Computing&AI Connect
ISSN: 3006-4163
2025, Article ID. 2025.0018, Cite as: https://doi.org/10.69709/CAIC.2025.177363

SCIFINITI

Research Article

OPEN ACCESS

A Novel MLLM-Based Approach for Autonomous Driving in Different Weather Conditions

Sonda Fourati[✉,1] **Wael Jaafar**[✉,2,*] **Noura Baccar**[✉,1]

[1] The Computer Systems Engineering Department, Mediterranean Institute of Technology (MedTech), South Mediterranean University (SMU), Tunis 1053, Tunisia
[2] The Software and IT Engineering Department, Ecole de Technologie Supérieure (ÉTS), Montreal, QC H3C 1K3, Canada

Article History
Submitted: February 09, 2025 Accepted: June 30, 2025 Published: July 21, 2025

Abstract
Autonomous driving (AD) technology promises to revolutionize daily transportation by making it safer, more efficient, and more comfortable. Its role in reducing traffic accidents and improving mobility is vital to the future of intelligent transportation systems. AD systems (ADS) are expected to function reliably across diverse and challenging environments. However, existing solutions often struggle under harsh weather conditions such as foggy, rainy, or stormy circumstances, and mostly rely on unimodal inputs, thus limiting their adaptability and performance. Meanwhile, multimodal large language models (MLLMs) have shown remarkable capabilities in perception, reasoning, and decision-making, yet their application in AD, particularly under extreme environmental conditions, remains largely unexplored. Consequently, this paper proposes MLLM-AD-4o, a novel AD agent that leverages prompt engineering to integrate camera and LiDAR inputs for enhanced perception and control. MLLM-AD-4o dynamically adapts to available sensor modalities and is built upon GPT-4o to ensure contextual reasoning and decision-making. To support realistic evaluation, the agent was developed using the LimSim++ framework, which integrates the SUMO and CARLA driving simulators. Experiments are conducted under harsh conditions, including bad weather, poor visibility, and complex traffic scenarios. The MLLM-AD-4o agent's robustness and performance are assessed for decision-making, perception, and control. The obtained results demonstrate the agent's ability to maintain high levels of safety and efficiency, even in extreme conditions, using different perception components (e.g., cameras only, cameras with LiDAR, etc.). Finally, this work provides valuable insights into integrating MLLMs with AD frameworks, paving the way for fully safe ADS.

Keywords:
autonomous driving; multimodal large language models; harsh environment; CARLA; LimSim++; GPT-4o

1. Introduction

1.1. General Context

Recent advances that integrate blockchain, artificial intelligence (AI), and Internet-of-Things (IoT) technologies have substantially enhanced the design of secure and intelligent systems across a wide range of domains, including smart cities, industrial automation, and vehicular ecosystems, by ensuring data integrity, operational efficiency, and resilience [1–6]. Autonomous Driving Systems (ADS) are poised to become the cornerstone of next-generation transportation infrastructures. Their safe and efficient deployment critically depends on robust AI-driven frameworks, including those that synergize with IoT and blockchain technologies for real-time decision-making and secure data exchange [7].

In addition, ADS are expected to operate safely and reliably in any environment, including urban, rural, and highway. However, real-world ADS deployment presents several challenges, particularly for adverse weather conditions such as rain, fog, and low-light, in which traditional ADS's perception and control systems struggle to operate. To address this, modern ADS typically rely on a combination of heterogeneous sensor modalities, e.g., cameras,

* **Corresponding Author:**
Wael Jaafar, The Software and IT Engineering Department, Ecole de Technologie Supérieure (ÉTS), Montreal, QC H3C 1K3, Canada, wael.jaafar@etsmtl.ca

LiDAR, and radar, to capture and interpret complex driving scenes.

To meet the growing demands for safety, efficiency, and adaptability in ADS, recent research has increasingly turned to AI-driven approaches, supported by large language models (LLMs) and multimodal LLMs (MLLMs) [8,9]. These models offer advanced perception, reasoning, and decision-making capabilities. Specifically, MLLMs combine multimodal data, including images, video, and audio data, with the advanced reasoning capabilities of LLMs [10]. Hence, they can act as decision-making agents for various applications, including medicine, education, and intelligent transport systems (ITS). MLLMs can provide accurate and timely responses to dynamic driving conditions by processing large volumes of driving data to understand and predict complex traffic patterns, thus enabling autonomous vehicles (AVs) to learn from their environment and enhance their driving performance. In addition, MLLMs contribute significantly to the development of ADS by providing the computational power required to analyze the driving scene in real time and make accurate decisions [11].

In this context, various approaches have been proposed to use MLLMs for AD. Indeed, several strategies have been proposed based on prompt engineering [12–14], fine-tuning [15–19], and reinforcement learning with human feedback (RLHF) [20–22]. Furthermore, MLLMs toward AD provisioning could be utilized for perception and scene understanding [23–26], question answering [27,28], planning and control [16,29], and multitasking [30].

Despite their diverse methods and applications, MLLMs experience limitations in ADS. For instance, fine-tuning-based approaches are computationally expensive and may result in models overfitting for specific tasks or environments. Also, RLHF can align models with human preferences, however, it still faces scaling and generalization issues beyond particular training data instances. In contrast, MLLMs, particularly vision language models (VLMs) such as GPT-4o, have shown promising capabilities in tasks that involve perception, reasoning, and decision-making. The latter can simultaneously integrate and interpret visual and textual information, enabling interactive and adaptive AI systems.

1.2. Motivation and Problem Statement

Despite the growing use of MLLMs in perception and decision-making, no existing approach integrates multimodal inputs (e.g., LiDAR and camera) using prompt engineering to enable robust autonomous driving in harsh environmental conditions. This work addresses this gap by proposing MLLM-AD-4o, a flexible sensor-adaptive AD agent. This work aims to address the following ADS issues:

- **Need for weather-robust AD:** Even after the rapid progress in ADS, robust perception and reasoning under adverse weather conditions remain an open issue. Accordingly, this work aims to close this gap by introducing an MLLM-based ADS system that is resilient to varying environmental conditions.
- **Multimodal adaptability and sensor awareness**: Real-world driving requires adaptability to available sensor inputs. A model that can reason effectively using LiDAR, front/rear cameras, or any combination thereof is necessary.
- **Harnessing the reasoning capabilities of MLLMs:** The ability of models like GPT-4o to perform complex reasoning and decision-making when properly prompted presents an opportunity to enhance ADS, especially through prompt engineering tailored for driving tasks.
- **Lack of open, reproducible frameworks for AD testing in harsh conditions:** Many existing benchmarks do not sufficiently simulate diverse weather and traffic conditions or provide extensibility. In contrast, this work extends LimSim++ to introduce a new benchmark within the CARLA simulator for reproducible testing in challenging weather environments.

1.3. Contributions and Paper Organization

To tackle the above challenges, this work introduces MLLM-AD-4o, a novel prompt engineering MLLM-based approach for AD, which is capable of handling autonomous driving under any weather condition. The proposed approach integrates different sensor modalities, including LiDAR and cameras, allowing the system to adapt according to available sensor data. Based on prompt engineering, MLLM-AD-4o provides enhanced environment perception, scene understanding, reasoning, and decision-making in ADS. Hence, our contributions can be summarized as follows:

- Unlike previous studies, this contribution integrates harsh environmental driving conditions into the CARLA simulator for the first time by leveraging functions and modules from the LimSim++ framework. This work has been documented and distributed on GitHub for open public access and reproducibility [31].
- The proposed MLLM-AD-4o is a novel MLLM-based AD agent that leverages GPT-4o [32] to drive perception and control decision-making.

2025, Vol. 2, Article ID. 2025.0018
https://doi.org/10.69709/CAIC.2025.177363

- Through extensive experiments, a performance analysis of the proposed MLLM-AD-4o in terms of safety, comfort, efficiency, and speed score metrics, and under different weather conditions and using different combinations of sensor data (e.g., front and/or rear cameras and/or LiDAR).
- The cumulative distribution functions (CDFs) are used for deeper insight into the agent's performance variability.

