IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

# DriveAgent: Multi-Agent Structured Reasoning with LLM and Multimodal Sensor Fusion for Autonomous Driving

Xinmeng Hou[2,*], Wuqi Wang[1,*], Long Yang[1], Hao Lin[3], Jinglun Feng[4,†], Haigen Min[1,†], Xiangmo Zhao[1]

***Abstract*— We introduce DriveAgent, a modular multi-agent autonomous driving framework that leverages large language model (LLM) reasoning combined with multimodal sensor fusion for autonomous driving. DriveAgent orchestrates specialized agents operating on camera, Light Detection and Ranging (LiDAR), Inertial Measurement Unit (IMU), and Global Positioning System (GPS) with LLM-driven analytical processes to deliver temporally aligned perception, causal reasoning, and action recommendations. The framework operates through a modular agent-based pipeline comprising four principal modules: (i) a descriptive analysis agent identifying critical sensor data events based on filtered timestamps, (ii) dedicated vehicle-level analysis conducted by LiDAR and vision agents that collaboratively assess vehicle conditions and movements, (iii) environmental reasoning and causal analysis agents explaining contextual changes and their underlying mechanisms, and (iv) an urgency-aware decision-generation agent prioritizing insights and proposing timely maneuvers. This modular design empowers the LLM to effectively coordinate specialized perception and reasoning agents, delivering cohesive, interpretable insights into complex autonomous driving scenarios. Extensive experiments demonstrate that DriveAgent substantially outperforms baseline methods, achieving a 26.31% improvement in vehicle reasoning and consistent enhancements of up to 2.85% in environmental reasoning. These results highlight the effectiveness of our LLM-driven multi-agent sensor fusion framework in boosting the robustness and reliability of autonomous driving systems.** [1]

## I. Introduction

Promising progress has been made in autonomous driving (AD) in recent years; however, some challenging problems in AD have yet to be solved, especially under dynamic, multimodal environments, such as contextual understanding and interpretability [1]. Commonly adopted AD architectures, whether modular or end-to-end, often struggle to integrate insights across heterogeneous sensor modalities—such as cameras, Light Detection and Ranging (LiDAR), Inertial Measurement Unit (IMU), and Global Positioning System (GPS)—especially in edge cases where visual information is ambiguous or missing [2].

Similarly to AD problems, strong capabilities in reasoning across diverse domains have been demonstrated with large language models (LLM) and vision-language models (VLM) in the past few years [3], [4]. Nevertheless, a key challenge, how to apply LLM into multimodal sensor fusion in driving scenarios, remains underexplored [5], [6]. Thus, an opportunity to enhance autonomous decision-making is presented by incorporating LLM-driven reasoning into sensor-rich driving pipelines.

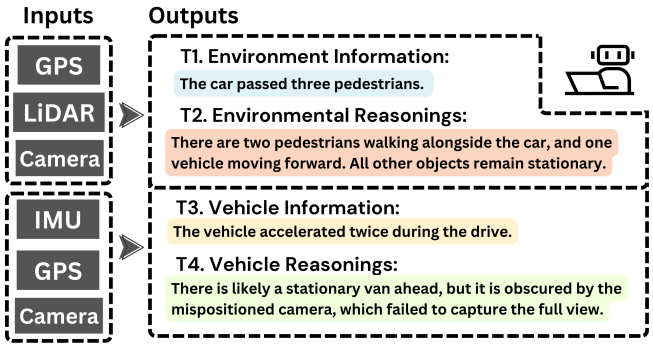


Fig. 1. Overview of the inputs and outputs for the proposed DriveAgent framework. DriveAgent takes multimodal sensor data as inputs, including camera, Lidar GPS, and IMU data. The input data are processed through four structured tasks (T1 to T4), supporting comprehensive reasoning tasks at both the environment level and the vehicle level as outputs.

Recent research has begun integrating LLMs into AD [7], primarily through structured visual reasoning [8], cooperative vehicle communication [9], and instruction-based fine-tuning [10]. However, these approaches often rely heavily on single-modality inputs, closed-loop planning, or basic reasoning techniques such as Chain-of-Thought (CoT) [11], limiting their effectiveness in challenging scenarios involving unreliable visual sensors.

Motivated by the aforementioned limitations, we introduce **DriveAgent**—a modular, LLM-driven multi-agent framework designed to reason over multimodal sensor streams in autonomous driving scenarios. DriveAgent integrates camera, LiDAR, GPS, and IMU data through a hierarchy of specialized agents that perform perception, reasoning, and decision-making tasks in a coordinated manner. Our framework uniquely leverages the structured compositionality of LLMs alongside domain-specific sensor processing modules, enabling clear and reliable decision-making in diverse and complex driving situations. Unlike prior works that focus on end-to-end planning or vision-language alignment alone [12],

[1] Wuqi Wang, Long Yang, Haigen Min, and Xiangmo Zhao are with Chang'an University, Xi'an, Shaanxi, China.

[2] Xinmeng Hou is with Chang'an University, Xi'an, Shaanxi, China and Agency for Science, Technology and Research (A*STAR), Singapore.

[3] Hao Lin is with University of California, Davis, USA.

[4] Jinglun Feng is with CCNY Robotics Lab, The City College of New York, USA. * Equally contributed. † Corresponding authors.

Jinglun Feng:jfeng1@ccny.cuny.edu,Haigen Min:hgmin@@chd.edu.cn

This work is supported in part by the National Natural Science Foundation (52441205), National Natural Science Foundation of China (No.52372426), Shaanxi Province Innovation Capability Support Plan-Innovative Talent Promotion Plan (No.2023KJXX-020), Natural Science Foundation of Shaanxi Province (No.2022JQ-663), Key Research and Development Program of Shaanxi Province (2024GX-YBXM-261), and Fundamental Research Funds for the Central Universities, CHD (No.300102243202, 300102244713).

[1]Code available at https://github.com/Paparare/DriveAgent

[13], DriveAgent provides a generalized approach to explain vehicle behavior, environmental dynamics, and causal events across multiple sensor modalities.

