

# A Unified Perception-Language-Action Framework for Adaptive Autonomous Driving

Yi Zhang, Erik Leo Haß, Kuo-Yi Chao, Nenad Petrovic, Yinglei Song, Chengdong Wu and Alois Knoll

*Chair of Robotics, Artificial Intelligence and Embedded Systems*

*Technical University of Munich (TUM)*

Munich, Germany

{yi1228.zhang, erik-leo.hass, kuoyi.chao, nenad.petrovic, yinglei.song, chengdong.wu, k}@tum.de

**Abstract**—Autonomous driving systems face significant challenges in achieving human-like adaptability, robustness, and interpretability in complex, open-world environments. These challenges stem from fragmented architectures, limited generalization to novel scenarios, and insufficient semantic extraction from perception. To address these limitations, we propose a unified Perception-Language-Action (PLA) framework that integrates multi-sensor fusion (cameras, LiDAR, radar) with a large language model (LLM)-augmented Vision-Language-Action (VLA) architecture, specifically a GPT-4.1-powered reasoning core. This framework unifies low-level sensory processing with high-level contextual reasoning, tightly coupling perception with natural language-based semantic understanding and decision-making to enable context-aware, explainable, and safety-bounded autonomous driving. Evaluations on an urban intersection scenario with a construction zone demonstrate superior performance in trajectory tracking, speed prediction, and adaptive planning. The results highlight the potential of language-augmented cognitive frameworks for advancing the safety, interpretability, and scalability of autonomous driving systems.

**Index Terms**—Autonomous Driving, Multi-Sensor Fusion, Large Language Models (LLMs), Vision-Language-Action (VLA), Scene Understanding, Trajectory Planning

## I. INTRODUCTION

Autonomous driving systems are rapidly evolving, yet fundamental challenges persist in achieving human-like cognitive fidelity during the integration of perception, decision-making, and control under stochastic real-world conditions. Unlike humans, current systems struggle to seamlessly integrate sensory inputs with contextual cues, hindering their ability to manage uncertainties in complex, dynamic environments. This limitation restricts adaptability in partially observable settings, such as urban traffic, where comprehensive scene understanding is critical. These challenges highlight the need for integrated frameworks that emulate human cognitive adaptability, enabling robust, context-aware decision-making.

A promising approach to address these challenges is multi-sensor fusion, integrating LiDAR, cameras, and radar to enhance perceptual robustness. While these technologies improve raw data acquisition, a critical gap persists in bridging perceptual inputs with contextual reasoning and actionable control outputs. Vision-language models attempt to mitigate this by grounding visual data in textual instructions, yet they often lack the situational awareness and adaptability required for complex and long-tail driving scenarios.

Despite these technological advancements, autonomous driving systems continue to face limitations that hinder their ability to achieve human-like adaptability and safety in complex, dynamic environments. These primary limitations include:

- **Poor connectivity between functional modules** – Isolated perception, language, and planning subsystems hinder cohesive scene understanding, reducing contextual coherence and safety in dynamic driving scenarios like merging or vehicle-following.
- **Lack of structured semantic understanding in perception data** – Raw sensor data (e.g., LiDAR, images) lacks inherent meaning, requiring complex multi-stage processing (e.g., object detection, semantic segmentation, reasoning), which compromises reliability and interpretability in safety-critical decisions.
- **Limited generalization to unseen scenarios** – Rule-based or narrowly trained models struggle with unfamiliar scenarios, such as construction zones or erratic pedestrian behavior, lacking the robust, human-like reasoning needed for diverse conditions.

However, large language models (LLMs) demonstrate exceptional capabilities in reasoning, planning, and interactive understanding. So to address these challenges, we propose a unified framework that integrates multi-sensor fusion with an LLM-augmented vision-language-action (VLA) agent. This framework enables autonomous systems to combine low-level perception with high-level cognitive reasoning, achieving explainable, adaptive, and safety-bounded decision-making in open-world driving environments. Our key contributions are outlined below:

- **Integrated Cognitive Framework** – We introduce a Perception-Language-Action (PLA) framework that tightly couples multi-modal perception with LLM-based reasoning and motion planning, enabling coherent and adaptive decision-making in complex urban environments.
- **Multi-Sensor Semantic Fusion** – We develop a robust fusion module that combines LiDAR, radar, and camera data into structured scene descriptions, enhancing both spatial accuracy and semantic richness for downstream reasoning.