This combination of prompt engineering, sensor adaptability, and resilience to weather is the core contribution that separates this work from existing MLLM-based AD approaches.

The remainder of the paper is organized as follows. Section 2 presents related work that integrate LLMs/MLLMs into driving agents in a closed-loop using the CARLA simulator. Section 3 explains the basic concepts of ADS and LLM/MLLMs with a focus on the architecture and functions of GPT-4o. Section 4 presents the architecture, main modules, and technical details to integrate an MLLM agent driver in a closed-loop environment using LimSim++. Section 5 evaluates the performance of MLLM-AD-4o in diverse scenarios and conditions. Finally, Section 6 closes the paper.

2. Related Work

Recently, there has been a significant effort to enhance the efficiency and reliability of ADS using LLMs and MLLMs. Among the proposed driving agents, this paper focuses on related work using prompt engineering. The authors of [33] proposed "LLM-driver", a framework integrating numeric vector modalities, into pre-trained LLMs for question-answering (QA). LLM-driver uses object-level 2D scene representations to fuse vector data into a pre-trained LLM with adapters. A language generator (LanGen) is used to ground vector representations into LLMs. The model's performance was evaluated on perception and action prediction using mean absolute error (MAE) for predictions, accuracy of traffic light detection, and normalized errors for acceleration, brake pressure, and steering. In [34], the SurrealDriver framework has been proposed. It consists of an LLM-based generative driving agent simulation framework with multitasking capabilities that integrate perception, decision-making, and control processes to manage complex driving tasks. In [35], the authors developed LLM-based driver agents with reasoning and decision-making abilities, aligned with human driving styles. Their framework utilizes demonstrations and feedback to align the agents' behaviors with those of humans, utilizing data from human driving experiments and post-driving interviews. Also, the authors of [36] proposed the Co-driver framework for planning & control and trajectory prediction tasks. Co-driver utilizes prompt engineering to understand visual inputs and generate driving instructions, while it uses deep reinforcement learning (DRL) for planning and control. In [37], the authors introduced the PromptTrack framework, an approach to 3D object detection and tracking, by integrating cross-modal features within prompt reasoning. PromptTrack uses language prompts that act as semantic guides to enhance the contextual understanding of a scene. Finally, the authors in [13] proposed HiLM-D as an efficient technique to incorporate high-resolution information in ADS for hazardous object localization.

Despite the variety of proposed driving agents, to our knowledge, no study has fully assessed the performance of MLLM-based driving agents within a closed-loop framework under harsh environmental conditions. Table 1 summarizes the aforementioned contributions with a comparison to this work.

3. Background

The key modules associated with an AV are *perception*, *planning*, *localization*, *decision-making*, and *action or control* as highlighted in Figure 1. The *perception* module ensures sensing of the surrounding environment and identifies the driving scene. The main goal of the *localization* module is to estimate with high precision and accuracy the vehicle position on the map. *Path planning and decision* modules establish the waypoints that the vehicle should follow when moving through the surroundings. The set of waypoints corresponds to the vehicle trajectory. The *action and control* module deals with the system's actuation, such as braking, steering, and acceleration.

Several MLLMs have been developed [9], with a subset of them suitable for ADS, as discussed in Section II. In this work, GPT-4o [32] was selected. GPT-4o is an optimized version of GPT-4 built around the latter's foundational transformer architecture with enhancements in understanding, generating, and maintaining context across interactions. GPT-4o includes various specialized modules to improve its operation, namely (i) language understanding module, (ii) language generation module, (iii) context management module, (iv) task-specific modules, and (v) optimization and fine-tuning modules. Programming GPT-4o involves the use of application programming interfaces (APIs) or software development kits (SDKs), custom prompts, or fine-tuning and customization. In our work, interaction with GPT-4o is performed via API calls, where a customized prompt to guide the response of the model is utilized.

Table 1: Summary of related works.

Ref.	Contribution	Used Models	Used Data	AD Tasks	Performance Metrics	Harsh Environment
[33]	Design and validation of an LLM Driver that interprets and reasons about driving situations to generate adequate actions.	GPT-3.5; RL agent trained with PPO, lanGen, and trainable LoRA modules.	160k QA driving pairs dataset; Control Commands Dataset.	Perception and action prediction.	Traffic light detection accuracy; Acceleration, brake, and steering errors.	No
[34]	LLM-based agent in a simulation framework to manage complex driving tasks.	GPT-3 and GPT-4.	Driving Behavior Data; Simulation data from CARLA simulator; NLP libraries.	Perception, control, and decision-making.	Collision rate.	No
[35]	LLM-based driver agents with reasoning and decision-making, aligned with human driving styles.	GPT-4.	Private dataset from a human driver; NLP library; RL library.	Behavioral alignment; Human-in-the-Loop system.	Collision rate, average speed, throttle percentage, and brake percentage.	No
[36]	Co-driver agent, based on a vision language model, for planning, control, and trajectory prediction.	Qwen-VL (9.6 billion parameters), including a visual encoder, a vision-language adapter, and the Qwen LLM.	CARLA simulation data; ROS2; Customized dataset.	Adjustable driving behaviors; Trajectory and lane prediction; Planning and control.	Fluctuations frequency, and running time.	Foggy/gloomy; Rainy/-gloomy
[37]	PromptTrack framework for 3D object detection and tracking by integrating cross-modal features within prompt reasoning.	VoVNetV2 to extract visual features; RoBERTa to embed language prompts.	NuPrompt for 3D perception in AD.	3D object detection and tracking; Scene understanding.	Average multiple object tracking precision (AMOTA), and identity switches (IDS).	No
[13]	High-resolution scene understanding using proposed HiLM-D.	BLiP-2; Q-Former; MiniGPT-4.	DRAMA dataset; Pytorch; Visual Encoder; Query detection module; ST-adapter module.	Risky object detection; Vehicle intentions and motions' prediction.	Mean intersection over union (mIoU) for detection; BLEU-4, METEOR, CIDER, and SPICE for captioning.	No
This work	Implementation of harsh environment scenarios and performance evaluation of an MLLM-based driving agent.	GPT-4o.	Dedicated Prompt.	Perception and decision-making.	Safety, comfort, efficiency, and speed scores.	Heavy rain; Storm; Foggy; Wetness; Good

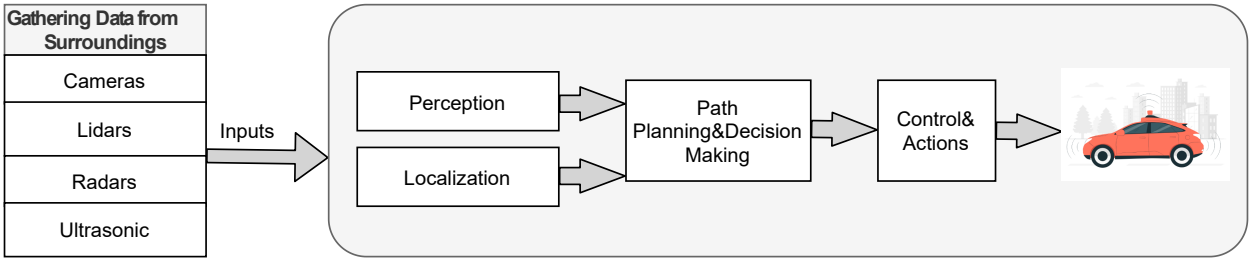


Figure 1: Typical architecture of ADS.

4. Materials and Methods: Proposed MLLM-Based Driving Agent in a Closed-Loop Environment

This section presents the tools used to integrate the proposed MLLM-based driving agent, also known as MLLM-AD-4o.

4.1. LimSim++ Framework

LimSim++ is a closed-loop framework for deploying MLLMs for autonomous driving, where multimodal data, such as text, audio, and video, can be analyzed, and then accurately react to driving scenarios [38]. LimSim++ integrates algorithms for real-time decision-making, obstacle detection, and navigation, making it an interesting tool for designing, developing, testing, and refining ADS.