Fig. 1 illustrates our proposed study's scope, showing how multimodal sensor inputs (e.g., camera, LiDAR, GPS, and IMU data) and text data support both vehicle-level and environmental-level tasks. Our contributions include:

1) **Multi-Modal Agent System**: The proposed multi-modal agent system enables cohesive, end-to-end reasoning in complex driving contexts.
2) **Vision-Language Model Fine-tune Strategy**: The proposed fine-tuned VLM enables abilities including object detection and traffic interpretation for the proposed system.
3) **Self-Reasoning Benchmarks**: Autonomous driving performance is evaluated based on tasks such as data analysis, visual reasoning, and integrated environment understanding.
4) **Three-Tier Driving Dataset**: The collected dataset represents standard, typical, and challenged AD scenarios, offering distinct challenges for comprehensive training and evaluation.

## II. Related Work

AI agents for environmental understanding have evolved from traditional SLAM to modern foundation model-based systems. Vision-Language-Action models unify multimodal perception with action generation, with systems like OpenVLA achieving superior performance through dual vision encoders and large-scale training data [14]. Embodied scene understanding requires spatial awareness and egocentric reasoning for safe action selection, with benchmarks like MetaVQA showing improvements in spatial reasoning and successful sim-to-real transfer [15]. Multi-agent systems enable distributed environmental understanding through cooperative perception and coordinated decision-making, excelling in vehicle-to-vehicle and vehicle-to-infrastructure scenarios [16], [17]. Foundation models contribute through multimodal reasoning and code generation, while generative AI addresses rare event synthesis for robust environmental understanding [18], [19]. Modern AI-SLAM integration combines geometric approaches with semantic understanding, though challenges remain in real-time processing and dynamic adaptation [20], [21]. Current challenges include cross-domain generalization, rare event handling, and simulation-to-reality transfer, with future research focusing on uncertainty quantification and standardized evaluation frameworks for embodied AI systems. Despite significant advances in AI agents for environmental understanding, existing approaches lack integrated multi-modal reasoning systems that can seamlessly combine visual perception, language understanding, and decision-making in complex driving scenarios. Current methods typically employ modular architectures that process different modalities separately, limiting their ability to perform cohesive end-to-end reasoning across diverse driving contexts and challenging scenarios.

## III. Methodology

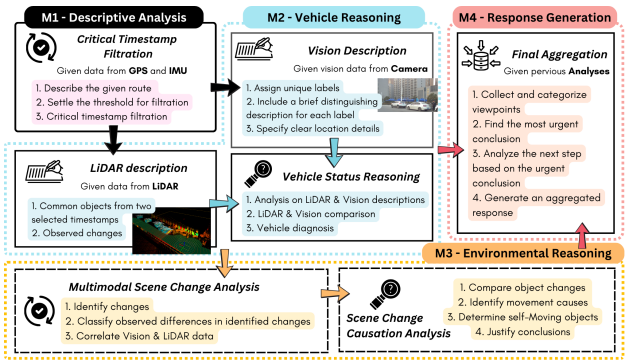


Fig. 2. An overview of the proposed architecture which is consisting of four modules (M1 to M4), where multimodal sensor inputs—camera, IMU, GPS and LiDAR—enable both environment-level tasks (e.g., information retrieval, environmental change detection, and reasoning) and vehicle-level tasks (e.g., vehicle status analysis, motion evaluation, and behavior pattern recognition).

Our approach addresses four key tasks through a structured reasoning process. Given an input instruction $\mathcal{I}$, the module $\mathcal{M}$ produces a response $\mathcal{R}$ in adherence to the prompt. To facilitate driving analysis, we design four sequential modules as demonstrated in Fig. 2: **(1) Descriptive Analysis**, **(2) Vehicle Reasoning**, **(3) Environmental Reasoning**, and **(4) Response Generation**.

In the initial phase, the system identifies $n$ critical timestamps corresponding to significant events. We represent these timestamps and their triggering factors as $\{(T_i, F_i)\}_{i=0}^{n}$, where $T_i$ indicates the timestamp and $F_i$ describes why it was selected. This set serves as the foundation for all subsequent analysis. The vehicle reasoning phase involves two sensor-specific agents and one aggregator agent. The LiDAR agent $\mathcal{M}_L$ generates triplets $\{(T_i, F_i, L_i)\}_{i=0}^{n}$, where $L_i$ is the LiDAR description at timestamp $T_i$. Similarly, the vision agent $\mathcal{M}_V$ produces triplets $\{(T_i, F_i, V_i)\}_{i=0}^{n}$, where $V_i$ is the vision-based description. The aggregation agent $\mathcal{M}_D$ compares each LiDAR description $L_i$ with its corresponding vision description $V_i$ to detect potential vehicle anomalies, denoted as $D_i$. Simultaneously, an environmental reasoning agent analyzes consecutive sensor descriptions ($V_i$ and $L_i$) to detect environmental changes, producing variations $\{E_2, E_3, \ldots, E_n\}$. A causal analysis agent $\mathcal{M}_C$ then examines these variations to determine underlying causes, highlighting objects requiring caution as $C_i$. Finally, the response aggregation agent $\mathcal{M}_R$ combines vehicle diagnostics ($D_i$) and caution flags ($C_i$) into a cohesive response $\mathcal{R}_i$ for each critical timestamp $T_i$. Each response $\mathcal{R}_i$ thus integrates both vehicle status information and environmental insights necessary for informed decision-making.

### A. *Module 1: Descriptive Analysis*

Determining which information is essential for accurate route description is a fundamental challenge. We address this by using a self-referential filtration system that automatically selects critical timestamps based on vehicle motion. The

filtration threshold is set by an LLM agent that analyzes standard route descriptions from both real-world and simulated autonomous drives on predefined paths. A single agent handles both route classification and threshold determination.

