Natural Language–Driven Navigation for Autonomous Rideshare Vehicles

Po-Hsun Chang, Siddhant Tandon and Saumit Vedula

***Abstract*— Existing autonomous ride-share systems typically rely on an explicit navigation destination, restricting their flexibility and intuitiveness. In contrast, natural language offers a richer means for specifying nuanced, context-dependent navigation goals such as "drop me near the coffee shop entrance" or "stop next to the red letter box".**

This project develops a natural language interface for final-mile navigation in autonomous ride-share scenarios. Our system integrates vision–language models with the CARLA [1] autonomous driving simulator to convert passenger text prompts and live camera imagery into actionable navigation way-points. The resulting interface enables context-aware, fine-grained control of drop-off locations. We analyze performance across navigation tasks with the Qwen 245b paramter VLM. Results demonstrate the systems' ability to navigate to simple targets within close view of the vehicle, but their inability to perform significant inference and reasoning when encountered with more complicated, less explicit directions.

I. INTRODUCTION

Autonomous vehicles (AVs) are increasingly expected to support human-centered modes of interaction, especially in shared mobility and ridesharing services. Shared autonomous mobility has the potential to transform transportation systems by reducing traffic congestion, lowering parking demand, and improving road safety [2], [3], [4].

As AVs transition into real-world deployment, the interface between passengers and systems becomes increasingly important. Current AV platforms rely on rigid input mechanisms for destinations, such as predefined addresses or preselected map locations. All of these methods are inferior in situations where passengers need to provide fine-grained control, such as approaching a particular entrance to a building, stopping beside an object, or adjusting the drop-off point based on the environment and scenarios.

Natural language is an intuitive modality for expressing specific navigation goals of a passenger, which is conventionally used in human-driven situations, like taxi services. Integrating natural language understanding along with perception and control, however, in the AV context, specifically for ride-share services, is a largely unexplored area of research. A language command must be grounded in the current visual scene, translated into an executable action, and carried out accurately and safely by the vehicle. Achieving this requires vision–language grounding and seamless coordination with navigation and action systems.

This project investigates a natural language navigation interface for autonomous ride-sharing vehicles using the CARLA simulator [1]. The proposed system converts passenger prompts and observed environment images into 3D waypoints that the vehicle can act upon. We evaluate performance using the Qwen VLM and test the system on tasks requiring visual grounding and spatial reasoning. Our goal is to assess the feasibility and limitations of utilizing VLMs for simple AV navigation.

II. RELATED WORK

Research regarding language-guided autonomous navigation can be mainly split into three sections, including (1) language grounding for vehicle actions, (2) vision-language models for driving control, (3) autonomous driving datasets with natural language supervision.

A. Language Grounding for Vehicle Actions

Earlier studies explored linguistic grounding for vehicle guidance. Roh et al. [5] proposed a supervised-learning pipeline that maps natural language instructions to driving actions by manipulating throttle control and steering angle. Kim et al. [6] investigated how human-provided advice can help the model attend to features in its environment, and modify control. These works show the usefulness of having language guidance as a complementary modality, but they rely on traditional visual encoders and a pre-defined grounding mechanism.

B. Vision-Language Models for Driving Control

As research on vision-language models (VLMs) has increased, it has demonstrated that VLMs can meaningfully influence autonomous driving behaviors. CarLLaVA [7] introduces a camera-only, closed-loop driving system whcih does not require complex or expensive labelling, in which the VLM outputs both path predictions and waypoints. SimLingo [8] further aligns vision and language with the action space. In the current state-of-the-art, however, VLMs have difficulty aligning high-level reasoning with precise control.

