LLM-Enhanced Traffic Editor for Accelerated Testing of Autonomous Vehicles under Various Pedestrian Behaviors

Aizierjiang Aiersilan[✉][†], Mingzhe Liu[†][‡]
[†]Faculty of Science and Technology, University of Macau, Macau SAR, China 999078
[‡]Dept. of Mechanical Engineering, University of California, Berkeley, CA USA 94720-1740

ABSTRACT

Achieving diverse pedestrian behaviors in traffic simulation is crucial for enhancing the realism and complexity necessary for autonomous vehicle (AV) testing. Human drivers adjust behavior based on context, balancing caution with errors that may contribute to accidents. Real-world scenarios — such as congestion, overtaking, lane changes, and interactions between human-driven and AVs — often create adversarial conditions that challenge autonomous driving (AD), especially in heterogeneous and densely populated environments. While recent studies have advanced AV capabilities in perception, risk assessment, adaptive control, and decision-making, a significant gap remains in the realistic and context-sensitive simulation of pedestrian behaviors, particularly in pedestrian-dense settings. To address this, we propose a large language model (LLM)-enhanced traffic editing strategy that dynamically configures diverse pedestrian behaviors in AD simulations. By harnessing the generative and interpretive capacities of LLMs, our approach facilitates real-time, prompt-based customization of pedestrian scenarios, thereby enhancing the adaptability and complexity of simulations. Broadly compatible with existing simulators, this strategy elevates the rigor of AV testing. Experimental results in CARLA and Udacity simulators indicate that AVs tested in LLM-enhanced environments demonstrate higher collision rates under autopilot compared to default scenarios, highlighting the increased realism and adversarial conditions introduced. Once implemented, the approach manages pedestrian behaviors through the LLM, substantially improving the efficiency of handling complex traffic patterns and fostering robust AD testing. The primary source code of this project is publicly available at https://aizierjiang.github.io/LLM4TrafficEditor .

Keywords: traffic, autonomous driving, pedestrian behavior, simulation, large language model

1. INTRODUCTION

Autonomous driving (AD) is increasingly dependent on robust and realistic simulations to evaluate the safety and reliability of autonomous vehicles (AVs). These simulations must replicate complex, multi-agent traffic environments, capturing not only vehicle behaviors but also the unpredictable dynamics introduced by pedestrian interactions. In dense, heterogeneous environments, pedestrian behaviors vary due to intent, context, and disturbances. However, simulations often fail to capture these dynamic, context-sensitive and disturbing behaviors, creating a critical gap in AD testing.

The core challenge lies in modeling diverse pedestrian behaviors that closely resemble real-world actions, including spontaneous crossings, erratic movements, and context-driven adjustments in response to surrounding traffic conditions. From an optimization perspective, the problem can be formulated as a function $S{:}\{P,E\} \rightarrow T$, where P represents the set of possible pedestrian behaviors, E represents the set of environmental and traffic conditions, and T denotes the resulting set of traffic scenarios. This function S must maximize realism and complexity while ensuring the scenarios remain computationally tractable for simulation. The objective function can thus be framed to optimize scenario diversity D and complexity C, expressed as follows: $\boldsymbol{Maximize\, J = \alpha D + \beta C \quad subject\ to \quad C\,(P,E) \leq \epsilon}$, where J represents the scenario quality, α and β are weighting parameters, and ϵ constrains the feasibility of simulating high-dimensional pedestrian behaviors within reasonable performance limits. Here, D measures the variety of pedestrian interactions, and C quantifies complex behaviors in various traffic situations.

To bridge the gap, our study introduces an LLM-driven traffic editing strategy that allows for the generation and real-time customization of pedestrian behaviors within AV simulations. Leveraging the generative capabilities of LLMs, this strategy can dynamically adjust pedestrian interactions, creating a versatile and adaptive simulation environment.

[✉] Contact Aizierjiang Aiersilan via mc25101@um.edu.mo and Mingzhe Liu via liu_mingzhe@berkeley.edu.

Specifically, the LLM's interpretive function enables context-sensitive pedestrian configurations that respond to evolving traffic conditions, thus closely mirroring real-world unpredictability.

Experiments in CARLA and Udacity simulators show that our approach increases AV collision rates in pedestrian-dense environments, highlighting the realism and complexity of these simulations. This work offers a flexible strategy for AV testing, incorporating dynamic pedestrian behaviors to enhance traditional vehicle-centric simulations. By adding this layer of complexity, we aim to contribute to the advancement of AV safety and reliability in real-world applications.

2. RELATED WORKS

Recent advancements in AD research have underscored the significance of simulating complex pedestrian behaviors to improve AV safety in urban settings. With the diversity of pedestrian actions in crowded environments, realistic modeling is essential for meaningful AV testing[1][2][3]. Existing studies have explored key aspects, such as human-vehicle interaction[4][5][6], adaptive pedestrian modeling[7][8], and behavior prediction[9][10], contributing to a foundation that enhances the realness of AV simulations.

