for the timing and contextual relevance of decisions [2]. These metrics mainly focus on general performance rather than the specific demands of high-stakes decision-making. For instance, accuracy does not account for how timely a model's response is during critical moments, which is crucial in crash situations where rapid actions can significantly affect outcomes. Similarly, metrics including precision and recall may not effectively evaluate a model's performance in rare or unusual edge cases, leading to potential shortcomings in real-world scenarios [55]. This is because traditional metrics measure the frequency of correct predictions without considering the temporal urgency or situational context in which those predictions are made, which can be crucial for avoiding crashes.

Additionally, existing metrics, such as mean time to collision (MTTC), offer limited insight into the effectiveness of a model's actions in preventing crashes, as MTTC primarily measures the remaining time before a collision but does not evaluate whether the actions taken by the model were timely enough to effectively avoid the crash. This limitation highlights the need for a more comprehensive evaluation framework that captures not only prediction accuracy but also the critical timing of decisions. To address these gaps, a new metric is needed to assess the performance of VLMs, human drivers, and AVs in crash scenarios. This metric should emphasize the timing of detection and the actions taken relative to critical points in a driving situation, providing a clearer understanding of how these systems perform under pressure and how they can be improved for enhanced safety. *What criteria should be included in a new metric to effectively compare the decision-making performance of VLMs, human drivers, and AVs in crash scenarios?*

5. Evaluation of VLMs Performance

In this section, we address the research questions by presenting the method used to evaluate the decision-making capabilities of VLMs, humans drivers and DNN-based AV systems. We begin by describing the dataset used in our analysis, which includes real-world crash scenarios involving both autonomous and human-driven vehicles. Next, we present the analysis performed to answer **RQ1** and **RQ2**. Finally, we present the a novel metric called CPE to evaluate the performance of VLMs and address **RQ3**.

To compare the decision-making capabilities of VLMs with those of humans and DNNs, we compiled a dataset of ten real-world crash videos. These scenarios cover a range of situations that challenge both autonomous and human-driven vehicles, enabling an in-depth comparison of decision-making performance. The videos in the dataset are classified in two categories:

- **AV Crashes (AVC)**: This category includes videos of crashes involving AVs, that rely on DNNs for driving decisions. The videos were obtained from a Wall Street Journal report [51].
- **Human-driven Vehicle Crashes (HDC)**: This category features crashes involving human drivers. The videos were scraped from the Instagram account "dashcam.nation" which posts dashcam videos recorded in North America and has over 200,000 followers [24].

We used a Python script to extract a still image from each video at intervals of every five frames. Although the original frame rate of the videos was 30 fps, we opted for five-frame intervals to balance capturing enough detail about the crash scenario with the need to manage computational resources efficiently. Through experiments, we confirmed that increasing the frame extraction rate did not yield additional data and only resulted in longer processing times and increased storage demands.

For each video, the speed of the vehicle and the distance traveled within the video were calculated. These calculations relied on the standard road markings specified by the Federal Highway Administration (FHWA) in the Manual on Uniform Traffic Control Devices (MUTCD). According to this manual, each dashed lane marking is approximately 3.05 m long, with a 9.14 m gap between each marking, resulting in a total distance of 12.19 m between the start of a lane marking and the start of the next [17]. By measuring the time it takes for a vehicle to cover this 40-foot distance, the speed can be calculated. Using this speed and the total duration of the video, the total distance traveled is then determined. The scenarios are presented in Table 1.

5.1. RQ1: Comparing VLM and Human Driver Responses

To analyze and compare the performance of humans and VLMs, we considered two metrics: the Point of no Return (PNR) and detection times. We define the PNR as the moment when a vehicle reaches a position from which it can no longer avoid a collision. It depends on the speed, distance and the time available to take evasive action. Detection time, on the other hand, measures how quickly a human driver or VLM identifies a potential crash situation, which is crucial for initiating timely corrective actions.

We compared the decision-making performance of human drivers and VLMs, specifically the LLaVA-7B and

TABLE 1 Summary of the collected dataset of crash scenarios.