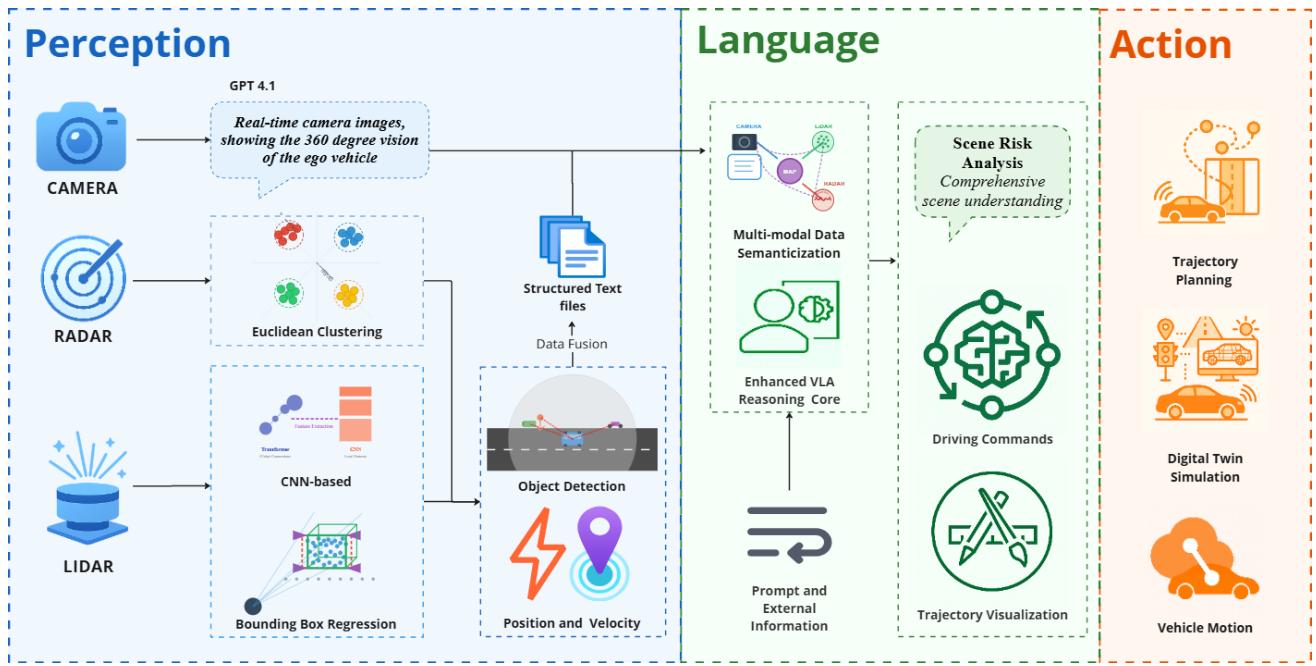


Fig. 1: Detailed workflow of the proposed framework for complex scene interpretation and motion control in autonomous driving: (1) Perception-layer, (2) Language-layer, and (3) Action-layer.

- **Enhanced Generalization through Contextual Reasoning** – By incorporating LLM-driven reasoning, our framework improves generalization to unseen scenarios, such as construction zones or unpredictable pedestrian behaviors, enabling robust decision-making in complex environments.
- **Empirical Validation in Complex Urban Scenarios:** We demonstrate the effectiveness and real-time adaptability of our framework through a case study on the nuScenes dataset, focusing on a challenging urban intersection with a construction zone, achieving low prediction errors and robust navigation performance.

The rest of the paper is structured as follows. Section II reviews related work, discussing advances in multi-sensor fusion, vision-language models, and large language models for autonomous driving. Section III introduces the Perception-Language-Action framework, including multi-sensor fusion and vision-language-action architecture. Section IV details the experimental setup and a case study on an urban intersection scenario. Section V presents the results, and Section VI concludes with findings and future work.

## II. RELATED WORK

The development of autonomous driving systems has spurred significant research across perception, decision-making, and control, with notable advances in multi-sensor fusion, vision-language integration, and reasoning-augmented architectures. This section surveys these efforts, emphasizing their strengths and limitations in addressing modular compartmentalization, data scarcity, and contextual rigidity.

### A. Multi-Sensor Fusion for Perception

Multi-sensor fusion has emerged as a cornerstone for robust perception in autonomous driving. Techniques integrating LiDAR, cameras, and radar enhance spatial and temporal scene understanding. For instance, *PointPainting* [1] fuses LiDAR with camera-derived semantic features for improved 3D object detection, while *TransFuser* [2] employs transformers to align multimodal features for trajectory prediction. Radar's resilience to adverse weather has been leveraged in fusion frameworks like *RadarNet* [3] to complement vision-based systems. However, these approaches often treat perception as an isolated module, leading to potential error propagation in dynamic scenarios, such as occlusions or sudden obstacles.