To do so, LimSim++ simulates a closed-loop system with specifications of the environment, including road topology, dynamic traffic flow, navigation, and traffic control. The MLLM agent is based on prompts that provide real-time scenario information in visuals or text. The MLLM-supported driving agent system can process information, use tools, develop strategies, and assess itself. Figure 2 depicts the architecture of Limsim++ and explains its main modules as follows:

- *Simulation system*: It gathers the scenario information from SUMO and CARLA, and packages it as in-

2025, Vol. 2, Article ID. 2025.0018
https://doi.org/10.69709/CAIC.2025.177363

put for the MLLM, such as visual content, scenario cognition, and task description.

- *Driver Agent*: It communicates with the "simulation system" to make driving judgments and uses MLLMs to interpret them. Data is represented as prompts for MLLMs to make suitable driving decisions. The outputs of MLLMs influence vehicle behavior and are kept in the case log system for knowledge accumulation. Following the simulation, the decisions are evaluated as follows: Those that performed well are immediately added to the memory, while poor decisions are added after reflection. Building upon the results of the previous experiment, the reflection module refines the incorrect decision-making process by incorporating expert experience. The modified decision results are then added to the Driver Agent's memory modules, serving as driving experiences to provide a few-shot examples for the agent in subsequent decision-making processes. This ensures continuous improvement and knowledge accumulation within the system.

The MLLM-AD-4o architecture was designed by customizing the above modules as follows:

(1) The original "MPGUI.py" file within LimSim++ and its related dependencies are modified to include the visualization of six cameras (rather than the default three front cameras) in the autonomous driving simulation. Hence, a comprehensive view is provided, which is crucial to evaluating the perception capability of the driving system, analyzing/debugging the AV's sensor inputs, and decision-making.

(2) The default "VLM-Driver-Agent" was updated to allow changes in the scenario and to integrate new sensors and environmental conditions.

(3) To automate the setup of different weather conditions, a novel function in the CARLA simulator was created .

5. Experimental Results

5.1. Definition of Metrics and Parameters

To accurately assess the AD performance of the proposed MLLM-AD-4o, the following performance metrics (called scores) are introduced:

- *Safety score:* The safety level is commonly assessed through Time-to-Conflict (TTC), which is a measure of the time remaining until a potential collision occurs between two moving objects, assuming no changes in their current trajectories. It serves as an indicator of the urgency to take corrective actions to avoid conflict. When TTC falls below a specific threshold, a potential risk warrants penalties. The safety score can be given by

$$\text{Safety\_score} = \begin{cases} 1, & \text{if } \tau_e \geq \tau_{\text{th}} \\ \tau_e/\tau_{\text{th}}, & \text{otherwise,} \end{cases} \quad (1)$$

where τ_e is the TTC of the AV and τ_{th} is the threshold value of TTC. Higher values reflect safer traveling.

- *Comfort score:* It evaluates the smoothness of the AV ride. Using empirical data, reference values can be determined for different driving styles, such as cautious, normal, and aggressive. It is computed as the average of the summation of acceleration score (acc_score), jerk score, lateral acceleration score (lat_acc_score), and lateral jerk score (lat_jerk_score), as shown below

$$\begin{aligned} \text{Comfort\_score} &= \tfrac{1}{4}(\text{acc\_score} + \text{jerk\_score} \\ &+ \text{lat\_acc\_score} + \text{lat\_jerk\_score}). \end{aligned} \quad (2)$$

A higher comfort score means that traveling is smoother.

- *Efficiency score:* Driving efficiency is computed according to the vehicle's speed. In regular traffic conditions, the AV should maintain a speed at least equivalent to the average speed of surrounding vehicles. In sparse traffic conditions, the AV should approach the road's speed limit for efficiency. This score is computed as (3).

$$\text{Efficiency\_score} = \begin{cases} 1.0, & \text{if } v_e \geq v^* \\ \frac{v_e}{v^*}, & \text{otherwise,} \end{cases} \quad (3)$$

where v_e represents the speed of the AV and $v^* \in \{v_{avg}, v_{limit}\}$ is the targeted speed for efficiency, being v_{avg} as the average speed of surrounding vehicles, and v_{limit} as the road's speed limit. A high value means efficient traveling.

- *Speed score:* The speed limit score (or speed score) penalizes the vehicle for exceeding the speed limit. This score is set to 0.9 when the AV exceeds the speed limit and 1 otherwise. This score is expressed as

$$\text{Speed\_score} = 1 - 0.1 \times \frac{\text{Nbr\_frames\_with\_speeding}}{\text{Total\_nbr\_frames}}, \quad (4)$$

where "Nbr_frames_with_speeding" is the number of captured frames while speeding, and "Total_nbr_frames" is the total number of captured frames.

In the following simulations, we set up $\alpha_1 = \alpha_2 = 0.25$, $\alpha_3 = 0.5$, and $v_{limit} = 50$ km/h (i.e., 13.89 m/s).

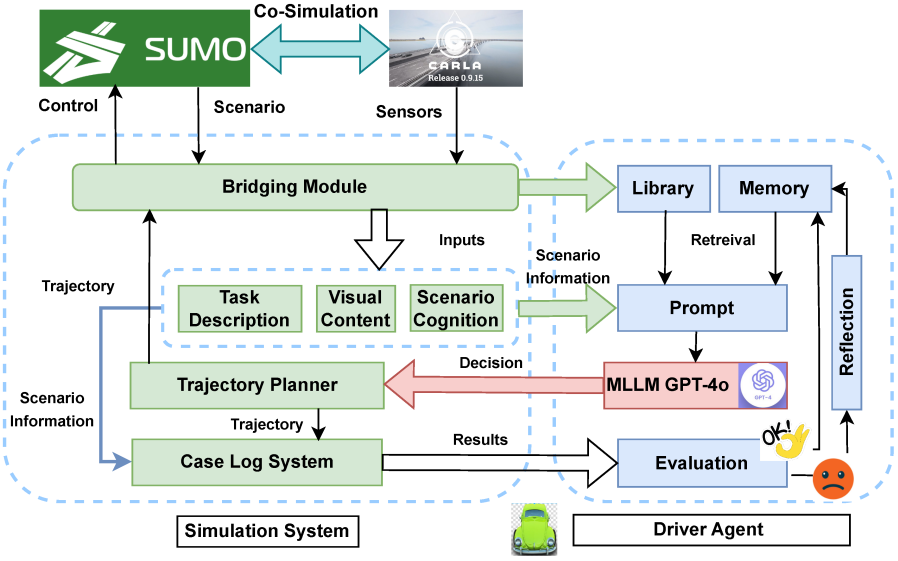


Figure 2: Integration of MLLM-AD-4o in LimSim++ (adapted with permission from [38]).

5.2. Interaction with GPT-4o API

To use the GPT-4o API with prompts to process the driving environment, with data mainly collected from CARLA and SUMO, the following steps are conducted: [31]:

(1) Install the OpenAI package and get an OpenAI API key.

(2) Capture images from the front and back cameras (in CARLA), save them in a suitable format, and then convert them for API submission.

(3) Prepare the API request, i.e., send the prompt with the image data to the GPT-4o API. The prompt describes what we want the model to do with the images and environmental data. Specifically, we ask the model to analyze the surroundings and make a driving decision. Figure 3 presents an example of using prompts in simulated scenarios that use six cameras. As illustrated, the prompts start with a clear role definition through the use "You are GPT-4o, a large multi-modal model..."; which enforces the model to behave like a decision-making assistant in autonomous driving, emphasizing responsibility, situational awareness, and accuracy. Then, the prompt explicitly defines the inputs (Front-view image, Back-view image, Navigation information, and a list of possible actions) and instructs GPT-4o to describe the visual scene, to reason about the next move using navigation data and visual inputs, and to decide on the next action. Finally, the prompt requires the output format to be structured into three clear sections corresponding to *Description*, *Reasoning*, and *Decision*. This structure encourages explainability and transparency in the model's decisions. Moreover, the decision must match exactly one item in the action set, e.g., *Idle* (IDLE), *Acceleration* (AC), *Deceleration* (DEC), *Turn_Left* (TL), and *Turn_Right* (TR). This is critical to avoid any ambiguity and to reduce the number of consumed tokens.