We categorize driving routes based on their speed $S$ and an urban complexity indicator $U$. Specifically, we define the function $\mathcal{R}(S, U)$, which outputs both a route category $r_i$ and a corresponding threshold $\theta_i$. Formally (the double colon :: indicates this correspondence):

$$\mathcal{R}(S, U) \in \{\, r_1 :: \theta_1,\ r_2 :: \theta_2,\ r_3 :: \theta_3\}, \tag{1}$$

where $r_1$ represents high-speed, low-complexity routes, $r_2$ represents medium-speed, medium-complexity routes, and $r_3$ represents variable-speed, high-complexity routes. For each category $r_i$, $\theta_i$ is computed by an agent function $G$:

$$\theta_i = G(S, U, r_i), \tag{2}$$

which tailors standard kinematic baselines (angular velocity of $10^\circ/s$, linear acceleration of $8\,m/s^2$, and yaw rate of $10^\circ/s$) to the specific speed $S$ and urban complexity $U$. By monitoring these kinematic signals, such as turning, acceleration/braking, and orientation changes, the filtration agent efficiently pinpoints critical timestamps reflecting significant motion changes.

To support vehicle motion understanding and sensor synchronization, the DriveAgent framework integrates GPS and IMU data in addition to camera and LiDAR. IMU measurements—such as angular velocity, acceleration, and yaw rate—are used to compute kinematic signals that guide trajectory segmentation and critical timestamp selection.

GPS data contribute by (1) providing global motion references through GNSS/INS fusion and (2) offering timestamps for multi-sensor synchronization. These two modalities ensure accurate motion profiling and consistent temporal alignment across sensors, enabling reliable reasoning in subsequent modules.

### B. *Module 2: Vehicle Reasoning*

The Vehicle Reasoning module comprises three agents: one processing vision data, one processing LiDAR data, and an analyzer agent that synthesizes both to detect vehicle abnormalities. The designed reasoning pipeline is shown in Algorithm 1.

*1) Vision Descriptor:* The vision agent assigns unique indices to all detected objects in the camera view. For two consecutive frames at times $t$ and $t+1$, it records each object's positions as $p_i(t)$ and $p_i(t+1)$, respectively. By comparing these positions, the agent calculates the displacement for each object, noted as $p_i(t) \sim p_i(t{+}1)$. This analysis summarizes object motion between the frames, identifying which objects moved and by how much.

*2) LiDAR Descriptor:* The LiDAR agent initially identifies objects and their positions relative to the vehicle from the LiDAR point cloud. If multiple objects share the same label, it resolves ambiguities using spatial separation or distinct features, ensuring each object $i$ is uniquely identified. For consecutive timestamps $t$ and $t+1$, the agent records positions $p_i(t)$ and $p_i(t+1)$ for each object and computes their displacement as:

$$\Delta p_i = p_i(t+1) - p_i(t) \tag{3}$$

*3) Vehicle Status Reasoning:* The analyzer agent combines outputs from the vision and LiDAR descriptors to assess the vehicle's condition and sensor reliability. It first filters out objects beyond a 100 m LiDAR range, forming the object set $\Omega = \{i \mid \|\mathbf{L}_i(t)\| \le 100\}$, where $\mathbf{L}_i(t)$ denotes the LiDAR-based position of object $i$ at time $t$. This filtering step focuses on relevant nearby objects and identifies potential LiDAR anomalies (e.g., missing or ghost objects, range errors). For each object $i \in \Omega$, the agent compares LiDAR positions $\mathbf{L}_i(t)$ with camera-derived positions $\mathbf{C}_i(t)$, quantifying cross-sensor consistency through the Euclidean distance:

$$\Delta_i(t) = \big\|\mathbf{L}_i(t) \sim \mathbf{C}_i(t)\big\| \tag{4}$$

If $\Delta_i(t)$ is large for an object, it indicates a discrepancy between LiDAR and camera data, potentially due to calibration errors or sensor noise. The agent also monitors if many objects simultaneously exhibit large $\Delta_i(t)$ values, suggesting broader sensor misalignment or issues like camera blur or drift. After these analyses, the agent compiles a comprehensive status report highlighting specific LiDAR anomalies and camera issues (e.g., inaccurate object localization).

**Algorithm 1** Vehicle Reasoning

**Require:** $\{p_i(t)\}$ from vision at $t = 1, \ldots, T$, $\{p_i(t)\}$ from LiDAR at $t = 1, \ldots, T$, $\mathbf{L}_i(t), \mathbf{C}_i(t)$ (LiDAR/camera positions), $R$ (distance threshold, e.g. 100 m)

**for all** $t \in \{1, \ldots, T-1\}$ **do**
  **for all** $i$ **do**
    $p_i(t) \sim p_i(t+1)$ {Associate object $i$ positions at time $t$ and $t+1$ (track across frames).}
  **end for**
  **for all** $i$ **do**
    $\Delta p_i = p_i(t+1) - p_i(t)$ {Compute displacement of object $i$ between $t$ and $t+1$.}
  **end for**
  $\Omega \leftarrow \{\, i \mid \|\mathbf{L}_i(t)\| \le R \,\}$ {Identify objects within LiDAR range $R$ for analysis.}
  **for all** $i \in \Omega$ **do**
    $\Delta_i(t) = \big\|\mathbf{L}_i(t) - \mathbf{C}_i(t)\big\|$ {Calculate LiDAR–camera positional discrepancy for object $i$.}
  **end for**
**end for**
**return** $\{\Delta p_i,\ \Omega,\ \Delta_i(t)\}$ {Return all computed displacements, the in-range object set, and cross-sensor differences.}

### C. *Module 3: Environmental Reasoning*

The environmental reasoning module comprises two coordinated agents: one detecting and characterizing environmental changes, and the other analyzing the underlying causes of

**Algorithm 2** Environmental Reasoning

**Require:** $\mathcal{V}(t), \mathcal{V}(t-1)$ (vision detections at times $t$ and $t-1$), $\mathcal{L}(t), \mathcal{L}(t-1)$ (LiDAR detections at $t$ and $t-1$), $\mathbf{O}_i(t)$ (state of object $i$ at time $t$), $\Delta t$ (time interval for long-term change detection)