### B. Vision-Language Models for Scene Understanding

Vision-Language Models (VLMs) have gained traction for grounding visual perception in natural language, offering a bridge between sensory inputs and decision-making. Models like CLIP [4] and ViLBERT [5] enable semantic alignment between images and text. In autonomous driving, *DriveVLM* [6] explores VLMs to interpret driving scenes via textual prompts and vision inputs, while *VLMaps* [7] employs language-guided spatial reasoning for navigation. However, these models often struggle with real-time adaptability and fine-grained situational awareness, such as anticipating pedestrian intent or interpreting ambiguous traffic signals.

### C. Large Language Models in Decision-Making

Large Language Models (LLMs) like GPT-4 [8] and LLaMA [9] have demonstrated potential in reasoning and planning for autonomous systems. *GPT-Driver* [10] utilizes

in-context learning to generate explainable driving trajectories, while *DriveLLM* [11] aligns multimodal motion prediction with language prompts. Despite these advances, LLM-based approaches often operate in isolation from low-level perception and control loops, limiting their ability to adapt to real-time stochastic conditions such as urban intersections or adverse weather [12].

#### D. Integrated Architectures for Autonomous Driving

Efforts to unify perception, reasoning, and action have led to hybrid architectures that blend modular and end-to-end learning approaches. Transformer-based frameworks like *UniAD* [13] integrate perception, prediction, and planning into a unified pipeline, reducing error propagation while maintaining interpretability. *Reason2Drive* [14] and *RDA-Driver* [15] introduce reasoning-based decision-making but remain constrained by a lack of joint optimization with perception modules. Meanwhile, *Text2motion* [16] integrates neural motion planners with language interfaces, underscoring the growing need for explainability and adaptability in autonomous systems.

### III. METHODOLOGY AND WORKFLOW

In this section, we introduce a framework to tackle the challenges of poor functional module connectivity, limited generalization to unseen scenarios, and evaluation difficulties in autonomous driving systems. Our approach integrates multi-sensor fusion with a large language model-augmented Vision-Language-Action (VLA) architecture to enhance system performance and adaptability.

#### A. Overview of the Proposed Framework

The pipeline of our framework enables advanced scene interpretation and motion control for dynamic driving scenarios, as depicted in Figure 1. The pipeline comprises three primary layers:

- **Perception Layer:** Raw sensor data from cameras, radar, and LiDAR are processed for a cohesive environmental representation. 360-degree camera images are interpreted using advanced models like GPT-4.1 for interpretation. Concurrently, radar data undergoes Euclidean clustering for object delineation based on spatial proximity. LiDAR point clouds are processed via CNNs with bounding box regression for 3D object detection. A data fusion mechanism then integrates LiDAR and Radar outputs, producing structured text files that contain precise position and velocity information for detected objects. This fusion process leverages the complementary strengths of each modality to enhance the accuracy of object detection and multi-target tracking.

- **Language Layer:** The language layer processes structured text files and camera images, converting fused perception data into semantically rich representations. An enhanced Vision-Language-Action (VLA) Reasoning Core conducts comprehensive scene risk analysis and understanding, enabling context-aware decision-making.

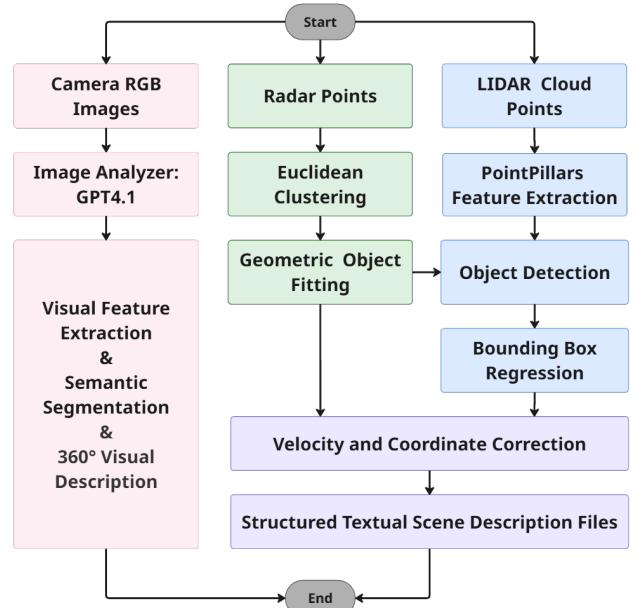


Fig. 2: Multimodal sensor fusion flow

Furthermore, the system integrates prompt-based inputs and external information (e.g., real-time traffic alerts, prior experience) to enrich the contextual understanding. Based on this comprehensive analysis, the language layer generates precise driving commands and visualizes the planned trajectories.