(4) GPT-4o API processes the gathered data (e.g., navigation info, lane info, vehicle info, and surrounding info) to make a decision, then sends a response to the MLLM agent in the form of a driving action.

(5) The agent converts the action received from the GPT-4o API to behavior, then the trajectory planner module updates the path according to the decision made by the driver agent, to be sent to the CARLA system and executed. Figure 4 below illustrates an example of GPT-4o's decision-making in a simulation under "Stormy weather". Accordingly, the decision taken is "Accelerate" despite the stormy weather with reduced visibility. In making this decision, the GPT-4o API considered the absence of obstacles directly ahead or in adjacent lanes in the processed images, along with the vehicle's current speed and location as specified in the prompt. It concluded that acceleration is appropriate in this situation due to the lack of nearby hazards, the safe road surface visibility, and the need to maintain efficient traffic flow.

2025, Vol. 2, Article ID. 2025.0018
https://doi.org/10.69709/CAIC.2025.177363

Prompt:: in case of using Front and back cameras

You are GPT-4o, a large multi-modal model trained by OpenAI. Now you act as a mature driving assistant, who can give accurate and correct advice for human driver in complex urban driving scenarios. You'll receive some images from the onboard camera.
You'll need to make driving inferences and decisions based on the information in the images.
At each decision frame, you receive navigation information and a collection of actions.
You will perform scene description, and reasoning based on the navigation information, the front-view image and back-view image. Eventually you will select the appropriate action output from the action set.
Make sure that all of your reasoning is output in the `## Reasoning` section, and in the `## Decision` section you should only output the name of the action, e.g. `AC`, `IDLE` etc.

Your answer should follow this format:
Description
Your description of the front-view image, and your description of the back-view image.
Reasoning
reasoning based on the navigation information, the front-view image and the back-view image.
Decision
one of the actions in the action set.(SHOULD BE exactly same and no other words!)
"""

Figure 3: Prompt used by GPT-4o (For AV with six cameras).

5.3. Setup of Weather Conditions

This work considered, in addition to "Good weather", four harsh weather conditions, namely "Heavy rain", "Storm", "Foggy", and "Wetness". A new function in the CARLA simulator, named "set_weather", is developed to configure different weather conditions. This function has eight parameters:

- *Cloudiness:* The amount of cloud cover in the sky. A high value means more clouds.
- *Precipitation:* The intensity of the rain. A high value means heavier rain.
- *Precipitation_deposits:* The amount of water that accumulates on the ground due to precipitation. A high value means an important accumulation.
- *Wind_intensity:* The strength of the wind. A high value means stronger wind.
- *Sun_altitude_angle:* The angle of the sun above the horizon. Lower values can simulate dawn, dusk, or low-light conditions.
- *Fog_density:* The density of fog in the environment. A high value implies thicker fog.
- *Fog_distance:* The distance at which the fog starts to become noticeable.
- *Wetness:* The wetness of the road surfaces. A high value makes the roads wetter.

In Table 2, we summarize the parameters' values to set up each weather condition using the function "set_weather" [31].

5.4. Integration of Semantic LiDAR in LimSim++

CARLA supports various sensing tools, including traditional LiDAR and semantic LiDAR. Traditional LiDAR provides raw distance measurements in the form of a point cloud, without an inherent understanding of the objects in the scene. In contrast, semantic LiDAR provides both the point cloud and an understanding of what each point represents, thus simplifying tasks that involve scene understanding, object recognition, and data processing reduction. Semantic LiDAR is selected within this work. To add the latter in the perception environment within Limsim++, several steps should be followed [31]:

(1) Attach "Semantic_LiDAR" function to the AV and gather data in CARLA.
(2) Save and process LiDAR data.
(3) Add "Semantic_LiDAR" data to the prompt request.

Table 2: Parameters to set up the weather conditions.

Case	Set_Weather Parameters Values
Heavy Rain	self.set_weather(cloudiness=80.0, precipitation=70.0, precipitation_deposits=60.0, wind_intensity=30.0, sun_altitude_angle=45.0, fog_density=10.0, fog_distance=10.0, wetness=80.0)
Storm	self.set_weather(cloudiness=80.0, precipitation=100.0, precipitation_deposits=100.0, wind_intensity=100.0, sun_altitude_angle=20.0, fog_density=20.0, fog_distance=10.0, wetness=80.0)
Fog	self.set_weather(cloudiness=40.0, precipitation=5.0, precipitation_deposits=5.0, wind_intensity=10.0, sun_altitude_angle=60.0, fog_density=70.0, fog_distance=3.0, wetness=10.0)
Wetness	self.set_weather(cloudiness=30.0, precipitation=0.0, precipitation_deposits=0.0, wind_intensity=0.0, sun_altitude_angle=70.0, fog_density=0.0, fog_distance=0.0, wetness=100.0)
Good Weather	self.set_weather(cloudiness=00.0, precipitation=00.0, precipitation_deposits=00.0, wind_intensity=00.0, sun_altitude_angle=60.0, fog_density=00.0, fog_distance=20.0, wetness=00.0)

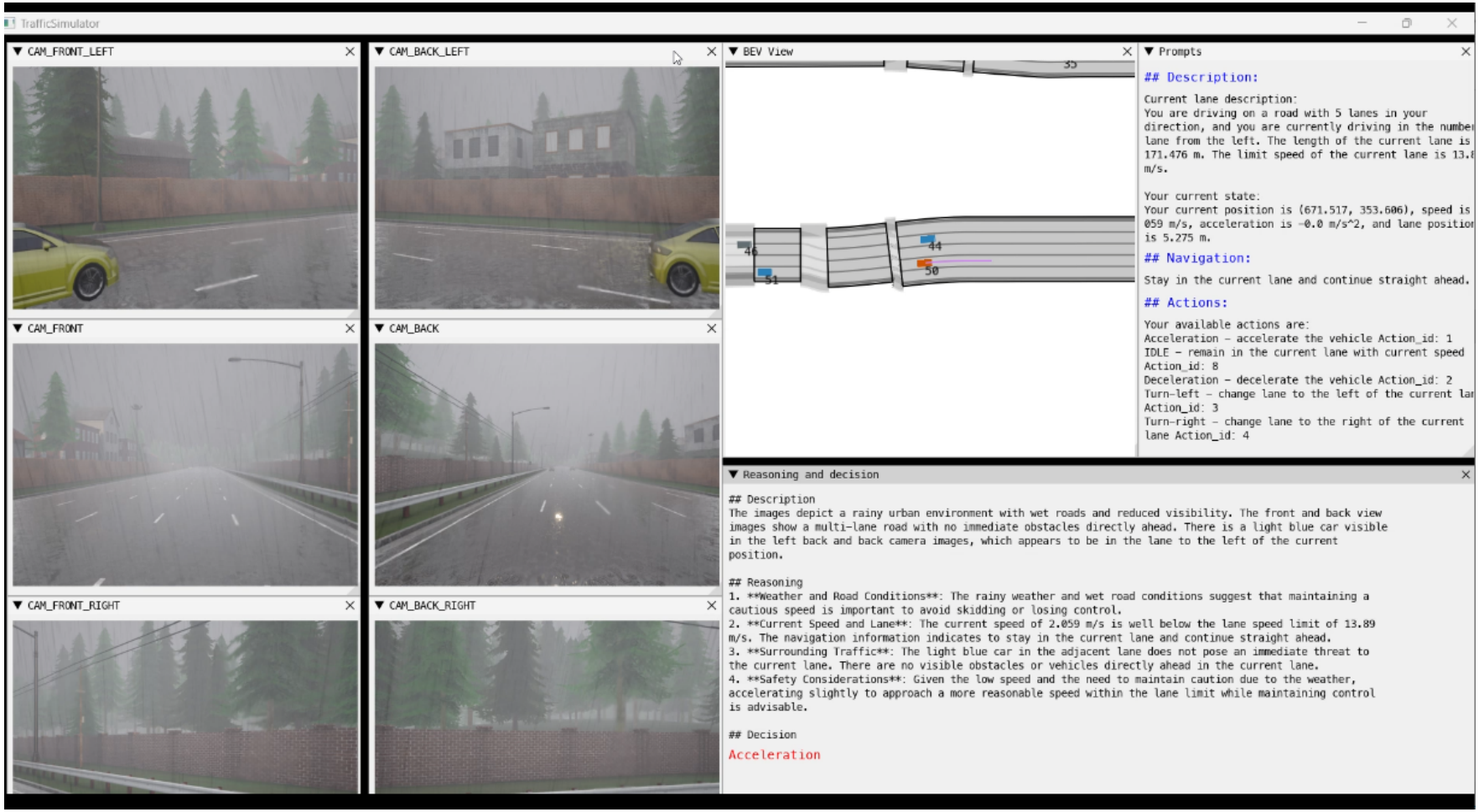


Figure 4: Example of autonomous driving in stormy weather (For AV with six cameras).