**for all** $v_i(t) \in \mathcal{V}(t)$ **do**
  $\Delta v_i(t) = v_i(t) - v_i(t-1)$ {Compute change in vision detection $v_i$ since the last timestep.}
**end for**
**for all** $\ell_j(t) \in \mathcal{L}(t)$ **do**
  $\Delta \ell_j(t) = \ell_j(t) - \ell_j(t-1)$ {Compute change in LiDAR detection $\ell_j$ since the last timestep.}
**end for**
$\Delta_{i,j}(t) = \left\| v_i(t) - \ell_j(t) \right\|$ {Calculate cross-sensor difference for a matched vision–LiDAR observation.}
**for all** $\mathbf{O}_i(t)$ **do**
  $\Delta \mathbf{O}_i(t) = \mathbf{O}_i(t) - \mathbf{O}_i(t - \Delta t)$ {Measure long-term state change of object $i$ over interval $\Delta t$.}
**end for**
**return** $\{\Delta v_i(t),\ \Delta \ell_j(t),\ \Delta_{i,j}(t),\ \Delta \mathbf{O}_i(t)\}$ {Return detected object changes and consistency metrics for analysis.}

these changes. Together, these agents provide comprehensive insights into environmental dynamics, as detailed in Algorithm 2.

*1) Environmental Reasoning:* This agent detects environmental changes by comparing sensor readings at consecutive timestamps. Let $\mathcal{V}(t) = \{v_1(t), v_2(t), \ldots, v_m(t)\}$ and $\mathcal{L}(t) = \{\ell_1(t), \ell_2(t), \ldots, \ell_n(t)\}$ represent vision and LiDAR detections at time $t$, respectively. By examining differences between $\mathcal{V}(t)$ and $\mathcal{V}(t-1)$, and between $\mathcal{L}(t)$ and $\mathcal{L}(t-1)$, the agent identifies new, missing, or significantly relocated objects. Changes are categorized by type (static or dynamic) and severity. For each identified change, cross-sensor consistency is assessed. Specifically, for an object observed by both sensors, the sensor agreement is measured by the Euclidean distance:

$$\Delta_{i,j}(t) = \|v_i(t) - \ell_j(t)\| \tag{5}$$

Note that a small $\Delta_{i,j}(t)$ indicates that the vision and LiDAR agree on the object's position, whereas a large $\Delta_{i,j}(t)$ could signal sensor misalignment, calibration issues, or an actual abrupt environmental change that one sensor registers differently than the other.

While both **Vehicle Reasoning** (Module 2) and **Environmental Reasoning** (Module 3) use Euclidean distances between LiDAR and camera observations, the former evaluates sensor consistency at a single time point, while the latter focuses on validating environmental changes over time.

*2) Causal Analysis:* This agent investigates the causes behind the detected environmental changes. It first retrieves each object's state from previous reasoning steps or raw sensor data, represented by $\mathbf{O}_i(t)$ at time $t$. The agent then examines how each object's state evolves over an extended interval $\Delta t$ by computing:

$$\Delta \mathbf{O}_i(t) = \mathbf{O}_i(t) - \mathbf{O}_i(t - \Delta t) \tag{6}$$

For each flagged change, the agent infers plausible causes by analyzing temporal patterns (e.g., sudden vs. gradual), environmental cues (e.g., wind or collisions), and surrounding context (e.g., nearby object motion). It classifies each change as either *self-moving* (e.g., vehicles or pedestrians) or *externally influenced* (e.g., displaced by force), using cues like mobility features and motion behavior. To resolve whether a large $\Delta_{i,j}(t)$ in Eq. (5) is due to sensor error or actual scene change, our framework combines short-term sensor consistency checks with longer-term motion tracking. Specifically, persistent discrepancies across time suggest sensor issues, while consistent object displacement confirmed by causal analysis indicates genuine environmental change. This cross-checking enables robust and interpretable reasoning. The agent then compiles a causal report summarizing the changes, inferred origins, and confidence levels, enabling informed downstream decision-making with interpretable reasoning.

### D. *Module 4: Response Generation*

This module synthesizes outputs from previous agents to generate a prioritized response, each insight $a_i$ is paired with a category $c_i$ (e.g., safety, efficiency), forming the set $\mathcal{A} = \{(a_i, c_i)\}_{i=1}^{N}$. A scoring function $\Psi(a_i, c_i)$ evaluates urgency, and the highest-priority issue, $\hat{a}$, can be identified as:

$$\hat{a} = \arg \max_{(a_i, c_i) \in \mathcal{A}} \Psi(a_i, c_i) \tag{7}$$

The agent then selects the best response $\phi^*$ from a candidate set $\Phi(\hat{a}) = \{\phi_1, \ldots, \phi_M\}$ by maximizing a utility function:

$$\phi^* = \arg \max_{\phi_j \in \Phi(\hat{a})} \text{Score}(\phi_j) \tag{8}$$

The final response is:

$$\mathcal{R} = \left(\hat{a}, \phi^*, \mathcal{A}^-\right) \tag{9}$$

where $\mathcal{A}^- = \mathcal{A} \setminus \{\hat{a}\}$ denotes secondary insights. This structured output can gather the top-priority issue, proposed action, and remaining considerations to support transparent and interpretable decision-making.

## IV. Experiments

### A. Datasets

Due to the lack of public datasets for evaluating an agent's understanding of driving environments, we introduce a new dataset collected from an autonomous vehicle in real-world scenarios [22]. As shown in Fig. 3, the vehicle was equipped with multiple sensors and a navigation system [2].