Category	Scenario	Description	Speed (m/s)	Video Duration (s)	Frame Count	Distance (m)
AVC	Fail to Stop at Accident (AVC-FTSA)	Secondary collision after failing to stop.	25.43	6.00	35	152.63
	Off-Road Stop (AVC-ORS)	Vehicle halts off the road unexpectedly.	24.42	6.00	34	146.48
	Lane Veering (AVC-LV)	Vehicle drifts out of its lane.	22.76	2.00	9	45.74
	Fail to Stop at Intersection (AVC-FTSI)	Misses stopping at a busy intersection.	1.51	4.00	25	6.09
	Fail to Stop Sudden Brake (AVC-FTSSB)	Rear-end collision after sudden brake.	20.00	2.00	12	40.23
HDC	Fail to Stop on Merge (HDC-FTSM)	Collision due to improper merging.	21.91	10.00	50	219.78
	Fail to Yield at Cross-Traffic (HDC-FYCT)	Crash from not yielding to cross-traffic.	23.06	11.00	55	254.32
	Fail to Brake on Time (HDC-FBT)	Delayed braking leads to impact.	22.35	10.00	51	223.47
	Red Light Violation (HDC-RLV)	Runs red light, causing a collision.	3.36	9.00	44	6.06
	Unsafe U-Turn (HDC-UUT)	Unsafe U-turn results in crash.	8.05	11.00	52	16.76

MoE-LLaVA models. The setup we used for inference consisted of a system equipped with 62 GB of RAM and an Intel 13th Gen Core i9-13900KF processor, featuring 32 CPU threads (24 cores) and a maximum clock speed of 5.8 GHz. This configuration provided sufficient computational resources for the VLM inference tasks. We used the stock version of the LLaVA-7B and MoE-LLava models without fine-tuning to ensure results reflected baseline performance.

To analyze the PNR in the HDC scenarios of our dataset, we calculated the differences in detection times between humans and VLMs, as summarized in Table 2. Our results show that, on average, humans identified the PNR 0.88 seconds later than the LLaVA-7B model and 0.81 seconds later than the MoE-LLaVA model.

Regarding the detection of potential crash scenarios, the models generally outperformed humans drivers, except for the **HDC-FBT** scenario using MoE-LLaVA. On average, humans were 1.13 seconds slower than the LLaVA-7B model and 1.33 seconds slower that the MoELLaVA model. For example, in the **HDC-FTSA** scenario (Figure 2), both models not only detected the potential crash earlier than the human driver but also recommended appropriate actions. These recommendations closely matched the expected responses at optimal detection points and PNR, showcasing the models' ability to provide timely and contextually relevant guidance during potential crash situations.

TABLE 2 Comparison of detection times between humans and VLMs (LLaVA-7B and MoE-LLaVA) for HDC scenarios.

Scenario	Δ PNR Time (Human - LLaVA-7B) [s]	Δ PNR Time (Human - MoE-LLaVA) [s]	Δ Detection Time (Human - LLaVA-7B) [s]	Δ Detection Time (Human - MoE-LLaVA) [s]
HDC-FTSM	1.55	1.22	1.40	2.57
HDC-FYCT	0.90	0.90	1.68	1.68
HDC-FBT	1.43	1.43	0.30	-0.20
HDC-RLV	0.44	0.44	1.20	1.37
HDC-UUT	0.08	0.08	1.08	1.25
Average	**0.88**	**0.81**	**1.13**	**1.33**

These findings suggest that VLMs possess a speed advantage in processing driving-related cues, which could be crucial for decision-making in high-pressure situations where even small time differences can significantly affect the outcome of a crash.

5.2. RQ2: Comparing VLM and AV Driver Responses

Next, we examined the frames where VLMs identified the PNR and potential crash situations and compare them to the *optimal frame*. We define the optimal frame as the ideal moment for detecting a possible crash situation. We evaluated the actions suggested by the VLMs to assess not only whether they could correctly identify when a dangerous situation was detected but also the appropriateness of their suggested actions under these conditions.

As shown in Table 3 and Table 4, our results reveal that not only did the VLMs consistently identify both the optimal frame and PNR earlier than in our manual analysis, but also suggested correct actions to be taken. For example, in the **AVC-FTSA** scenario, as shown in Figure 1, the MoE-LLaVA model detected the PNR at frame 21, compared to the optimal frame of 25. More importantly, unlike the AV system, the VLM was able to recognize the presence of a flipped truck on the road. This highlights one of the critical advantages of VLMs over traditional DNNs: The VLMs could identify a hazardous action and suggest an appropriate action without being explicitly trained on it. The model suggested slowing down or stopping, demonstrating its ability to provide timely and relevant guidance in complex, unforeseen crash situations.

FIGURE 2 Human-driven Vehicle Crash Red Light Violation (HDC-RLV) scenario with the MoE-LLaVA model description. In this scenario, the human driver accelerates the light turns green, but another vehicle runs the red light and makes a turn, causing a collision. The human driver fails to react in time to avoid the crash.