2025, Vol. 2, Article ID. 2025.0018
https://doi.org/10.69709/CAIC.2025.177363

5.5. Simulation Results and Discussion

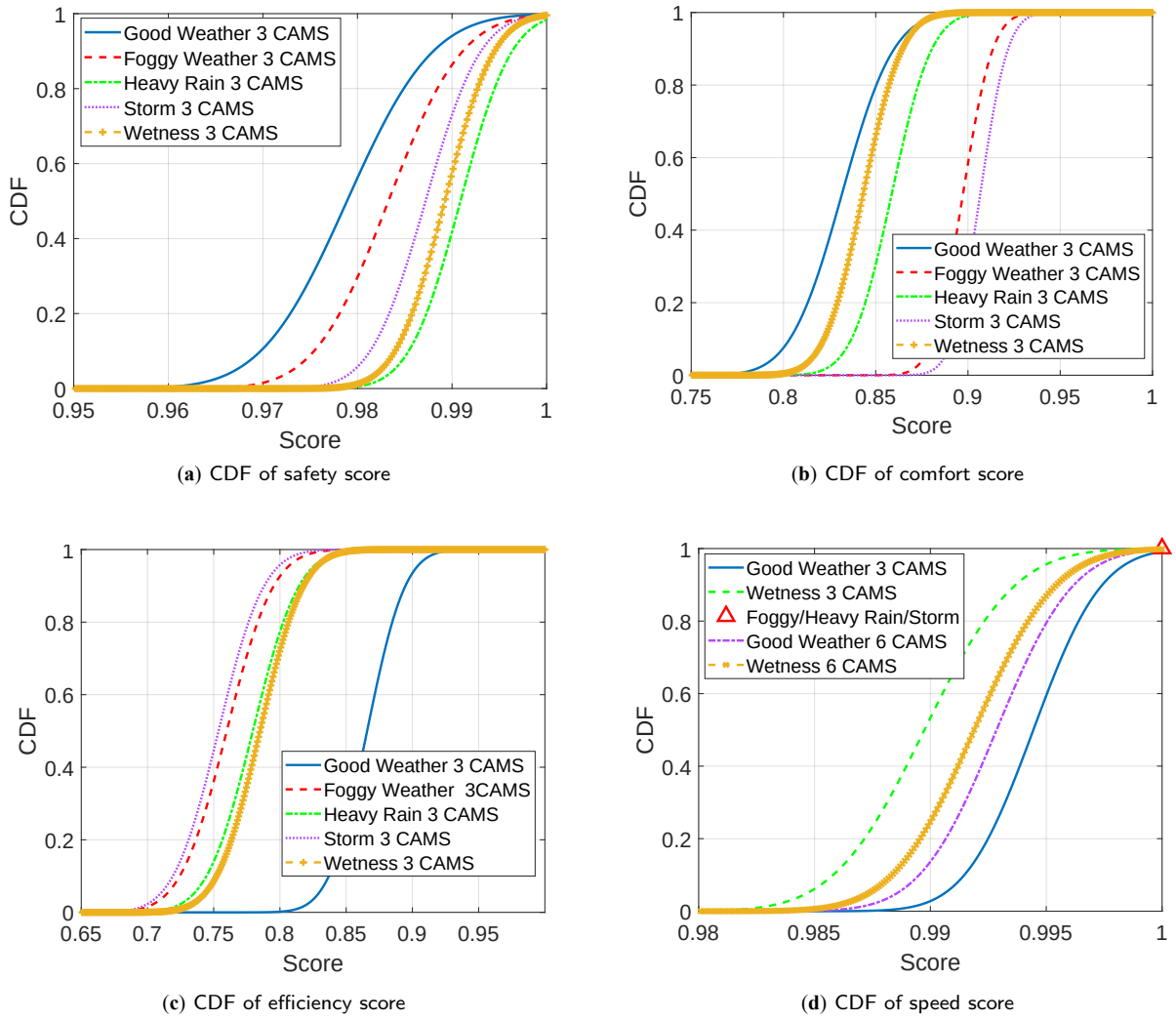
This section assesses the performance of the proposed MLLM-AD-4o autonomous driving agent in different conditions. Each experiment is built with the same scenario of our AV driving between points A and B, where it travels on multilane main roads and has to maneuver in curves. The autonomous driving agent has to make one decision among five possible decisions in each time frame, including idle (no change in current behavior), acceleration, deceleration, turn left, or turn right.

Figure 5a–d presents the cumulative distribution functions (CDFs) for the safety, comfort, efficiency, and speed scores/metrics, respectively, under various weather conditions, when 3 cameras are used for autonomous driving by the proposed MLLM-AD-4o method[1]. Figure 5a, illustrates that when the weather is good (blue line), the agent takes a riskier behavior that may cause collisions than in a harsh condition. This also impacts the passengers' comfort during travel as shown in Figure 5b. Nevertheless, in good weather, the driving behavior is the most efficient as shown in Figure 5c and it aligns with the speed performance depicted in Figure 5d, This trade-off highlights that while the AV may be driven more aggressively and with less comfort, its efficiency improves as it aligns more closely with the surrounding traffic's speed. In contrast, in harsh weather conditions, e.g., heavy rain (green line) or wetness (dark yellow line), safety is enforced through slower driving, which translates into a better comfort score, offering a smoother and more cautious ride, at the expense of degraded efficiency and speed performances.

Among the weather conditions, we notice in Figure 5 that the best safety CDF is achieved in heavy rain (green line), the best comfort CDF in stormy weather (pink line), the best efficiency CDF in good weather, and the best speed CDF in all of the foggy, heavy rain, and stormy weather conditions (red triangle). This suggests that the MLLM-AD-4o driving agent considers heavy rain the most unsafe environment, requiring low-speed traveling. Also, traveling can be smooth in stormy weather, while driving is most efficient in good weather. Finally, in foggy/heavy rain/storm conditions, the agent is more likely to respect the speed limits as it favors slower driving to guarantee security, which is not the case in good and wet weather conditions. These results reflect the trade-offs the MLLM-AD-4o agent makes between various performance scores, under different weather patterns.

Figure 6a–c, depict the CDFs of the safety, comfort, and efficiency scores when 6 cameras (3 front + 3 rear) are used by the AD agent. In good weather, MLLM-AD-4o now achieves the best safety CDF performance (blue triangle), at the expense of comfort and efficiency. Indeed, the additional information brought by the 3 rear cameras makes the agent more cautious regarding the behavior of the surrounding vehicles, which was not possible when it only relied on 3 cameras. This is confirmed with the speed CDF result (purple line) in Figure 5d, which is higher than that of the CDF of the 3 cameras. According to Figure 6b, the best CDFs are realized in stormy/foggy weather (with a preference for stormy) at the expense of efficiency and safety. Indeed, in these situations, the AD agent is more likely to drive smoothly with minor behavior changes in response to poor visibility, with a higher safety risk and reduced efficiency, due to misalignment with the surrounding road traffic behavior. Moreover, the best CDF efficiency is obtained in wet conditions, as shown in Figure 6c. Indeed, with the knowledge of its 6 cameras, the AV maintains a speed closer to the average of other vehicles by continuously adapting to the dynamics of the road and surrounding traffic. This is validated through the results in Figure 5d where the 6 cameras' speed CDF is better than the 3 cameras' one in wetness. However, as illustrated in Figure 6a,b, this behavior negatively impacts the safety and comfort scores. In summary, the addition of the rear cameras enhances the vehicle's ability to perceive and respond to potential collision threats, particularly in challenging weather conditions.