[2]Datasets available at `https://huggingface.co/datasets/Parechan/driveagent`

| Sensor | Model | Frequency | Specifications |
|---|---|---|---|
| LiDAR | Robosense Ruby-128 | 10 Hz | 128-beam, Maximum detection range: 230 m, Vertical Field-of-View (FOV): -25° to +15°, Horizontal FOV: 360° |
| Camera | Basler acA2440-75uc | 10 Hz | Default resolution: 2448×2048 |
| GNSS/INS | Xsens MTI-680G | 4 Hz (GNSS), 400 Hz (IMU/RTK ) | 0.2° roll/pitch, 0.5° Yaw/Heading, cm-level position accuracy, Internal u-blox ZED F9 RTK-enabled GNSS receiver |

TABLE I

SENSOR SPECIFICATIONS FOR THE CHANG'AN UNIVERSITY (XINDA) AUTONOMOUS VEHICLE, INCLUDING GLOBAL NAVIGATION SATELLITE SYSTEM/INERTIAL NAVIGATION SYSTEM (GNSS/INS), REAL-TIME KINEMATIC (RTK), AND OTHER MODALITIES.

All sensor data were time-synchronized for consistent multi-modal observations. Sensor specifications are provided in Table I.


Fig. 3. Data collection vehicle sensor configuration and satellite images of recorded driving routes. There are three routes being recorded in total at Chang'an University, Xi'an, China. Route 1 (R1) is shown in red trajectory, Route 2 (R2) is shown in purple trajectory, while Route 3 (R3) is shown in green trajectory.

| Attribute | R1 | R2 | R3 |
|---|---|---|---|
| Length (m) | 1277.76 | 969.19 | 1125.91 |
| Max Speed (m/s) | 13.90 | 11.40 | 12.09 |
| Average Speed (m/s) | 7.30 | 4.17 | 4.29 |
| Environment Dynamic Level | Small | Large | Medium |
| Roadside Obstructions | ✗ | ✗ | ✓ |
| Righ & Left-side Camera Views | ✗ | ✓ | ✓ |

TABLE II

DETAILED ATTRIBUTES FOR ROUTES R1, R2, AND R3.

Moreover, as summarized in Table II, our dataset covers three distinct driving routes: R1, R2, and R3. R1 spans 1277.76 meters and was recorded in a controlled environment, serving as the baseline scenario. The ego vehicle reached a maximum speed of $13.90m/s$ with an average speed of $7.30m/s$, and only forward-facing images were captured. The environment dynamic level for R1 is qualitatively described as Small, reflecting relatively simple traffic conditions. R2, measuring 969.19 meters in length, features a loop around an urban square and is qualitatively described as having a Large environment dynamic level, indicating a more complex and active driving environment. The maximum and average speeds along R2 were $11.40m/s$ and $4.17m/s$ respectively. Compared to R1, R2 includes right and left-side camera views, providing a broader field of view. R3, at 1125.91 meters, introduces additional environmental complexity, with roadside obstructions and is qualitatively described as having a Medium environment dynamic level, indicating moderately active traffic with added structural challenges. The maximum speed recorded was $12.09m/s$, with an average speed of $4.29m/s$. Similar to R2, R3 captures views from the right, left, and front cameras.

In addition, an enhanced detection method, combined with PointPillars architecture [23] and a clustering strategy, was used to perform real-time perception on LiDAR observation results and detect objects.

### *B. Task and Evaluation Metrics*

We define three primary tasks: (1) **Object and Category Detection**, (2) **Vehicle Reasoning** (LiDAR and visual understanding), and (3) **Environmental Reasoning**. Each task is validated by its contribution to scene understanding, decision-making, and system robustness, with results discussed in Section V.

For object identification task, we consider seven key categories: *four-wheel vehicles* (the principal motorized participants on roads), *non-four-wheel vehicles* (e.g., bicycles and scooters, which often pose higher risk due to less coverage), *pedestrians* (vulnerable road users who commonly receive priority), *signs* (official traffic instructions and regulations), *fixed installations* (permanent structures, barriers, or buildings), *plants* (vegetation that may obscure visibility or mark boundaries), and *monitors* (electronic displays or cameras supporting traffic supervision). This task is trained on datasets R2 and R3 and evaluated on R1, using precision, recall, and F1 as metrics; its importance lies in ensuring the accurate classification of objects critical for traffic safety.

The vehicle-reasoning task include two tasks: a LiDAR understanding task, evaluated by comparing the model's output with ground-truth labels in R2, and a vision-based reasoning task, assessed on R2 and R3, where misaligned camera views serve as distractors. These evaluations measure real improvements in perception accuracy and prevent false gains

from random guessing. Finally, the environmental reasoning task tests the system's ability to tell apart stationary objects from independently moving ones (like pedestrians), with improvements validated through better situational awareness, collision avoidance, and safer navigation in dynamic traffic.

### C. Baseline Approaches

For *task 1*, we benchmark five leading vision-language models including LLaMA-3.2-Vision-Instruct [24], GPT-4o-mini [25], Pixtra-large [26], GPT-4o [25], and Claude-3.7-Sonnet [27]—selected for their strong performance, diverse architectures, and proven effectiveness on vision tasks.

For *tasks 2 & 3*, we adopt three baseline methods: (1) Zero-Shot [28], which has been applied in car-following prediction [10]; (2) CoT, employed in autonomous driving applications [29], [30]; and (3) CoT + Self-Refine [31], which approximates our multi-agent structure by incorporating self-refinement capabilities. Zero-Shot tests direct inference ability, CoT adds step-by-step reasoning, and CoT + Self-Refine further improves reasoning through iterative refinement.

### D. Reasoning Instructions

Fig. 4 presents the structured annotation guidelines used to define high-quality responses, emphasizing three aspects: (1) accurate identification of dynamic objects (e.g., vehicles, bicycles), (2) inclusion of static infrastructure such as lane markings, traffic signs, and signals, and (3) objective, concise descriptions without subjective or irrelevant content.

To evaluate output quality, model-generated descriptions are compared to reference responses based on these guidelines. The evaluation focuses on content accuracy and category coverage across five key scene elements: Trees, Buildings, Vehicles, Pedestrians, and Signs—chosen for their importance in road-scene understanding and prevalence in standard AD datasets.