C. Datasets for Language-Guided Driving

To support training of models with direct links between language instructions and driving actions, some research has focused on developing datasets that pair natural language

*This work was not supported by any organization

Po-Hsun Chang is in Department of Electrical & Computer Engineering, University of Michigan, Ann Arbor, MI 48105, USA pohsun@umich.edu

Siddhant Tandon is in Robotics Department, University of Michigan, Ann Arbor, MI 48105, USA siddtan@umich.edu

Saumit Vedula is in Robotics Department, University of Michigan, Ann Arbor, MI 48105, USA savedula@umich.edu

Code available at: https://github.com/CharlesChang012/Natural-Language-Navigation-for-Rideshare-Vehicles.git

with driving scenes to support the training of grounded VLMs. The doScenes dataset [9] provides natural-language instructions accompanied by referential tags and multimodal sensor data, enabling detailed spatial grounding training. OmniDrive [10], in addition to its model baselines, proposes counterfactual multimodal annotations to help evaluate robustness in perception and language alignment.

III. Methodology

Our system enables an autonomous ridesharing vehicle to interpret natural language commands and translate them into actionable 3D waypoints within the CARLA autonomous driving simulator. The methodology consists of four key components: (1) environment observation and sensor configuration, (2) language-vision grounding using VLM, (3) 3D waypoint projection and navigation, and (4) task design and evaluation metrics. The flow chart is shown in figure 1.

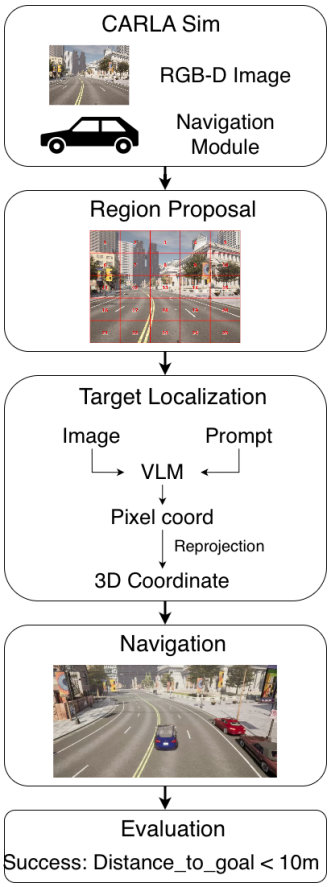

Fig. 1: Flowchart of system architecture

A. Sensor and Simulator Configuration

Experiments are conducted in CARLA using the default urban map. The simulated vehicle is equipped with the following components:

- Forward-facing RGB camera (resolution 1280×720)
- Depth camera sharing the same intrinsic parameters
- CARLA's navigation module

RGB images from camera provide vision information for the VLM. Depth information and camera intrinsics allow the recovery of 3D scene points from image coordinates. In order to focus on the significance of VLMs, we utilize CARLA's basic navigation agent to handle way-point following, traffic rules obedience, and collision avoidance.

B. Language–Vision Grounding

Passenger instructions are given in sentences such as "navigate to the bin". A 10×10 patch is overlaid on the RGB image as region proposal to help locate the target object. This path-based approach is utilized as early tests suggested the VLM struggled to extract the exact pixel coordinates of the object of interest in the image, with it often picking entirely the wrong area of the image as the target of interest. The system uses the pixel coordinates at the center of each patch as the object coordinates. It is assumed that for the largely , creating a direct mapping between a patch index and its location in the image.

The resulting gridded image and text prompt are provided to a vision–language model. The prompt instructs the model to identify which patch contains the referenced object and to return only the cell index. The model output is parsed to obtain a single identifier. This index is translated into pixel coordinates using the stored dictionary.

The prompt being sent to the VLM is as follows:

```
You are a navigation assistant. You are
     looking at a driver's view with a RED
     GRID overlay.
Each grid cell has a WHITE NUMBER.

Your Task:
Identify the grid cell number that contains
     the: "{target}"

Rules:
1. Select the cell containing the center of
     the object.
2. If the object spans multiple cells, pick
     the one closest to the road surface.
3. If the object is not visible, reply with
     ID: 0.

Output Format:
Return ONLY the JSON: { "id": <number> }
```

C. 3D Waypoint Reprojection

Given a pixel location (u, v) on the RGB image, the depth value $D(u, v)$ and the intrinsic matrix K are used to project a 3D point in the camera frame:

$$P_c = D(u, v)K^{-1}[u, v, 1]^\top.$$

Using CARLA's transformation utilities, the point is then re-projected into world coordinate via the extrinsic matrix:

$$P_w = T_w^c \, P_c,$$

where T_w^c is obtained from the simulator's sensor characteristics. The 3D point is then snapped to the nearest legal/drivable location the AV is able to drive to, hence producing a waypoint that is consistent with the vehicle's environment.