Figure 7a–c illustrate the CDFs of the safety, comfort, and efficiency scores, when the AD agent is equipped with 3 (front) cameras only, 6 (3 front + 3 rear) cameras only, 3 (front) cameras + LiDAR, and 6 (3 front + 3 rear) cameras + LiDAR, respectively, operating in heavy rain conditions. The best safety CDF is obtained when the LiDAR is used with 3 cameras (green triangle in Figure 7a). Indeed, activating the LiDAR in adverse weather conditions enhances the system's awareness of its environment, leading to more responsible driving actions and improved safety. Nevertheless, the safety and efficiency of CDFs degrade when LiDAR is activated with 6 cameras. This suggests that rear cameras' information might distort the agent's comprehension of its surroundings when combined with LiDAR. The best comfort score is also achieved when 3 cameras are activated with LiDAR (green line in Figure 7b), while the best efficiency score is realized with a 3-camera system (blue line in Figure 7c). Nevertheless, the activation of LiDAR with the 3 cameras (green line in Figure 7c) realizes similar efficiency CDF performance to the latter. Based on these results, the activation of LiDAR with cameras should be triggered promptly, under particular conditions, and with specific configurations, as the full combination of cameras with LiDAR does not necessarily lead to better AD performances.



(a) CDF of safety score

(b) CDF of comfort score

(c) CDF of efficiency score

(d) CDF of speed score

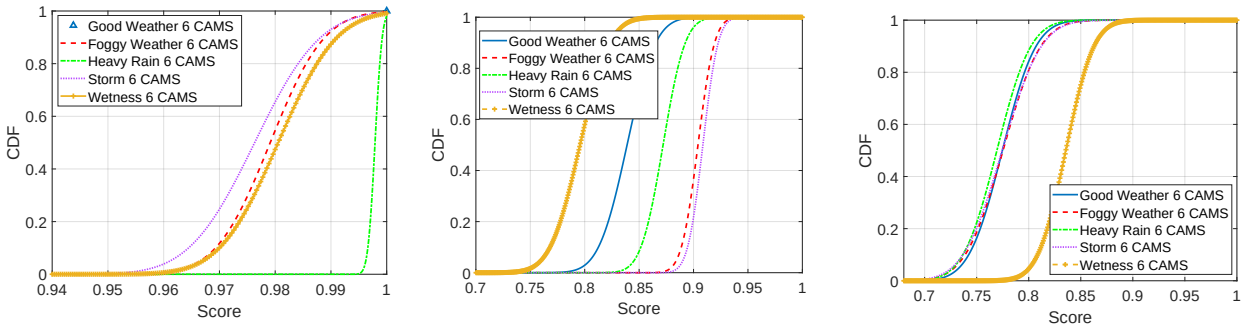
Figure 5: CDFs of MLLM-AD-4o scores (Front cameras).

To quantitatively assess the scalability and latency implications of querying GPT-4o under different sensor configurations, we plot in Figure 8 the prompt token load per frame for different sensor configurations (3 cameras, 6 cameras, 3 cameras plus LiDAR, and 6 cameras plus LiDAR) in heavy rain weather.

As shown, the prompt size increases with the number of sensors. Specifically, configurations that include LiDAR contribute significantly to the token count, reaching around 1850 tokens for the "6Cam+LiDAR" setup, while simpler configurations, e.g., "3 Cams", consume under 800 tokens. This trend illustrates a linear scalability pattern in prompt size with sensor complexity, which affects the API querying time. Given that OpenAI's GPT-4o API latency is partially a function of input prompt size, these findings allow us to quantify a trade-off between perception fidelity and real-time responsiveness. For example, at a 30 frame-per-second (FPS) rate, using "6Cam+LiDAR" could saturate the API token limits or introduce delays depending on the network bandwidth and batching strategies.

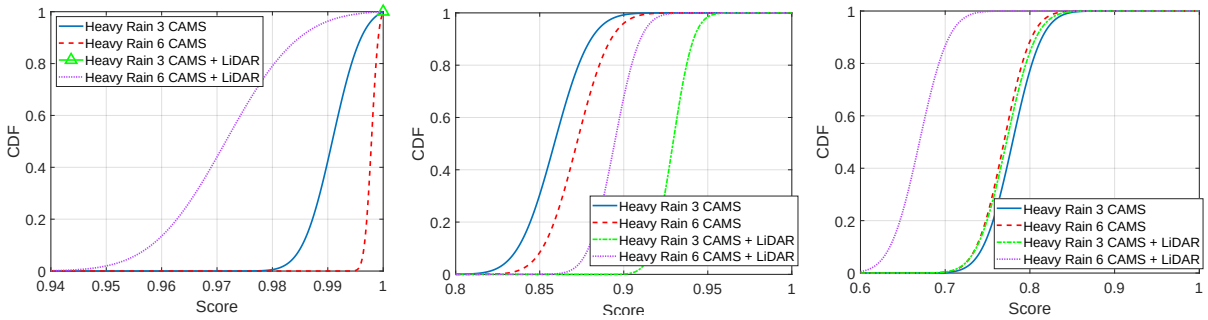
To further explore the latency trade-off in a realistic low-bandwidth setting, we conduct an additional experiment to evaluate token usage and inference time when querying GPT-4o through its API. The results, summarized in Figure 9, present four subplots capturing prompt tokens (top-left), completion tokens (top-right), total tokens (bottom-left), and total response time (bottom-right) for 3-camera and 6-camera sensor configurations. We noticed that increasing the number of cameras leads to higher prompt and completion token counts, which correlate with increased total token counts. Nevertheless, the total re-

2025, Vol. 2, Article ID. 2025.0018
https://doi.org/10.69709/CAIC.2025.177363



(a) CDF of safety score (b) CDF of comfort score (c) CDF of efficiency score

Figure 6: CDFs of MLLM-AD-4o scores (Front+rear cameras).



(a) CDF of safety score (b) CDF of comfort score (c) CDF of efficiency score

Figure 7: CDFs of MLLM-AD-4o scores (Combinations of cameras with LiDAR; Heavy rain).

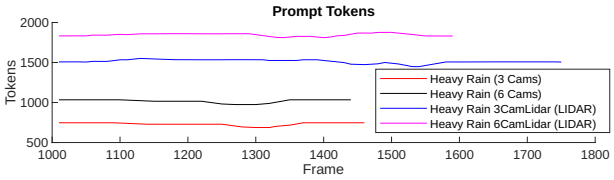


Figure 8: Comparison of token load (different sensor settings, heavy rain).

sponse time is shorter with 6-camera data, probably due to the model producing shorter completions, because of a richer initial context.

These results emphasize the nuanced trade-off between richer multimodal context and token budget constraints, a key consideration for real-time deployment of LLM-based autonomous driving agents.

6. Conclusions

This work introduced MLLM-AD-4o, a novel AD agent based on GPT-4o, capable of performing accurate AV perception and safe control decisions through prompt engineering across diverse environmental conditions. Harsh weather scenarios were integrated into the CARLA simulator through the LimSim++ framework, enabling the evaluation of MLLM-AD-4o with different sensor configurations. For clarity, the performances have been assessed in terms of safety, comfort, efficiency, and speed score CDFs. The obtained results prove that combining front and rear cameras' data significantly improves performance compared to single-view configurations, in clear weather. Moreover, integrating LiDAR with a moderate number of cameras (e.g., only three) yields optimal performance in different weather conditions. In contrast, excessive sensor inputs, e.g., from LiDAR and six cameras simultaneously, can be counterproductive. These findings demonstrate the practicality of MLLM-AD-4o for AD and the importance of context-aware sensor configuration. In future work, we will expand our framework to more complex traffic scenarios and explore the use of lighter MLLMs that are better adapted for real-time decision-making.