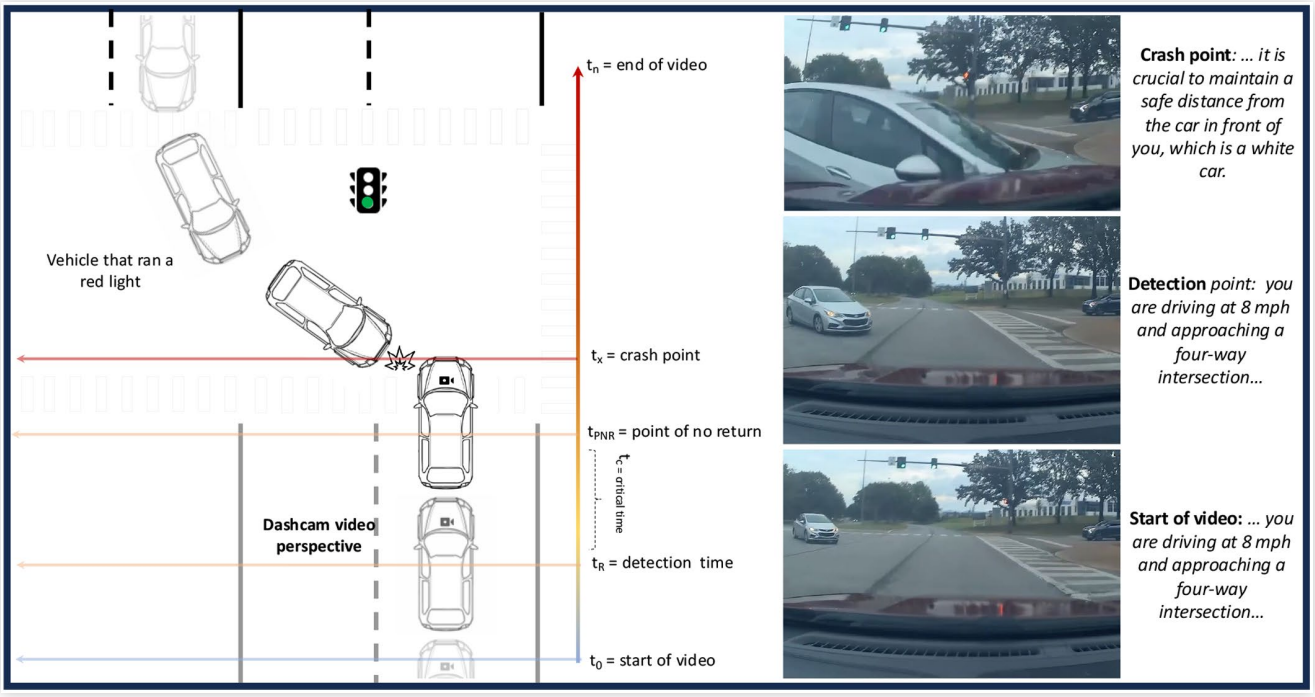


TABLE 3 Detection frame comparison for AVC scenarios using LLaVA-7B and MoE-LLaVA.

Scenario	Total Frames	Detection of Optimal Frame	Detection Frame LLaVA-7B	LLaVA-7B Description	Detection Frame MoE-LLaVA	MoE-LLaVA Description
AVC-FTSA	35	12	2	...there is a truck in front of you on the highway...	3	maintain a safe distance from the vehicle ahead
AVC-ORS	34	6	5	...there is a stop sign ahead	4	...there is a stop sign ahead
AVC-LV	9	3	1	...there is a sign on the road that says "clear keep right"	1	There is a car...driving on the same lane as you
AVC-FTSI	25	4	3	...there is a stop sign... a car is visible in the background	3	..there is a pedestrian crossing sign
AVC-FTSSB	12	4	6	...there are several cars on the road...	6	truck is positioned ahead of you prepare for sudden braking...

TABLE 4 PNR detection for AVC scenarios using LLaVA-7B and MoE-LLaVA.

Scenario	Total Frames	Optimal PNR	PNR LLaVA-7B	LLaVA-7B Description	PNR MoE-LLaVA	MoE-LLaVA Description
AVC-FTSA	35	25	22	there is a truck in front of ...you should slow down	21	...there is a truck on the road ... in a precarious position...
AVC-ORS	34	13	12	...you should come to a complete stop	12	...you need to come to a complete stop
AVC-LV	9	5	5	...there is a car in front of you	5	car ahead of you, and you should react
AVC-FTSI	25	10	6	...you should come to a complete stop...	4	...slow down and come to a complete stop at the crosswalk
AVC-FTSSB	12	6	7	slow down and exercise caution...	9	...maintain a safe distance from the truck ahead

5.3. RQ3: Crash Prevention Efficiency (CPE)

In this section, we introduce a new metric called *CPE* to evaluate the performance of VLMs in detecting crash scenarios and taking the appropriate actions to prevent crashes. The proposed metric measures how well a VLM can respond to a potential crash by assessing both the timing of the detection and the proximity to a predefined PNR. The PNR is defined as the latest moment when action must be taken to avoid a crash. It is determined by the vehicle's speed and distance from the crash location.