- **Action Layer:** The action layer receives driving commands and trajectory visualizations from the language layer. It handles detailed trajectory planning, converting high-level commands into precise, actionable vehicle paths. These paths are validated using high-fidelity digital twin simulations, which replicate real-world scenarios to ensure safety and efficiency. The layer's final output is direct control of vehicle motion, guiding the autonomous system effectively in dynamic driving environments.

#### B. Multi-Sensor Fusion Module

The multi-sensor fusion module integrates LiDAR, and radar data to create a structured representation of the ego vehicle's state and surrounding obstacles within 50 meters, enabling robust scene understanding, object tracking, and trajectory planning. Cameras provide visual data for feature extraction and semantic segmentation, LiDAR offers 3D point clouds for geometric understanding, and radar ensures reliable velocity estimation in adverse conditions. The process, shown in Figure 2, uses parallel pipelines to combine data for downstream tasks.

The camera pipeline processes RGB images using an advanced AI-based module (ChatGPT-4.1) for visual feature extraction, semantic segmentation, and generating a 360-degree visual description, identifying elements such as traffic signs, pedestrians, and lane boundaries. LiDAR point clouds are processed using the PointPillars [17] architecture for feature

TABLE I: An Example of Ego Vehicle State and Relative Obstacles in a structured textual scene description file

Ego Vehicle Information	
Label	ego_vehicle
Dimension (m)	(3.99, 2.06, 1.84)
Position (m)	(0, 0, 0)
Distance (m)	0
Velocity (m/s)	$v_x = 8.28, v_y \approx 0, v_z = 0$
Speed (m/s)	8.28
Obstacle Information	
Obstacle 1	
Label	human.pedestrian.adult
Partition	Front-right
Position (m)	(25.17, -21.64, 0.86)
Distance (m)	33.20
Velocity (m/s)	$v_x = 1.26, v_y = -0.06, v_z = -0.03$
Speed (m/s)	1.26
Obstacle 2	
Label	vehicle.truck
...	...

extraction, followed by object detection and bounding box regression to localize and estimate object dimensions in 3D space. Radar data, provided as point measurements, undergoes Euclidean clustering to form object-level groupings, followed by geometric object fitting. Radar and LiDAR pipelines are integrated in a velocity and coordinate correction block to ensure temporal and spatial consistency.

The fused scene information is processed into structured text files, detailing a scene relative to an ego vehicle, comprising two sections:

- **Ego Vehicle Information:** Attributes include the label (object type), dimensions (m for length, width, height), position ( $x, y, z$ ) in the ego-vehicle coordinate system (m), distance to origin (m), velocity ( $v_x, v_y, v_z$ ) (m/s), and scalar speed (m/s).
- **Obstacles:** A list of objects within a 50 m radius of the ego vehicle, each with attributes: label (obstacle type, e.g., pedestrian, vehicle, barrier, or unknown), partition (relative position, e.g., front, back, right), position ( $x, y, z$ ) (m), distance (m), velocity ( $v_x, v_y, v_z$ ) (m/s), and scalar speed (m/s).

Table I illustrates examples of these outputs, detailing the ego vehicle’s state and the attributes of surrounding obstacles.

### C. Augmented Vision-Language-Action (VLA) Architecture

We propose an integrated Augmented Vision-Language-Action (VLA) architecture that unifies perception, scene analysis, and action planning for autonomous navigation. Leveraging large language models (LLMs) as core reasoning engines, our approach enables intuitive interpretation of complex driving scenarios, thereby enhancing explainability and robust handling of ambiguous cases that challenge conventional modular systems.