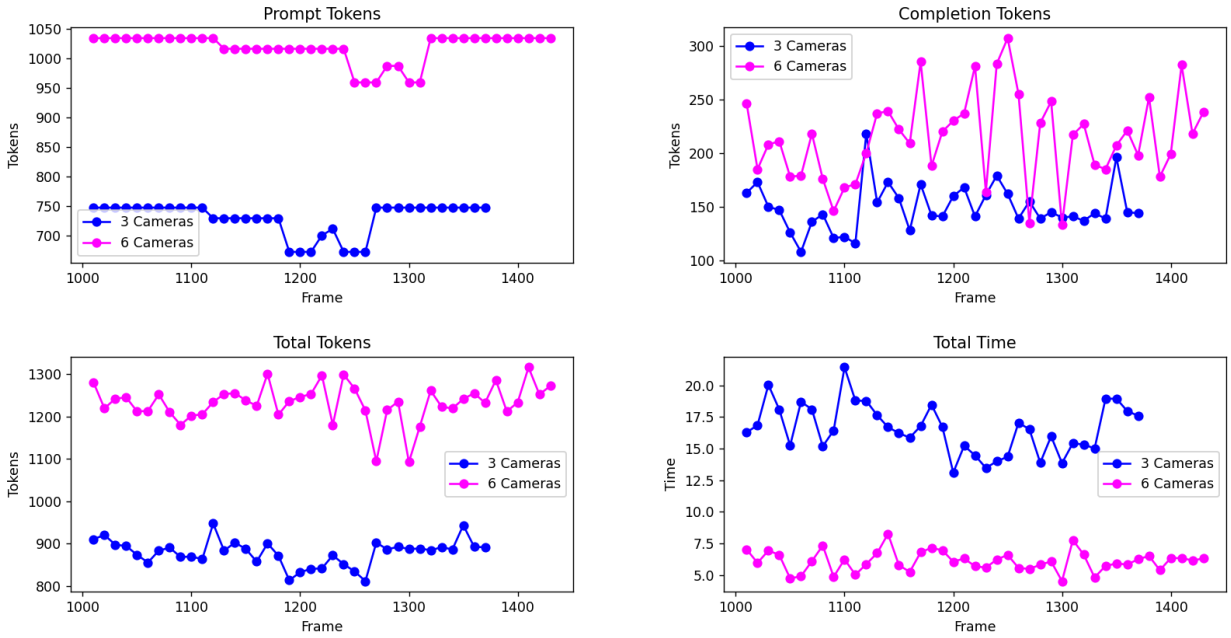


Figure 9: Comparison of prompt tokens (top-left), completion tokens (top-right), total tokens (bottom-left), and total response time per frame (bottom-right) when using 3 or 6 cameras.

List of Abbreviations

We present here the list of acronyms used in this paper.

AD	Autonomous Driving
ADS	Autonomous Driving Systems
API	Application Programming Interface
AV	Autonomous Vehicles
CDF	Cumulative Distribution Functions
DRL	Deep Reinforcement Learning
ITS	Intelligent Transport Systems
LanGen	Language Generator
LiDAR	Light Detection and Ranging
LLM	Large Language Models
MAE	Mean Absolute Error
MLLM	Multimodal Large Language Models
NLP	Natural Language Processing
QA	Question-Answering
SDK	Software Development Kit
TTC	Time-to-Conflict
VLM	Vision-Language Model

Note

[1] Figure 5d illustrates also speed score results for the 6 cameras case, which will be discussed later.

Author Contributions

S.F.: Conceptualization, methodology, software, investigation, data curation, writing, original draft preparation; W.J.: Validation, Formal analysis, writing, review, and editing, supervision, project administration; N.B.: Formal analysis, Visualization, writing, review, and editing, supervision. All authors have read and agreed to the published version of the manuscript.

Availability of Data and Materials

The data used in this study was generated through extensive simulations using the LimSim++ tool. The simulation outputs, including traffic scenarios, sensor data, and decision logs, are available from the corresponding author upon reasonable request.

Consent for Publication

All authors have read and approved the final version of the manuscript and consent to its submission for publication. Where applicable, consent has been obtained from individuals or institutions for the use of data, images, or other materials included in the manuscript.

Conflicts of Interest

The authors have no competing interests to declare relevant to this article's content. All authors certify that they have no affiliations with or involvement in any or-

2025, Vol. 2, Article ID. 2025.0018
https://doi.org/10.69709/CAIC.2025.177363

ganization or entity with any financial or non-financial interest in the subject matter or materials discussed in this manuscript.

Funding

This work is funded in part by a Mitacs Globalink Research Award scholarship and in part by NSERC Discovery Grant RGPIN-2023-03757.

Acknowledgments

Declared None.

References

[1] A. A. Khan, A. A. Laghari, A. M. Baqasah, R. Bacarra, R. Alroobaea, M. Alsafyani, et al., "BDLT-IoMT—a novel architecture: SVM machine learning for robust and secure data processing in internet of medical things with blockchain cybersecurity," *J. Supercomput.*, vol. 81, no. 1, pp. 1–22, 2025. [CrossRef]

[2] A. A. Khan, J. Yang, A. A. Laghari, A. M. Baqasah, R. Alroobaea, C. S. Ku, et al., "BAIoT-EMS: Consortium network for small-medium enterprises management system with blockchain and augmented intelligence of things," *Eng. Appl. Artif. Intell.*, vol. 141, p. 109838, 2025. [CrossRef]

[3] A. A. Khan, A. A. Laghari, S. A. Inam, S. Ullah, L. Nadeem, "A review on artificial intelligence thermal fluids and the integration of energy conservation with blockchain technology," *Discov. Sustain.*, vol. 6, no. 1, pp. 1–18, 2025. [CrossRef]

[4] A. A. Khan, A. A. Laghari, A. M. Baqasah, R. Alroobaea, T. R. Gadekallu, G. A. Sampedro, et al., "ORAN-B5G: A next generation open radio access network architecture with machine learning for beyond 5G in industrial 5.0," *IEEE Trans. Green Commun. Netw.*, vol. 8, no. 3, pp. 1026–1036, 2024. [CrossRef]

[5] A. A. Khan, A. A. Laghari, R. Alroobaea, A. M. Baqasah, M. Alsafyani, R. Bacarra, et al., "Secure remote sensing data with blockchain distributed ledger technology: a solution for smart cities," *IEEE Access*, vol. 12, pp. 69383–69396, 2024. [CrossRef]

[6] A. A. Khan, S. Dhabi, J. Yang, W. Alhakami, S. Bourouis, L. Yee, "B-LPoET: A middleware lightweight proof-of-elapsed time (PoET) for efficient distributed transaction execution and security on blockchain using multithreading technology," *Comput. Electr. Eng.*, vol. 118, p. 109343, 2024. [CrossRef]

[7] A. A. Khan, J. Yang, S. A. Awan, A. M. Baqasah, R. Alroobaea, Y.-L. Chen, et al., "Artificial intelligence, internet of things, and blockchain empowering future vehicular developments: A comprehensive multi-hierarchical lifecycle review," *Hum.-Centric Comput. Info. Sci.*, vol. 15, p. 13, 2025. [CrossRef]

[8] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, et al., "A survey on evaluation of large language models," *ACM Trans. Intelli. Syst. Technol.*, vol. 15, no. 3, pp. 1–45, 2024. [CrossRef]

[9] S. Fourati, W. Jaafar, N. Baccar, S. Alfattani, "XLM for autonomous driving systems: A comprehensive review," *arXiv preprint arXiv:2409.10484*, 2024. [CrossRef]