The following terms are defined for clarity: v represents the vehicle's speed at a given frame, and d is the distance from the vehicle to the crash point. For this calculation, we assume v remains constant over the time interval. Based on these two factors, the time of the PNR t_b can be calculated as:

$$t_b = \frac{d}{v} \tag{1}$$

where t_b is the maximum allowable time to take action before the vehicle reaches the crash point. This represents the time before reaching the crash point where the VLM must act to prevent a collision. The time at which the VLM detects the crash is denoted as t_d, and the time required to react after detection is t_r. The total time for detection and reaction, called the critical time t_c, is given by:

$$t_c = t_d + t_r \tag{2}$$

If $t_c \leq t_b$, then the VLM successfully takes action before reaching the PNR. We define the time margin Δt, which represents the remaining time before the PNR when the VLM has completed its action:

$$\Delta t = t_b - t_c \tag{3}$$

Using these definitions, we now introduce the formula for the *CPE*. This metric evaluates the efficiency of the VLM's detection and action response relative to the PNR. The CPE is calculated as:

$$CPE = \begin{cases} e^{-\alpha(t_b - t_c)}, & \text{if } t_c < t_b \ (\text{Early Action}) \\ 1, & \text{if } t_c = t_b \ (\text{On-Time Action}) \\ -e^{-\beta(t_c - t_b)}, & \text{if } t_c > t_b \ (\text{Late Action}) \end{cases} \tag{4}$$

Where α is a scaling factor for early actions, controlling how quickly the score approaches 1 as the action gets closer to the PNR. β is a scaling factor for late actions, designed to make the score approach -1 more quickly as the action occurs further past the PNR. Here, $\beta > \alpha$ is selected to emphasize the critical nature of late actions.

The CPE metric evaluates the efficiency of the VLM's response based on the timing of its action relative to the PNR:

- Early Action ($t_c < t_b$): When the VLM acts before the PNR, the CPE value falls within the range [0, 1]. As t_c gets closer to t_b, the CPE approaches 1, indicating a highly effective response. If the action occurs much earlier than the PNR, the CPE moves toward 0, indicating a less effective (premature) action.
- On-Time Action ($t_c = t_b$): When the VLM acts exactly at the PNR, the CPE is 1, representing the ideal response.
- Late Action ($t_c > t_b$): When the VLM acts after the PNR, the CPE falls within the range [-1, 0]. As t_c just passes t_b, the CPE starts near -1, indicating a relatively effective but slightly late action. As t_c moves further away from t_b, the CPE approaches 0, reflecting increasingly poor performance due to the significant delay in taking action.

In this approach, CPE values close to 1 (early) or -1 (late) suggest effective actions taken near the PNR, whereas values close to 0 (either before or after) indicate poor timing and minimal effectiveness.

Next, we provide the mathematical proof of the metric's properties.

First, if $t_c < t_b$, the term $\alpha(t_b - t_c)$ is positive, and the CPE is given by:

$$CPE = e^{-\alpha(t_b - t_c)} \to 1 \quad \text{as} \quad t_c \to t_b \tag{5}$$

This demonstrates that the CPE approaches 1 when the VLM detects and acts just before the PNR, indicating a highly efficient response. For actions taken much earlier than the PNR, the CPE moves toward 0, indicating less effectiveness.

Now, if $t_c = t_b$, we have:

$$CPE = 1 \tag{6}$$

This shows that the metric evaluates to 1 when the VLM acts exactly at the PNR, representing an ideal response.

Finally, if $t_c > t_b$, the term $\beta(t_c - t_b)$ becomes positive, and the CPE is given by:

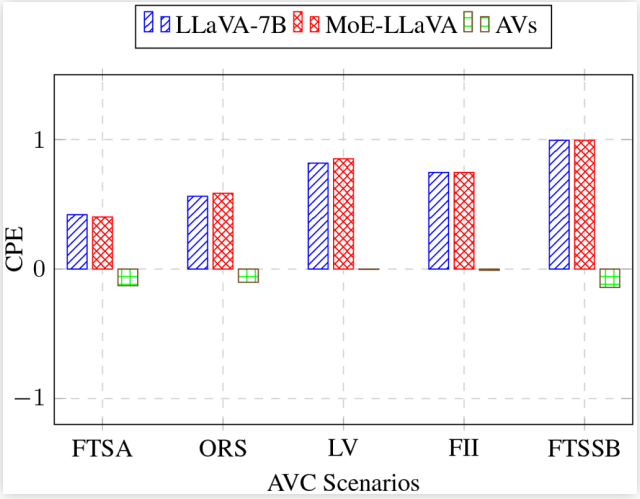
$$CPE = -e^{-\beta(t_c - t_b)} \to -1 \quad \text{as} \quad t_c \to t_b \tag{7}$$

This demonstrates that the CPE starts from -1 when the VLM takes action just after the PNR, indicating a delayed but still relatively effective response. As t_c increases further from t_b, the CPE approaches 0, reflecting increasingly poor performance due to significant delays in taking action.