D. Navigation Control

Navigation is handled using CARLA's built-in waypoint follower. The vehicle drives toward the projected 3D target, and the system features lane tracing, road rules-following and collision avoidance.

E. Task Design and Evaluation

We evaluate our system on direct grounding tasks in which the target object is visible from the initial view (e.g., "go to the trash can", "stop near the mailbox"). Scenarios vary in target location. Performance is quantified using the navigation Success Rate, defined as the fraction of trials in which the vehicle reaches within 10 m of the target.

IV. RESULTS

We evaluated our architecture in 3 distinct scenarios listed in Table III. For this experiment, literal target navigation tasks were used, with an explicit reference to the navigation objective in the prompt, as would be expected in an autonomous ride-share service context. This setup was designed to assess the model's capabilities in scene and language understanding.

Scenario	Target	Target Visible	Success
1	Go to the nearest car	Yes	✓
2	Go to the nearest street light	Yes	✓
3	Go to the nearest tree	Yes	✓
4	Go to the nearest bin	Yes	✓
5	Go to the bus stop	No	✗

TABLE I: Results for 5 scenarios at spawn point 1

Scenario	Target	Target Visible	Success
1	Go to the bin	Yes	✓
2	Go to nearest "no parking" sign	Yes	✓
3	Go to nearest street light	Yes	✗
4	Go to nearest traffic light	Yes	✓
5	Go to nearest bus stop	No	✗

TABLE II: Results for 5 scenarios at spawn point 2

Scenario	Target	Target Visible	Success
1	Stop by stairs on the right	Yes	✓
2	Stop by stairs on the left	Yes	✓
3	Stop behind motorcycle on right	No	✗
4	Stop near red car on left	No	✗
5	Stop by black car on left	Yes	✓

TABLE III: Results for 5 scenarios at spawn point 3

As an example in scenario 1 at spawn point 1, consider the instruction: *"go to the car"*. Figure 2 shows the third-person view from behind the ego vehicle, where the target car is clearly visible. Therefore, it is selected as the target object. Figure 3 presents the BEV (bird's-eye view) map of the vehicle in motion. The vehicle low level policy in CARLA plans its path, follows traffic regulations, and successfully reaches the target, showing the capability of how language can provide useful information for vehicle navigation combining with VLMs.

Fig. 2: Third person view of target car in CARLA

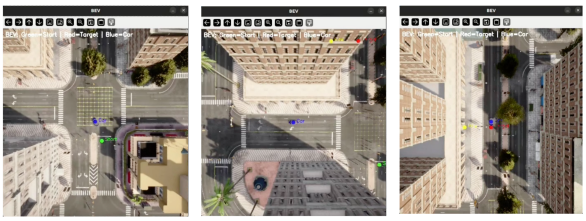

Fig. 3: BEV of vehicle motion to target car

On the other hand, in scenario 5 at spawn point 1 as shown in Figure 4, the target bus stop is not visible, making it impossible for the agent to localize its position. As a result, the agent is unable to select a correct waypoint or frontier to navigate toward. Instead, it falls back to choosing the waypoint that best matches the target "bus stop" based on the initial frame. Although the CARLA controller navigates the vehicle without collisions or traffic violations (Figure 5), this scenario highlights that the limited perception range of the system, and its lack of exploratory capabilities beyond the initial supplied image.

Fig. 4: Third person view of target bus stop in CARLA

Fig. 5: BEV of vehicle motion to target bus stop

The model also demonstrated the ability to perform spatial

reasoning. When given prompts such as "Stop by black car on left" or "Stop by stairs on the left" at spawn point 3, it was able to differentiate similar target objects based on their spatial position, and successfully navigate to them.