As illustrated in Fig. 3, the architecture accepts multimodal sensor streams—camera images, LiDAR point clouds, and

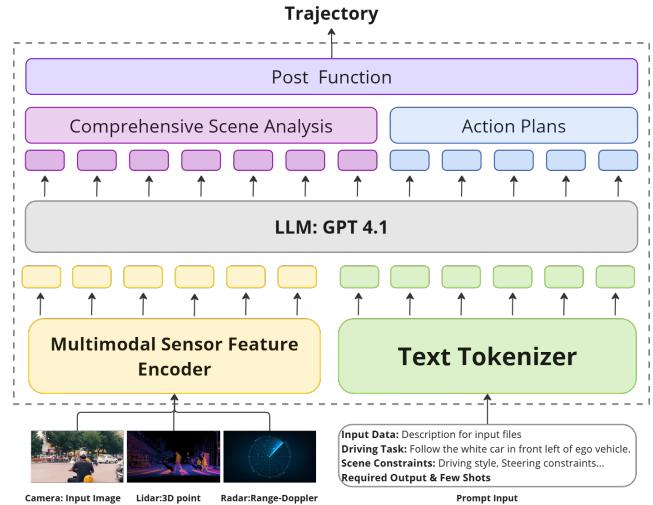


Fig. 3: The architecture of Augmented Vision-Language-Action model.

radar data—which are processed by a Multimodal Sensor Feature Encoder. This encoder extracts high-level representations from heterogeneous sensor modalities. Simultaneously, natural language inputs, including environmental conditions, driving tasks, and scene constraints, are processed via a dedicated Text Tokenizer.

Both feature streams are integrated and fed into an LLM (GPT 4.1), which serves as the central reasoning component. The LLM generates two main outputs: Comprehensive Scene Analysis, providing environmental interpretation, and Action Plans, defining navigation strategy. These outputs are further post-processed to produce the final vehicle trajectory.

This architecture unifies perception, language-based reasoning, and action planning within a single framework, thereby advancing the capabilities of autonomous systems toward more interpretable and adaptable behavior.

### IV. CASE STUDY: URBAN INTERSECTION FOLLOWING WITH CONSTRUCTION ZONE

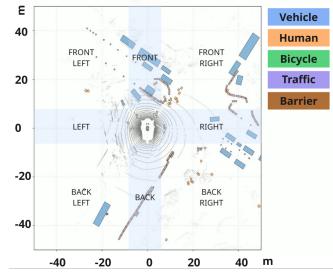
In this section, we present the experimental evaluation of our multi-layered framework for autonomous driving in complex environments. Although various scenarios were tested, this paper focuses on a representative case study: a following task at an urban intersection with an active construction zone. This challenging scenario enables a targeted assessment of the Perception-Language-Action (PLA) architecture’s capabilities in advanced scene understanding and robust decision-making. While the results demonstrate the effectiveness of our framework, further validation across more diverse scenarios will be conducted in future work.

#### A. Scenario Description

The selected case study focuses on a following task at an urban intersection with an active construction zone. Here, the ego vehicle navigates a signalized intersection where barriers and warning signs cause partial lane occlusion, temporary lane



(a) Front and rear camera views of the intersection with construction zones.



(b) Bird's-eye view visualization of the intersection with construction zones.

Fig. 4: Illustration of the selected urban intersection scenario.

TABLE II: Key Performance Metrics

Category	Metric	Value
Speed Prediction	MAE (m/s)	0.39
	R <sup>2</sup> Score	0.923
Steering Angle Prediction	MAE (°)	2.52
	R <sup>2</sup> Score	0.537
Trajectory	ADE (m)	1.013
	FDE (m)	2.026

shifts, and disrupted traffic flow. The environment is further complicated by irregular vehicle movements, dynamic obstacles (e.g., workers and equipment), unpredictable surrounding behavior, rapidly changing road conditions, and occlusions from construction machinery. This scenario was selected for its high unpredictability and complexity, posing significant challenges to perception, planning, and decision-making. It serves as a rigorous testbed to evaluate the robustness and adaptability of our framework under realistic, non-ideal conditions.

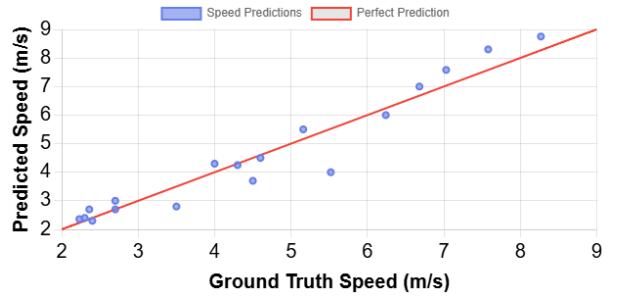
Figure 4 illustrates the selected urban intersection scenario with an active construction zone. Specifically, Figure 4a shows the front and rear camera perspectives, and Figure 4b provides a bird's-eye view visualization based on sensor fusion data.