Figure 3 compares the CPE values of VLMs (LLaVA-7B and MoELLaVA) with AVs across various crash scenarios. The results indicate that VLMs consistently achieve higher CPE scores than AVs, suggesting more effective crash prevention. In the **AVC-FTSA** scenario, LLaVA-7B and MoE-LLaVA achieved CPE scores of 0.419 and 0.402, respectively, while the AV recorded -0.129, reflecting a delayed response beyond the PNR. In **AVC-ORS**, the VLMs scored 0.562 and 0.585, while the AV's -0.102 again indicated a late response.

In **AVC-LV**, the VLMs outperformed the AV, with scores of 0.818 (LLaVA-7B) and 0.851 (MoE-LLaVA), compared to the AV's -0.002, showing a response past

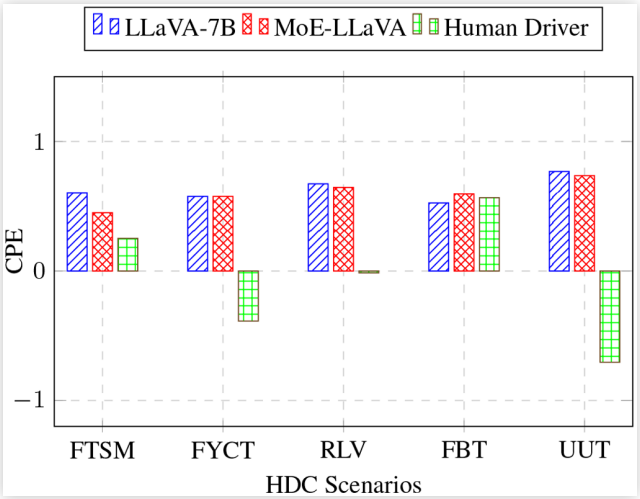
FIGURE 3 CPE Values for AVs and VLMs Across AVC Scenarios



the PNR. In **AVC-FII**, both VLMs achieved a CPE of 0.745, while the AV scored -0.010, failing to prevent the crash. In **AVC-FTSSB**, the VLMs reached near-optimal values of 0.994, while the AV's -0.142 underscored poor timing. These results show that VLMs detect crashes earlier and act closer to the PNR, while AVs often exhibit late responses, as indicated by their negative CPE scores. This finding was consistent when manually looking at the videos; in all cases, the AVs were late to act, which caused the accidents.

Figure 4 compares the CPE values for VLMs (LLaVA-7B and MoELLaVA) to those of human drivers across various high-demand crash (HDC) scenarios. The results show that VLMs consistently outperform human drivers in crash prevention. For example, in the **HDC-FTSM** scenario, LLaVA-7B achieved a CPE of 0.603 and MoE-LLaVA had 0.450, both higher than the human driver's 0.252, indicating more effective responses. In **HDC-RLV**, LLaVA-7B and MoE-LLaVA scored CPEs of 0.673 and 0.645, respectively, while the human driver's -0.015 CPE suggests a delayed response beyond the PNR.

FIGURE 4 CPE Values for Human Driver and VLMs Across HDC Scenarios



In scenarios where human drivers struggled the most, such as **HDCUUT**, VLMs maintained strong performance, with CPE values of 0.768 for LLaVA-7B and 0.736 for MoE-LLaVA, compared to the human driver's -0.705. The negative CPE indicates a delayed action past the PNR, resulting in poor crash prevention. Across all scenarios, VLMs maintain higher and more consistent CPE values, for instance, in **HDC-FYCT**, LLaVA-7B and MoE-LLaVA achieved a CPE of 0.576, while the human driver's CPE was -0.387, further demonstrating the VLMs' superior ability to anticipate and avoid collisions.

6. Limitations

While VLMs demonstrate strong potential in crash prevention scenarios, their primary limitation lies in latency. This latency arises from the significant computational resources and storage required for effective operation. VLMs typically need high-performance hardware, such as advanced GPUs and large amounts of RAM, to process complex datasets efficiently. When hardware is insufficient, processing times can increase, leading to delays that hinder timely decision-making and reduce the effectiveness of VLMs in critical applications. In contrast, advancements in AV technology, such as Tesla's HW4 system [42], illustrate how specialized hardware can enhance performance. The HW4 system is optimized for autonomous driving tasks and features 20 ARM cores, 2 GPUs, three neural network processors, and 16GB of RAM, all dedicated to handling demanding computational tasks efficiently.