**Reasoning Setup:**

The reasoning experiments follow the multi-phase methodology outlined in Section III, where DriveAgent executes four sequential modules: Descriptive Analysis, Vehicle Reasoning, Environmental Reasoning, and Response Generation. At each stage, the system produces a response based on outputs from the previous module, resulting in four stepwise generations per input case. Evaluation focuses on two key aspects: (1) the accuracy of vehicle diagnostic reasoning, and (2) the accuracy of environmental and causal inference.

**VLM Implementation Details:** The VLM in DriveAgent is built upon the LLaMA-3.2-Vision model (11B parameters), combining a pre-trained LLaMA vision encoder with the LLaMA-3.2 language model (11B). A learnable linear projection layer aligns visual features with the language model's input. Both the vision encoder and language model are fine-tuned using Low-Rank Adaptation (LoRA) [32], while the projection layer is trained from scratch without LoRA. Experiments are conducted on an NVIDIA H100 GPU server, optimized via AdamW with a $2 \times 10^{-4}$ learning rate and batch size of 2. Training spans 10 epochs with a cosine learning rate schedule and a 3% warm-up period. We utilize structured JSON-formatted prompts as instruction-response pairs containing special <Image> tokens for visual inputs, fine-tuning the model through supervised instruction tuning.

## V. Results and Analysis

### A. Object and Category Detection Performance

In this subsection, we first evaluate the task 1 introduced at Section IV-B. Table III illustrates the substantial performance gains achieved when training with structured annotation guidelines aimed at better and more accurate object identification. The baseline LLaMA-3.2-Vision-Instruct model achieves moderate performance (Precision = $64.33\%$, Recall = $35.26\%$, F1-score = $45.55$). However, once annotation guidelines are systematically applied, the VLM model in DriveAgent exhibits a significant leap in all key metrics—reaching a precision of $89.96\%$ and an F1-score of $71.62$, outperforming other models in the table.

| Model | Precision | Recall | F1-score |
|---|---|---|---|
| LLaMA-3.2-Vision-Instruct | 64.33 | 35.26 | 45.55 |
| GPT-4o-mini | 64.40 | 51.60 | 57.30 |
| pixtra-large | 76.12 | 54.63 | 60.86 |
| GPT-4o | 80.98 | 59.98 | 64.96 |
| claude-3.7-sonnet | 83.83 | 69.73 | 68.80 |
| DriveAgent | 89.96 | 63.52 | 71.62 |

TABLE III

Precision (%), Recall (%), and F1-score for each baseline methods and the proposed model on Object and Category Detection task. Red color highlights the best performance.

Fig. 5 shows that DriveAgent is the only model that consistently detects overhead monitors, which are often missed by other baselines due to their subtle visual cues. This highlights the value of precise and consistent annotations: by reducing ambiguity and enforcing uniform labeling standards, our new dataset enables more effective learning of object boundaries. As a result, DriveAgent achieves superior accuracy in object localization and identification, demonstrating that high-quality structured annotations are essential for robust detection performance.

### B. Reasoning Performance

**Vehicle Reasoning:** We first We first evaluate the Vehicle Reasoning task (LiDAR and Vision) introduced in Section IV-B, as shown in Table IV.

For LiDAR reasoning, the Zero-Shot approach achieves moderate accuracy ($47.50\%$–$65.05\%$) across routes, establishing a baseline for detecting sensor misplacement. CoT alone leads to substantial performance drops, suggesting basic sequential reasoning struggles with subtle errors. Adding Self Refine significantly improves accuracy, reaching $72.63\%$ on R2 and $63.89\%$ on R2-right. DriveAgent, however, achieves strong and stable performance, particularly on R2-left ($69.90\%$), demonstrating reliable LiDAR misplacement detection.

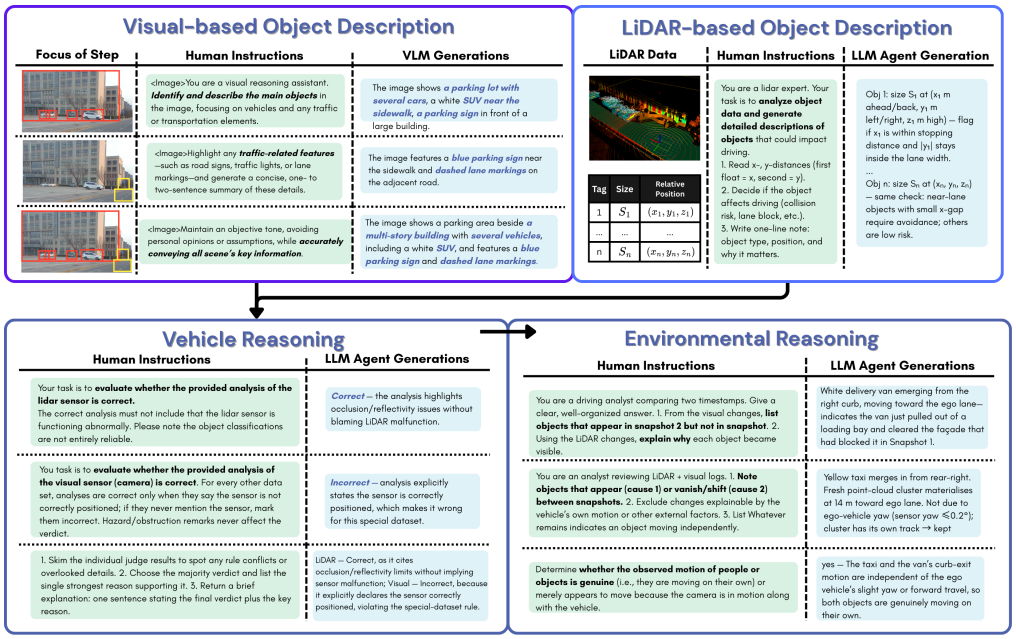


Fig. 4. Overview of the multimodal reasoning pipeline used for driving scene understanding. Visual descriptions are generated from camera images, focusing on identifying traffic-related objects and maintaining objective scene summaries. LiDAR-based descriptions analyze object sizes and relative positions to assess driving risk. In the reasoning stages, LLM agents evaluate the correctness of sensor-based analyses (vehicle reasoning) and identify environmental changes over time (environmental reasoning). Human instructions and corresponding LLM generations are provided for each step, supporting robust, explainable autonomous driving assessments.