### B. Experimental Setup

1) *Dataset:* We used the nuScenes dataset as the primary data source for our experiments. nuScenes is a large-scale, publicly available autonomous driving dataset that provides synchronized data from multiple sensors, including cameras, LiDAR, radar, and GPS/IMU, collected in urban environments.

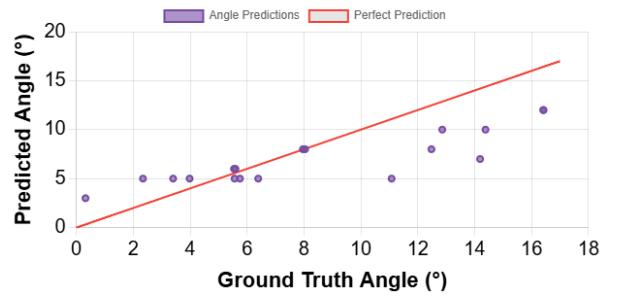
2) *Metrics:* We evaluate system performance using standard metrics for autonomous driving: mean absolute error (MAE)[18] and coefficient of determination ( $R^2$  score)[19] for speed and steering angle prediction, as well as average displacement error (ADE) and final displacement error (FDE) for trajectory accuracy [20], [21]. These metrics respectively capture absolute prediction error, regression fit, average trajectory deviation, and endpoint accuracy. They provide a comprehensive assessment of the system's ability to estimate motion

Speed: Ground Truth vs Predicted



(a) Speed: ground truth vs predicted. The red line indicates perfect prediction.

Steering Angle: Ground Truth vs Predicted



(b) Steering angle: ground truth vs predicted. The red line indicates perfect prediction.

Fig. 5: Comparison of model predictions with ground truth values for speed and steering angle.

parameters and maintain accurate trajectories in complex urban scenarios.

3) *Procedure:* For the scenes in the dataset, we sample various frames to obtain the original data. The proposed PLA framework is then applied to each sampled frame to predict the future trajectory for the next one second. And to emulate real-world perception-to-decision workflows, we design a structured prompt that guides GPT-4.1 to analyze traffic scenes and generate driving strategies. The input includes six surrounding camera views, a front-facing image for trajectory overlay, and a structured file with ego status and obstacle information. The prompt specifies the driving task, lane information, safe lateral deviation in a lane( $\pm 1.0$  m), and typical steering rate (5–15°/s). The model is instructed to assess risks and output:

- **Driving Commands:** speed action (accelerate/decelerate/maintain), steering direction (left/right) and angle for the next second;
- **Explanation:** reasoning based on perception and motion data.

### C. Results

The qualitative and quantitative results are summarized in Table II and Figures 5a, and 5b. For speed prediction, the model achieves a low mean absolute error (MAE) of 0.39 m/s and a high  $R^2$  score of 0.923, indicating accurate and reliable

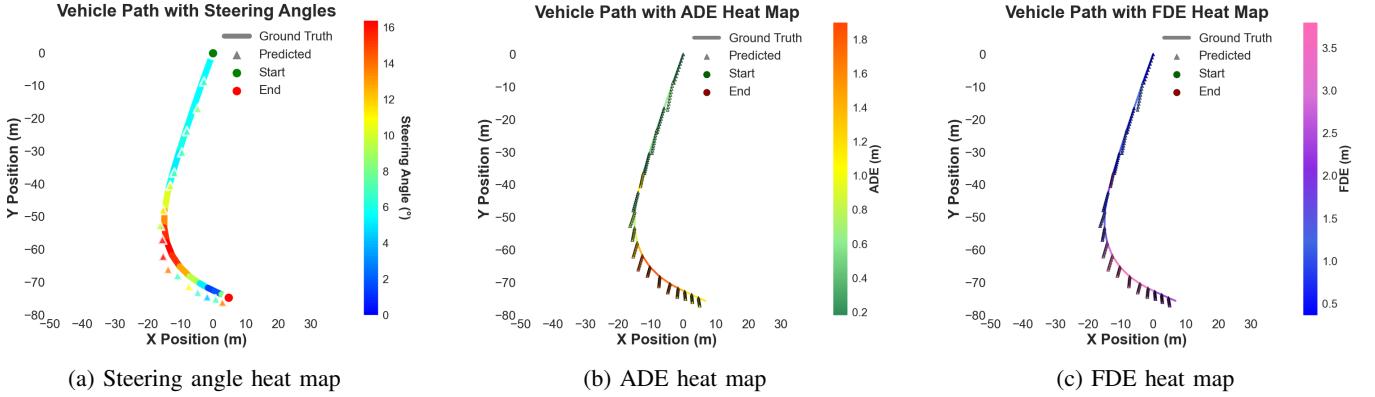


Fig. 6: Trajectory-based heat map visualizations for the same following task: (a) steering angle, (b) average displacement error (ADE), and (c) final displacement error (FDE). The color bars indicate the corresponding error or angle magnitude along the predicted path.