In our evaluation setup, we used an Intel 13th Gen Core i9-13900KF CPU with 32 cores and 64GB of RAM, along with an NVIDIA GPU. Although our hardware configuration is high-end, it may not match the specialized capabilities of Tesla's HW4 system, which could explain some latency differences observed during our experiments. This comparison highlights the importance of specialized hardware in optimizing VLM performance for crash prevention applications. Leveraging advanced hardware can enhance the performance of VLMs in such scenarios. Additionally, the relatively small sample size of ten crash videos used in this study may only partially represent the diverse range of real-world crash situations. Future research should expand the dataset to include a broader variety of crash scenarios, road conditions, and environmental factors to improve the generalizability of the results. Additionally, this study is limited by the proprietary nature of the DNN models implemented inside the AVs, which prevents detailed architectural comparisons. These models are not disclosed due to safety and security concerns. However, the study focuses on the overall performance characteristics of these systems to ensure a fair and comprehensive evaluation.

7. Conclusion

This study demonstrates the potential of VLMs, specifically LLAVA-7B and MoE-LLaVA, in crash prevention by comparing their performance against both human drivers and AV systems. The results showed that VLMs outperformed human drivers in making more accurate decisions during crash scenarios, particularly under high-pressure and unfamiliar conditions. Additionally, VLMs surpassed traditional AV systems by detecting critical crash points earlier and taking more effective actions to prevent accidents. Although the computational demands of VLMs present a limitation, ongoing improvements in hardware and optimization techniques are expected to address these challenges, further enhancing the real-time capabilities of these models. As the field progresses, integrating VLMs into ADS could significantly reduce road accidents and improve overall traffic safety.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I. et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
2. Nacho Al, "Model evaluation metrics: Precision, recall, f1-score, and beyond," 2024, accessed: 2024-10-20.
3. Aldeen, M., Mohajer Ansari, P., Ma, J., Chowdhury, M., Cheng, L., and Pesé, M.D., "An Initial Exploration of Employing Large Multimodal Models in Defending against Autonomous Vehicles Attacks," in *2024 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 2024, 3334-3341.
4. Aldeen, M., Mohajer Ansari, P., Ma, J., Chowdhury, M., Cheng, L., and Pesé, M.D., "Wip: A First Look at Employing Large Multimodal Models against Autonomous Vehicle Attacks," in *ISOC Symposium on Vehicle Security and Privacy (VehicleSec'24)*, 2024.
5. Amazon Web Services. Modular functions design for advanced driver assistance systems (adas) on aws, 2021, accessed: 2024-10-17.
6. Attaoui, M., Fahmy, H., Pastore, F., and Briand, L., "Black-Box Safety Analysis and Retraining of Dnns Based on Feature Extraction and Clustering," *ACM Transactions on Software Engineering and Methodology* 32, no. 3 (2023): 1-40.
7. Bahl, L., Brown, P., De Souza, P., and Mercer, R., "Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition," in *ICASSP'86. IEEE international conference on acoustics, speech, and signal processing*, 11, IEEE, 1986, 49-52.
8. Bain, M., Nagrani, A., Varol, G., and Zisserman, A.. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, 1728-1738.
9. Bao, H., Dong, L., Piao, S., and Wei, F.. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.
10. Brown, T., Mann, B., Ryder, N., Subbiah, M. et al., "Language Models Are Few-Shot Learners," *Advances in Neural Information Processing Systems* 33 (2020): 1877-1901.
11. Carletta, J., Ashby, S., Bourban, S., Flynn, M. et al. The Ami Meeting Corpus: A Pre-Announcement," in *International workshop on machine learning for multimodal interaction*, Springer, 2005, 28-39.
12. Chen, J., *Robust Deep Learning under Distribution Shift* (The University of Wisconsin-Madison, 2023).
13. Cheng, H., Zhang, M., and Shi, J.Q., A survey on deep neural network pruning: Taxonomy, comparison, analysis, and recommendations. *arXiv preprint arXiv:2308.06767*, 2024.
14. Chowdhery, A., Narang, S., Devlin, J., Bosma, M. et al., "Palm: Scaling Language Modeling with Pathways," *Journal of Machine Learning Research* 24, no. 240 (2023): 1-113.
15. Cui, C., Yunsheng Ma, X., Cao, W.Y., and Wang, Z., "Receive, Reason, and React: Drive as you Say, with Large Language Models in Autonomous Vehicles," *IEEE Intelligent Transportation Systems Magazine* (2024).
16. Curran, N.T., Cho, M., Feng, R., Liu, L., Tang, B.J. et al., "Achieving the Safety and Security of the End-to-End Av Pipeline," in *Proceedings of the 2024 Cyber Security in CarS Workshop, CSCS '24*, New York, NY, USA, 2024, Association for Computing Machinery, 13-24.
17. Federal Highway Administration. Manual on uniform traffic control devices (mutcd). https://mutcd.fhwa.dot.gov/, accessed: 2024-10-17.
18. Fourati, S., Jaafar, W., Baccar, N., and Alfattani, S.. XLM for Autonomous Driving Systems: A Comprehensive Review. *arXiv preprint arXiv:2409.10484*, 2024.
19. Fraunhofer IKS. Autonomous driving. https://www.iks.fraunhofer.de/en/topics/autonomous-driving.html, accessed: 2024-09-07.
20. Ghobrial, A., Zheng, X., Hond, D., Asgari, H., and Eder, K., "Dira: Dynamic Incremental Regularised Adaptation," in *2024 IEEE Conference on Artificial Intelligence (CAI)*, IEEE, 2024, 448-455.
21. Maggie Hennessy. The 94 accessed: 2024-10-17.
22. Huang, S., Dong, L., Wang, W., Hao, Y. et al., "Language Is Not all you Need: Aligning Perception with Language Models," *Advances in Neural Information Processing Systems* 36 (2024).
23. Huang, Y.. Leveraging large language models for enhanced nlp task performance through knowledge distillation and optimized training strategies. *arXiv preprint arXiv:2402.09282*, 2024.
24. Instagram - Dashcam Nation. Dashcam nation on instagram. https://www.instagram.com/dashcam.nation/. accessed: 2024-10-15.
25. Kühr, M., Hamad, M., MohajerAnsari, P., Pesé, M.D., and Steinhorst, S., Sok: Security of the image processing pipeline in autonomous vehicles. *arXiv preprint arXiv:2409.01234*, 2024.
26. KNR Legal. Self-Driving Car Accident Statistics, 2023, accessed: 2024-10-17.