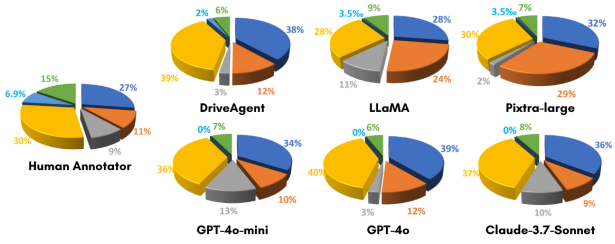


Fig. 5. Distribution of object categories in the human-annotated ground truth versus each model's predictions. Colour key: deep blue = fixed installations, orange = four-wheel vehicles, grey = non-four-wheel vehicles, yellow = plants, and light blue = monitors.

| Tasks | Method | R2 | R2-left | R2-right | R3 | R3-left | R3-right |
|---|---|---|---|---|---|---|---|
| Vehicle Understanding (Lidar) | Zero-Shot | 62.11 | 65.05 | 52.78 | 57.14 | 47.50 | 55.26 |
| | CoT | 15.79 | 22.33 | 15.74 | 18.10 | 20.00 | 13.16 |
| | CoT + Self Refine | 72.63 | 66.02 | 63.89 | 64.76 | 45.00 | 55.26 |
| | DriveAgent (ours) | 65.26 | 69.90 | 58.33 | 55.24 | 51.25 | 50.00 |
| Vehicle Reasoning (Vision) | Zero-Shot | 70.53 | 2.91 | 0.93 | 82.86 | 6.25 | 3.95 |
| | CoT | 69.47 | 25.24 | 27.78 | 79.05 | 31.25 | 40.79 |
| | CoT + Self Refine | 65.26 | 0.97 | 2.78 | 80.00 | 3.75 | 5.26 |
| | DriveAgent (ours) | 96.84 | 58.25 | 71.30 | 87.62 | 68.75 | 63.16 |
| Environmental Reasoning | Zero-Shot | 37.89 | - | - | 36.19 | - | - |
| | CoT | 56.84 | - | - | 62.86 | - | - |
| | CoT + Self Refine | 43.16 | - | - | 56.19 | - | - |
| | DriveAgent (ours) | 58.95 | - | - | 65.71 | - | - |
| Ablation | DriveAgent (w/o causation) | 56.28 | - | - | 63.77 | - | - |

TABLE IV

REASONING ACCURACY (%) IS REPORTED ACROSS MODALITIES AND REGIONS, WITH EACH TASK PRESENTING RESULTS FROM SEVERAL PROMPTING METHODS. THE LABELS R-LEFT* AND R-RIGHT* DENOTE THE LEFT- AND RIGHT-SIDE CAMERA VIEWS OF THE SAME ROUTE; THESE VIEWS ACT AS DISTRACTORS FOR THE VEHICLE-REASONING SUBTASK. RED COLOR HIGHLIGHTS THE BEST PERFORMANCE.

For Vision reasoning, detecting misaligned cameras is even more challenging. Zero-Shot and CoT show very low accuracies on left and right views. In contrast, DriveAgent achieves notable gains, including 96.84% accuracy on R2, and clear improvements across left and right variants (58.25% and 71.30%), confirming that modality-specific tuning is crucial for visual sensor reasoning.

**Environmental Reasoning:** At last, we evaluate the task 3 introduced in Section IV-B. The evaluation of environmental reasoning performance is based on the agent's ability to detect independently moving objects by comparing two selected timestamps. As shown in Table IV, the Zero-Shot performance is low (37.89% and 36.19%), indicating that without any additional reasoning cues the agent struggles with temporal object differentiation. The CoT method significantly improves performance, achieving accuracies of 56.84% and 62.86%. However, the performance of CoT + Self Refine strategy offers mixed results, where the performance drops to 43.16% for one set and recovers partially to 56.19% for the other, suggesting that the refinement process may not always synergize effectively with the inherent sequential reasoning of CoT in this task. Notably, our proposed DriveAgent model outperforms all baselines, obtaining the highest accuracies of 58.95% and 65.71% respectively. These results underscore the importance of a dedicated, well-tuned approach for integrating temporal and spatial reasoning, which is critical for accurately identifying independently moving objects in dynamic environments.

**Ablation Study:** To validate the effectiveness of our

change causation analysis module, we conduct an ablation study by removing this component from DriveAgent. This module distinguishes between changes caused by ego vehicle movement versus other independently moving objects. As shown in Table IV, removing this module results in performance drops from $58.95\%$ to $56.28\%$ and from $65.71\%$ to $63.77\%$ across both evaluation sets. This demonstrates that explicit causation analysis is crucial for accurately identifying independently moving objects in dynamic driving scenarios.

## VI. Conclusion

In this paper, we proposed *DriveAgent*, a modular, LLM-guided multi-agent framework for structured reasoning in autonomous driving. By fusing multimodal sensor data—camera, LiDAR, GPS, and IMU—through specialized perception and reasoning agents, DriveAgent enables robust interpretation of complex driving environments. Our four-stage pipeline—filtration, vehicle diagnostics, environmental reasoning, and response generation—delivers accurate, interpretable, and extensible decision-making. Experiments on real-world datasets show that DriveAgent outperforms baseline prompting approaches in both accuracy and stability. Furthermore, our three-tier dataset, self-reasoning benchmarks, and VLM tuning pipeline contribute broadly to the field. Overall, DriveAgent demonstrates a promising path toward generalizable and reflective sensor-aware autonomy.

## References


[1] J. Deichmann, E. Ebel, K. Heineke, R. Heuss, M. Kellner, and F. Steiner, "Autonomous driving's future: convenient and connected," *McKinsey & Company*, 2023. [Online]. Available: https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/autonomous-drivings-future-convenient-and-connected

[2] J. Xie, Z. Chen, R. Zhang, X. Wan, and G. Li, "Large multimodal agents: A survey," 2024. [Online]. Available: https://arxiv.org/abs/2402.15116

[3] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[4] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.