V. CONCLUSIONS

For autonomous rideshare services to become more convenient, user-friendly, and flexible, they must have the ability to parse and respond to natural-language human commands. In this project, we used the CARLA simulator and designed realistic user scenarios to evaluate how language can guide vehicle navigation through VLMs. Our results demonstrate that language grounding can enable effective navigation in many cases, especially when the navigation prompt is composed of explicit target objects and/or simple spatial relationships. The system has key limitations, however: it is currently only able to navigate to objects directly within its field of view. Future work could explore how it could respond to commands with multiple steps incumbent on their successful execution (such as driving around a corner to a drop of point). Furthermore, the prompts used in the project have limited complexity, and the system relies on explicit targets to navigate to. Though these would be the norm in most situations, more abstract prompts would further improve the flexibility of the system (e.g. drop em of somewhere out of the rain"). Testing yielded no success for such complicated prompts by the VLM, but it is possible higher resolution images with greater range would enable a VLM to perform such reasoning.

REFERENCES

[1] Dosovitskiy, Alexey; Ros, Germán; Codevilla, Felipe; López, Antonio; Koltun, Vladlen. "CARLA: An Open Urban Driving Simulator." In *Proceedings of the 1st Annual Conference on Robot Learning (CoRL)*, pp. 1–16 (2017).

[2] Hasan, Mohd Hafiz; Van Hentenryck, Pascal. "The Benefits of Autonomous Vehicles for Community-Based Trip Sharing." arXiv preprint arXiv:2008.12800 (2020).

[3] Kumakoshi, Yusuke; Hanabusa, Hisatomo; Oguchi, Takashi. "Impacts of Shared Autonomous Vehicles: Tradeoff Between Parking Demand Reduction and Congestion Increase." *Transportation Research Interdisciplinary Perspectives*, 12:100482 (2021).

[4] Ubbink, Angie; et al. "An Analysis of the Use of Autonomous Vehicles in the Shared Mobility Market: Opportunities and Challenges." *Sustainability*, 16(16):6795 (2021).

[5] Roh, J.; et al. "Conditional Driving from Natural Language Instructions." arXiv preprint arXiv:1910.07615 (2019).

[6] Kim, J.; et al. "Grounding Human-to-Vehicle Advice for Self-Driving Vehicles." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[7] Renz, Katrin; Chen, Long; Marcu, Ana-Maria; Hünermann, Jan; Hanotte, Benoit; Karnsund, Alice; Shotton, Jamie; Arani, Elahe; Sinavski, Oleg. "CarLLaVA: Vision-Language Models for Camera-Only Closed-Loop Driving." arXiv preprint arXiv:2406.10165 (2024).

[8] Renz, Katrin; Chen, Long; Arani, Elahe; Sinavski, Oleg. "SimLingo: Vision-Only Closed-Loop Autonomous Driving with Language-Action Alignment." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.

[9] Roy, Parthib; Perisetla, Srinivasa; Shriram, Shashank; Krishnaswamy, Harsha; Keskar, Aryan; Greer, Ross. "doScenes: An Autonomous Driving Dataset with Natural Language Instructions and Referential Tags." arXiv preprint arXiv:2412.05893 (2024).

[10] Wang, Shihao; Yu, Zhiding; Jiang, Xiaohui; Lan, Shiyi; Shi, Min; Chang, Nadine; Kautz, Jan; Li, Ying; Alvarez, Jose M. "OmniDrive: A Holistic Vision–Language Dataset for Autonomous Driving with Counterfactual Reasoning." arXiv preprint arXiv:2504.04348 (2025).

[11] Arai, Hidehisa; Miwa, Keita; Sasaki, Kento; Yamaguchi, Yu; Watanabe, Kohei; Aoki, Shunsuke; Yamamoto, Issei. "CoVLA: Comprehensive Vision–Language–Action Dataset for Autonomous Driving." arXiv preprint arXiv:2408.10845 (2024).