27. Li, Z., Pan, M., Zhang, T., and Li, X.. "Testing Dnn-Based Autonomous Driving Systems under Critical Environmental Conditions," in *International Conference on Machine Learning*, PMLR, 2021, 6471-6482.
28. Lien., J.J., Kanade, T., Cohn, J.F., and Li, C.-C., "Automated Facial Expression Recognition Based on Facs Action Units," in *Proceedings third IEEE international conference on automatic face and gesture recognition*, IEEE, 1998, 390-395.
29. Lin, B., Tang, Z., Ye, Y., Cui, J. et al., Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024.
30. Louisiana Transportation Research Center. Human factors in roadway safety, 2023, accessed: 2024-10-17.
31. Zhou, L., "A Theory of Multimodal Learning," *Advances in Neural Information Processing Systems* 36 (2024).
32. Luo, G., Huang, M., Zhou, Y., Sun, X. et al., Towards efficient visual adaption via structural re-parameterization. *arXiv preprint arXiv:2302.08106*, 2023.
33. Mao, J., Qian, Y., Zhao, H., and Wang, Y.. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*, 2023.
34. Mohajer Ansari, P., Domeke, A., de Voor, J., Mitra, A. et al., Discovering new shadow patterns for black-box attacks on lane detection of autonomous vehicles. *arXiv preprint arXiv:2409.18248*, 2024.
35. Mohammadi, M., and Salarpour, A.. Point-gn: A nonparametric network using gaussian positional encoding for point cloud classification. *arXiv preprint arXiv:2412.03056*, 2024.
36. Mukilan. The 6 autonomous driving levels explained, 2023, accessed: 2024-10-17.
37. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A.Y., "Multimodal Deep Learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 689-696, 2011.
38. Moein Peyghambarzadeh, S.M., Azizmalayeri, F., Khotanlou, H., and Salarpour, A., "Point-Planenet: Plane Kernel Based Convolutional Neural Network for Point Clouds Analysis," *Digital Signal Processing* 98 (2020): 102633.
39. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A. et al., "Learning Transferable Visual Models from Natural Language Supervision," in *International conference on machine learning*, PMLR, 2021, 8748-8763.
40. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M.. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.
41. Renals, S., Hain, T., and Bourlard, H., "Interpretation of Multiparty Meetings the Ami and Amida Projects," in *2008 Hands-Free Speech Communication and Microphone Arrays*, IEEE, 2008, 115-118.
42. AutoPilot Review. Tesla hardware 4 rolling out to new vehicles, 2023, accessed: 2024-10-20.
43. Reza, M.K., Prater-Bennette, A., and Asif, M.S.. Robust multimodal learning with missing modalities via parameter-efficient adaptation. *arXiv preprint arXiv:2310.03986*, 2023.
44. Salarpour, A., Khotanlou, H., and Mavridis, N., "Long-Term Estimation of Human Spatial Interactions through Multiple Laser Ranging Sensors," in *2014 International Conference on Robotics and Emerging Allied Technologies in Engineering (iCREATE)*, IEEE, 2014, 109-114.
45. Satoh, S., and Kanade, T., "Name-it: Association of Face and Name in Video," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, 1997, 368-373.
46. Schwarting, W., Alonso-Mora, J., and Rus, D., "Planning and Decision-Making for Autonomous Vehicles," *Annual Review of Control, Robotics, and Autonomous Systems* 1 (2018): 187-210.
47. Smith, B.W.. Human error as a cause of vehicle crashes, 2013, accessed: 2024-10-17.
48. Song, R., Ozmen, M.O., Kim, H., Muller, R., Celik, Z.B., and Bianchi, A., "Enhancing llm-based autonomous driving agents to mitigate perception attacks," *arXiv preprint arXiv:2409.14488*, 2024.
49. Song, R., Ozmen, M.O., Kim, H., Muller, R., Celik, Z.B., and Bianchi, A., "Discovering Adversarial Driving Maneuvers against Autonomous Vehicles," in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, 2957-2974.
50. Sriram, N.N., Maniar, T., Kalyanasundaram, J., Gandhi, V. et al., "Talk to the Vehicle: Language Conditioned Autonomous Navigation of Self Driving Cars," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2019, 5284-5290.
51. The Wall Street Journal. Inside the wsj's investigation of tesla's autopilot crash risksj. https://www.wsj.com/business/autos/tesla-autopilot-crash-investigation-997b0129, accessed: 2024-10-15.
52. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A. et al., "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.
53. Tunkelang, D. Llms and rag are great. what's next?, 2023, accessed: 2024-10-18.
54. Tur, G., Stolcke, A., Voss, L., Peters, S. et al., "The Calo Meeting Assistant System," *IEEE Transactions on Audio, Speech, and Language Processing* 18, no. 6 (2010): 1601-1611.
55. AI Upbeat, "Breaking down the science of benchmarking: Evaluating ai models," 2024, accessed: 2024-10-20.
56. Vemprala, S.H., Bonatti, R., Bucker, A., and Kapoor, A., "Chatgpt for Robotics: Design Principles and Model Abilities," *IEEE Access* (2024).
57. Wei, J., Wang, X., Schuurmans, D., Bosma, M. et al., "Chain-Ofthought Prompting Elicits Reasoning in Large Language Models," *Advances in Neural Information Processing Systems* 35 (2022): 24824-24837.
58. Wen, L.-H. and Jo, K.-H., "Deep Learning-Based Perception Systems for Autonomous Driving: A Comprehensive Survey," *Neurocomputing* 489 (2022): 255-270.
59. Jiayang W., Gan, W., Chen, Z., Wan, S., and Philip, S.Y., "Multimodal Large Language Models: A Survey," in *2023 IEEE International Conference on Big Data (BigData)*, IEEE, 2023, 2247-2256.