[5] J. Sun *et al.*, "A survey of reasoning with foundation models," *arXiv preprint arXiv:2401.12345*, 2024. [Online]. Available: https://arxiv.org/abs/2401.12345

[6] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[7] H. Shao, Y. Hu, L. Wang, S. L. Waslander, Y. Liu, and H. Li, "Lmdrive: Closed-loop end-to-end driving with large language models," 2023. [Online]. Available: https://arxiv.org/abs/2312.07488

[8] C. Sima *et al.*, "Drivelm: Driving with graph visual question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[9] Y.-C. Chiu *et al.*, "V2v-llm: Vehicle-to-vehicle cooperative autonomous driving with multi-modal large language models," *arXiv preprint arXiv:2502.12345*, 2025.

[10] Z. Lan, H. Li, L. Liu, B. Fan, Y. Lv, Y. Ren, and Z. Cui, "Genfollower: Enhancing car-following prediction with large language models," *arXiv preprint arXiv:2407.05611*, 2024. [Online]. Available: https://arxiv.org/abs/2407.05611

[11] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou, "Chain of thought prompting elicits reasoning in large language models," *arXiv preprint arXiv:2201.11903*, 2022.

[12] Z. Xu, Y. Zhang, E. Xie, Z. Zhao, Y. Gou, Y. Kwan, K. Wong, Z. Li, and H. Zhao, "Drivegpt4: Interpretable end-to-end autonomous driving via large language model. arxiv," *arXiv preprint arXiv:2310.01412*, 2023.

[13] S. Park, M. Lee, J. Kang, H. Choi, Y. Park, J. Cho, A. Lee, and D. Kim, "Vlaad: Vision and language assistant for autonomous driving," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 980–987.

[14] Y. Ma, Z. Song, Y. Zhuang, J. Hao, and I. King, "A survey on vision-language-action models for embodied ai," *arXiv preprint arXiv:2405.14093*, 2024.

[15] W. Wang, C. Duan, Z. Peng, Y. Liu, and B. Zhou, "Embodied scene understanding for vision language models via metavqa," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, June 2025, pp. 22 453–22 464.

[16] R. Zhang, J. Hou, F. Walter, S. Gu, J. Guan, F. Rohrbein, Y. Du, P. Cai, G. Chen, and A. Knoll, "Multi-agent reinforcement learning for autonomous driving: A survey," *arXiv preprint arXiv:2408.09675*, 2024.

[17] Y. Wu, R. Jiao, T. Kim, Y. Jin, Y. Kwon, Q. A. Chen, C. Huang, and Q. Zhu, "Multi–agent autonomous driving systems with large language models: A survey of recent advances," *arXiv preprint arXiv:2502.16804*, 2025. [Online]. Available: https://arxiv.org/abs/2502.16804

[18] H. Gao *et al.*, "A survey for foundation models in autonomous driving," *arXiv preprint arXiv:2402.12345*, 2024. [Online]. Available: https://arxiv.org/abs/2402.12345

[19] Y. Wang, S. Xing, C. Can, R. Li, H. Hua, K. Tian, Z. Mo, X. Gao, K. Wu, S. Zhou, H. You, J. Peng, J. Zhang, Z. Wang, R. Song, M. Yan, W. Zimmer, X. Zhou, P. Li, Z. Lu, C.-J. Chen, Y. Huang, R. A. Rossi, L. Sun, H. Yu, Z. Fan, F. H. Yang, Y. Kang, R. Greer, C. Liu, E. H. Lee, X. Di, X. Ye, L. Ren, A. Knoll, X. Li, S. Ji, M. Tomizuka, M. Pavone, T. Yang, J. Du, M.-H. Yang, H. Wei, Z. Wang, Y. Zhou, J. Li, and Z. Tu, "Generative ai for autonomous driving: Frontiers and opportunities," *arXiv preprint arXiv:2505.08854*, 2025.

[20] B. Al-Tawil, T. Hempel, A. Abdelrahman, and A. Al-Hamadi, "A review of visual SLAM for robotics: Evolution, properties, and future applications," *Frontiers in Robotics and AI*, vol. 11, 2024.

[21] W. He, W. Chen, S. Tian, and L. Zhang, "Towards full autonomous driving: challenges and frontiers," *Frontiers in Physics*, vol. 12, 2024.

[22] W. Wang, H. Min, X. Wu, Y. Fang, G. Li, and X. Zhao, "A modular loop closure detection scheme for autonomous driving-a loosely coupled approach," *IEEE Transactions on Vehicular Technology*, 2024.

[23] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 697–12 705.

[24] M. AI, "Llama 3.2: Advancing vision ai on edge and mobile devices," https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/, 2024, accessed: 2025-04-12.

[25] OpenAI, "Gpt-4o technical report," 2024, https://openai.com/research/gpt-4o.

[26] P. AI, "Pixtra: A large vision-language model," 2024, https://pixtra.ai.

[27] Anthropic, "Claude 3.7 sonnet model card," 2024, https://www.anthropic.com/index/claude-3-7-sonnet.

[28] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[29] Y. Ma, C. Cui, X. Cao, W. Ye, P. Liu, J. Lu, A. Abdelraouf, R. Gupta, K. Han, A. Bera, J. M. Rehg, and Z. Wang, "Lampilot: An open benchmark dataset for autonomous driving with language model programs," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. [Online]. Available: https://aclanthology.org/2024.findings-emnlp.287

[30] K. Jiang, X. Cai, Z. Cui *et al.*, "Koma: Knowledge-driven multi-agent framework for autonomous driving with large language models," *arXiv preprint arXiv:2407.14239*, 2024.

[31] A. Madaan, S. Lin, X. Liu, D. Zhou, Q. V. Le, D. Schuurmans, E. H. Chi, and J. Wei, "Self-refine: Iterative refinement with self-feedback," *arXiv preprint arXiv:2303.17651*, 2023.

[32] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.