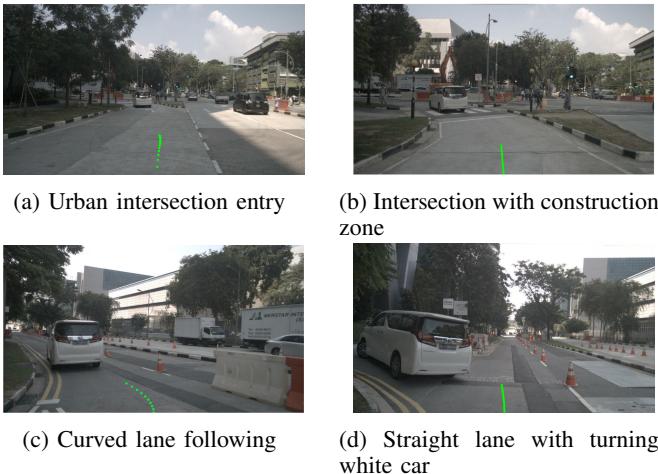


Fig. 7: Results of trajectory prediction by the PLA framework for the "following the front white car" task across diverse situations. Predicted trajectories are shown in green.

performance across the test scenes. Most predicted speed values are closely aligned with the ground truth, as shown by the clustering along the diagonal in Figure 5a.

Steering angle prediction, however, is more challenging, with an MAE of  $2.52^\circ$  and an  $R^2$  score of 0.537. Figure 5b shows a wider spread of points at higher steering angles, where the model tends to slightly underestimate the actual values. These results indicate that while the model captures general trends, there remains room for improvement in precise steering control.

For trajectory evaluation, the average displacement error (ADE) is 1.013 m and the final displacement error (FDE) is 2.026 m, demonstrating robust trajectory tracking overall, but highlighting the need for further improvement in steering accuracy. In summary, the results validate the effectiveness of the proposed model in speed and trajectory prediction, while identifying steering angle prediction under complex conditions as an area for future work.

#### D. Discussion

We present heat map visualizations in Fig. 6 to assess prediction performance. The steering angle heat map (Fig. 6a) shows predicted steering commands aligned with ground truth trajectories. The ADE (Fig. 6b) and FDE (Fig. 6c) heat maps illustrate spatial distributions of average and final displacement errors along predicted paths, respectively. These demonstrate low errors across most trajectory segments, validating robustness and accuracy. However, conservative steering angles during curves result in larger turning radii, causing slight deviations from inner lane alignment, though still safe for driving.

To further demonstrate the effectiveness of our PLA framework in real-world scenarios, Fig. 7 presents four representative frames from the "following the front white car" task across diverse urban settings. Predicted trajectories are shown in green. The selected scenes include: entering an urban intersection (Fig. 7a), navigating through a construction zone (Fig. 7b), curved lane following (Fig. 7c), and approaching a turning vehicle in a straight lane (Fig. 7d). In all cases, predictions closely follow the lead vehicle's path, demonstrating that the PLA framework effectively adapts to complex layouts and dynamic conditions. These qualitative results reinforce its robustness and practical applicability in autonomous driving.

#### V. CONCLUSION AND FUTURE WORK

This paper presents a Perception, Language, and Action (PLA) framework for autonomous driving, integrating multi-sensor fusion (radar, LiDAR, camera) with a GPT-augmented VLA agent for explainable, adaptive, and safety-bounded decision-making in complex urban environments, such as construction zones. The PLA system unifies perception, language, and action planning, achieving robust speed and trajectory prediction.

Future work will focus on enhancing steering control precision, optimizing real-time performance, and expanding validation to diverse scenarios, including rare edge cases, to advance trustworthy and human-aligned autonomous systems.