[10] Y. Wang, W. Chen, X. Han, X. Lin, H. Zhao, Y. Liu, et al., "Exploring the reasoning abilities of multimodal large language models (MLLMs): A comprehensive survey on emerging trends in multimodal reasoning," *arXiv preprint arXiv:2401.06805*, 2024. [CrossRef]

[11] C. Cui, Y. Ma, X. Cao, W. Ye, Y. Zhou, K. Liang, et al., "A survey on multimodal large language models for autonomous driving," in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, Waikoloa, HI, pp. 958–979, 2024. [CrossRef]

[12] L. Wen, D. Fu, X. Li, X. Cai, T. Ma, P. Cai, et al., "DiLu: A knowledge-driven approach to autonomous driving with large language models," *arXiv preprint arXiv:2309.16292*, 2023. [CrossRef]

[13] X. Ding, J. Han, H. Xu, W. Zhang, X. Li, "HiLM-D: Towards high-resolution understanding in multimodal large language models for autonomous driving," *arXiv preprint arXiv:2309.05186*, 2023. [CrossRef]

[14] J. Yuan, S. Sun, D. Omeiza, B. Zhao, P. Newman, L. Kunze, et al., "RAG-Driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model," in *Robotics: Science and Systems (RSS)*, pp. 1–14, Daegu, 2024. [CrossRef]

[15] T. Wang, E. Xie, R. Chu, Z. Li, P. Luo, "DriveCoT: Integrating chain-of-thought reasoning with end-to-end driving," *arXiv preprint arXiv:2403.16996*, 2024. [CrossRef]

[16] W. Wang, J. Xie, C. Hu, H. Zou, J. Fan, W. Tong, et al., "DriveMLM: Aligning multi-modal large language models with behavioral planning states for autonomous driving," *arXiv preprint arXiv:2312.09245*, 2023. [CrossRef]

[17] Y. Huang, J. Sansom, Z. Ma, F. Gervits, J. Chai, "DriVLMe: exploring foundation models as autonomous driving agents that perceive, communicate, and navigate," in *Vision and Language for Autonomous Driving and Robotics Workshop*, Seattle, WA, 2024. [CrossRef]

[18] G. Liao, J. Li, X. Ye, "VLM2Scene: Self-supervised image-text-LiDAR learning with foundation models for autonomous driving scene understanding," in *AAAI Conference on Artificial Intelligence*, Vancouver, BC, 2024, vol. 38, no. 4, pp. 3351–3359. [CrossRef]

[19] S. Wang, Z. Yu, X. Jiang, S. Lan, M. Shi, N. Chang, et al., "OmniDrive: A holistic LLM-agent framework for autonomous driving with 3D perception, reasoning and planning," *arXiv preprint arXiv:2405.01533*, 2024. [CrossRef]

[20] H. Tian, K. Reddy, Y. Feng, M. Quddus, Y. Demiris, P. Angeloudis, "Enhancing autonomous vehicle training with language model integration and critical scenario generation," *arXiv preprint arXiv:2404.08570*, 2024. [CrossRef]

[21] J. Mao, J. Ye, Y. Qian, M. Pavone, Y. Wang, "A language agent for autonomous driving," in *Conference on Language Modeling (COMS)*, Philadelphia, PA, 2024. [CrossRef]

[22] J. Zhang, Z. Huang, A. Ray, E. Ohn-Bar, "Feedback-guided autonomous driving," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, 2024, pp. 15000–15011. [CrossRef]

[23] H. I. Ashqar, T. I. Alhadidi, M. Elhenawy, N. O. Khanfar, "Leveraging multimodal large language models (MLLMs) for enhanced object detection and scene understanding in thermal images for autonomous driving systems," *Automation*, vol. 5, no. 4, pp. 508–526, 2024. [CrossRef]

[24] S. Luo, W. Chen, W. Tian, R. Liu, L. Hou, X. Zhang, et al., "Delving into multi-modal multi-task foundation models for road scene understanding: From learning paradigm perspectives," *IEEE Trans. Intelli. Veh. (Early Access)*, pp. 1–25, 2024. [CrossRef]

[25] R. Li, Z. Zhang, C. He, Z. Ma, V. M. Patel, L. Zhang, "Dense multimodal alignment for open-vocabulary 3D scene understanding," in *The European Conference on Computer Vision*. Berlin: Springer, 2024, pp. 416–434. [CrossRef]

[26] C. Zhong, S. Cheng, M. Kasoar, R. Arcucci, "Reduced-order digital twin and latent data assimilation for global wildfire prediction," *Natural Hazards Earth Syst. Sci.*, vol. 23, no. 5, pp. 1755–1768, 2023. [CrossRef]

[27] Z. Xu, Y. Zhang, E. Xie, Z. Zhao, Y. Guo, K.-Y. K. Wong, et al., "DriveGPT4: Interpretable end-to-end autonomous driving via large language model," *IEEE Robot. Autom. Lett.*, vol. 9, no. 10, pp. 8186–8193, 2024. [CrossRef]

[28] S. Cheng, H. Chassagnon, M. Kasoar, Y. Guo, R. Arcucci, "Deep learning surrogate models of jules-inferno for wildfire prediction on a global scale," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 9, no. 1, pp. 444–454, 2025. [CrossRef]

[29] C. Cui, Y. Ma, X. Cao, W. Ye, Z. Wang, "Receive, reason, and react: Drive as you say, with large language models in autonomous vehicles," *IEEE Intell. Transp. Syst. Mag.*, vol. 16, no. 4, pp. 81–94, 2024. [CrossRef]

[30] Z. Li, W. Wang, Y. Cai, X. Qi, P. Wang, D. Zhang, et al., "UnifiedMLLM: Enabling unified representation for multi-modal multi-tasks with large language model," *arXiv preprint arXiv:2408.02503*, 2024. [CrossRef]

[31] S. Fourati, (2024) *MLLM applied to autonomous driving across various weather conditions*. [Online]. Available: https://github.com/SondaFourati/MLLM-applied-to-autonomous-driving-across-various-weather-conditions/tree/main.

[32] R. Islam and O. M. Moushi, "GPT-4o: The cutting-edge advancement in multimodal LLM," *Authorea Preprints*, 2024. [CrossRef]

[33] L. Chen, O. Sinavski, J. Hünermann, A. Karnsund, A. J. Willmott, D. Birch, et al., "Driving with LLMs: Fusing object-level vector modality for explainable autonomous driving," in *IEEE International Conference on Robotics and Automation (ICRA)*, Yokohama, 2024, pp. 14093–14100. [CrossRef]

[34] Y. Jin, R. Yang, Z. Yi, X. Shen, H. Peng, X. Liu, et al., "SurrealDriver: Designing LLM-powered generative driver agent framework based on human drivers' driving-thinking data," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Abu Dhabi, 2024, pp. 966–971. [CrossRef]

[35] R. Yang, X. Zhang, A. Fernandez-Laaksonen, X. Ding, J. Gong, "Driving style alignment for LLM-powered driver agent," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Abu Dhabi, 2024, pp. 11318–11324. [CrossRef]

[36] Z. Guo, A. Lykov, Z. Yagudin, M. Konenkov, D. Tsetserukou, "Co-driver: VLM-based autonomous driving assistant with human-like behavior and understanding for complex road scenes," *arXiv preprint*, arXiv:2405.05885v1, 2024. [CrossRef]

[37] D. Wu, W. Han, Y. Liu, T. Wang, C.-z. Xu, X. Zhang, et al., "Language prompt for autonomous driving," in *AAAI Conference on Artificial Intelligence*, Philadelphia, PA, 2025, vol. 39, no. 8, pp. 8359–8367. [CrossRef]

[38] D. Fu, W. Lei, L. Wen, P. Cai, S. Mao, M. Dou, et al., "Limsim++: A closed-loop platform for deploying multimodal LLMs in autonomous driving," in *IEEE Intelligent Vehicles Symposium (IV)*, Jeju Island, 2024, pp. 1084–1090. [CrossRef]