60. Yagiz, M.A., Mohajer Ansari, P., Pesé, M.D., and Goktas, P., "Transforming in-Vehicle Network Intrusion Detection: Vae-Based Knowledge Distillation Meets Explainable Ai," in *Proceedings of the SixthWorkshop on CPS&IoT Security and Privacy, CPSIoTSec'24*, New York, NY, USA, 2024, Association for Computing Machinery, 93-103.
61. Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., and Chen, E., "A survey on multimodal large language models," *arXiv preprint arXiv:2306.13549*, 2023.
62. You, Q., Jin, H., Zhaowen Wang, C.F., and Luo, J., "Image Captioning with Semantic Attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4651-4659, 2016.
63. Zhai, Y., Bai, H., Lin, Z., Pan, J. et al., "Fine-tuning large vision-language models as decisionmaking agents via reinforcement learning," *arXiv preprint arXiv:2405.10292*, 2024.
64. Zhang, J., Huang, J., Jin, S., and Lu, S., "Vision-Language Models for Vision Tasks: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).
65. Zhang, Q., Hu, S., Sun, J., Chen, Q.A., and Mao, Z.M., "On Adversarial Robustness of Trajectory Prediction for Autonomous Vehicles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, 15159-15168.
66. Zhao, Z., Monti, E., Lehmann, J., and Assem, H., "Enhancing contextual understanding in large language models through contrastive decoding," *arXiv preprint arXiv:2405.02750*, 2024.

ISSN 0148-7191