To bridge the gap between simulation and real-world testing, we plan to integrate our framework with AutoFrame [22], a framework enabling low-cost, hardware-in-the-loop evaluation in real vehicles, which will refine steering accuracy and system reliability. Additionally, we aim to leverage LLM-based toolchains, such as those in [23], to generate dynamic driving scenarios from freeform textual requirements. This approach will enable tailored testing beyond standard datasets, offering fine-grained control over environmental conditions, vehicle configurations, and system constraints, thus improving generalization and robustness.

#### ACKNOWLEDGMENT

This research was funded by the Federal Ministry of Research, Technology and Space of Germany as part of the CeCaS project, FKZ: 16ME0800K.

#### REFERENCES

- [1] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, “Point-Painting: Sequential Fusion for 3D Object Detection,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [2] A. Prakash, K. Chitta, and A. Geiger, “Multi-modal fusion transformer for end-to-end autonomous driving,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 7077–7087.
- [3] B. Yang, R. Guo, M. Liang, S. Casas, and R. Urtasun, “Radarnet: Exploiting radar for robust perception of dynamic objects,” in *European conference on computer vision*, Springer, 2020, pp. 496–512.
- [4] A. Radford, J. W. Kim, C. Hallacy, *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*, PMLR, 2021, pp. 8748–8763.
- [5] J. Lu, D. Batra, D. Parikh, and S. Lee, “Vilbert: Pre-training task-agnostic visiolinguistic representations for vision-and-language tasks,” *Advances in neural information processing systems*, vol. 32, 2019.
- [6] X. Tian, J. Gu, B. Li, *et al.*, “Drivevlm: The convergence of autonomous driving and large vision-language models,” *arXiv preprint arXiv:2402.12289*, 2024.
- [7] C. Huang, O. Mees, A. Zeng, and W. Burgard, “Visual language maps for robot navigation,” in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2023, pp. 10 608–10 615.
- [8] J. Achiam, S. Adler, S. Agarwal, *et al.*, “Gpt-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.
- [9] H. Touvron, T. Lavig, G. Izacard, *et al.*, “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.
- [10] J. Mao, Y. Qian, J. Ye, H. Zhao, and Y. Wang, “Gpt-driver: Learning to drive with gpt,” *arXiv preprint arXiv:2310.01415*, 2023.
- [11] Y. Cui, S. Huang, J. Zhong, *et al.*, “DriveLLM: Charting the Path Toward Full Autonomous Driving With Large Language Models,” *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 1450–1464, 2024. DOI: [10.1109/TIV.2023.3327715](https://doi.org/10.1109/TIV.2023.3327715).
- [12] Z. Yang, X. Jia, H. Li, and J. Yan, “Llm4drive: A survey of large language models for autonomous driving,” *arXiv preprint arXiv:2311.01043*, 2023.
- [13] Y. Hu, J. Yang, L. Chen, *et al.*, “Planning-oriented autonomous driving,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 17 853–17 862.
- [14] M. Nie, R. Peng, C. Wang, *et al.*, “Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving,” in *European Conference on Computer Vision*, Springer, 2024, pp. 292–308.
- [15] Z. Huang, T. Tang, S. Chen, *et al.*, “Making large language models better planners with reasoning-decision alignment,” in *European Conference on Computer Vision*, Springer, 2024, pp. 73–90.
- [16] K. Lin, C. Agia, T. Migimatsu, M. Pavone, and J. Bohg, “Text2motion: From natural language instructions to feasible plans,” *Autonomous Robots*, vol. 47, no. 8, pp. 1345–1365, 2023.
- [17] 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.
- [18] C. J. Willmott and K. Matsura, “Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance,” *Climate research*, vol. 30, no. 1, pp. 79–82, 2005.
- [19] N. J. Nagelkerke *et al.*, “A note on a general definition of the coefficient of determination,” *biometrika*, vol. 78, no. 3, pp. 691–692, 1991.
- [20] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, “Social lstm: Human trajectory prediction in crowded spaces,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 961–971.
- [21] H. Caesar, V. Bankiti, A. H. Lang, *et al.*, “nuscenes: A multimodal dataset for autonomous driving,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.
- [22] S. Kirchner, N. Purschke, C. Wu, M. A. Khan, D. Dixit, and A. C. Knoll, “AUTOFRADE-A Software-Driven Integration Framework for Automotive Systems,” in *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*, IEEE, 2024, pp. 521–528.
- [23] N. Petrovic, F. Pan, V. Zolfaghari, K. Lebioda, A. Schamschurko, and A. Knoll, “GenAI for Automotive Software Development: From Requirements to Wheels,” *arXiv preprint arXiv:2507.18223*, 2025.