Research on traffic heterogeneity has emphasized how varied vehicle types, sizes, and capabilities impact traffic dynamics, which is crucial for realistic AV evaluation. Wang[10] et al. and Morales[11] et al. have shown that differences in vehicle types and road conditions can influence traffic flow and driver response times, respectively. These insights suggest that incorporating such diversity in simulation can help to represent the range of conditions AVs encounter in real-world scenarios. Efforts to introduce behavioral variety into traffic simulations have also progressed through traffic authoring frameworks. For instance, Li[12] et al. used Graph Neural Networks to model agent interactions, while "DeepMimic"[13] and other reinforcement learning-based approaches[14] enable the generation of diverse, realistic movements for agents in virtual environments.

In our work, we propose a strategy that leverages LLMs to enhance pedestrian behavior variability within simulators like CARLA[15] and Udacity[16]. Building on LLM capabilities explored by Dong[17] et al. and Zheng[18] et al., we employ LLM-generated pedestrian behaviors to introduce dynamic, unpredictable actions in simulated environments. This approach aims to bridge existing gaps by incorporating context-sensitive pedestrian behaviors, offering a new layer of realism for AV testing without claiming to fully address all challenges in AD simulation.

3. LLM-ENHANCED TRAFFIC EDITING

The proposed LLM-enhanced pedestrian behaviors simulation strategy for traffic simulation is structured in two primary phases (Fig. 1): the Traffic Editing Phase and the Pedestrian Control Phase. These phases collaboratively simulate diverse, dynamic scenarios essential for robust AV evaluation.

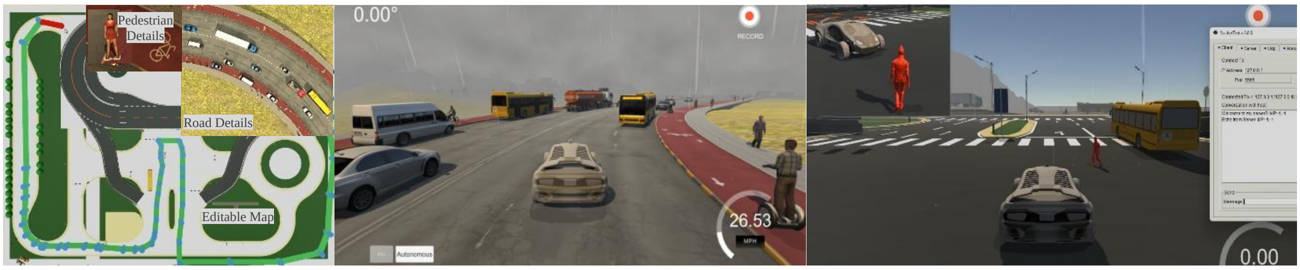

Figure 1. The screenshots of Traffic Editor. In panel (right), the green lines denote customizable traffic directions, with details included. In the traffic flow editing module, AV testers can define traffic flow and position pedestrians to create diverse interactions, simulating complex vehicle-pedestrian behaviors for comprehensive AV evaluation. The gray car in panels (middle) and (left) represents the AV, while the red figure in (left) represents a pedestrian controlled by the LLM.

3.1 Traffic Editing Phase

In the Traffic Editing Phase, the environment configuration and traffic elements, such as traffic lights, signs, and vehicle flow, can be managed both manually or by an LLM. This phase aims to maximize the heterogeneity of traffic flow, denoted H_f, which influences the level of challenge C posed to the AV. The complexity function for traffic in this phase, considering m vehicles and H_f^j as the heterogeneity of the j-th vehicle's flow, can be expressed as: $C_{traffic} = \sum_{j=1}^{m} H_f^j$, where

$C_{traffic}$ grows proportionally with the complexity of individual vehicle behaviors in the environment, adding realism to the AV testing scenarios.

3.2 Pedestrian Control Phase

In this phase, pedestrian behaviors are dynamically managed based on LLM-generated commands or by direct intervention from human players. The LLM first receives a description of the initial environment, and subsequently, it sends control commands to individual pedestrians, allowing them to interact dynamically with the AVs. The presence of diverse pedestrian actions introduces further layers of complexity and risk, essential for evaluating AVs performance in pedestrian-dense areas.

The behavior of each pedestrian i is modeled by H_p^i, representing individual heterogeneity within the pedestrian population. Let n denote the total number of pedestrians. The overall heterogeneity of pedestrian behaviors, $C_{pedestrian}$, is computed as: $C_{pedestrian} = \sum_{i=1}^{n} H_p^i$. The addition of these heterogeneous pedestrian behaviors contributes to the overall simulation complexity C, by introducing the potential for collisions or unexpected interactions with the AVs. Specifically, our approach leverages the LLM's command-driven control to simulate scenarios where the AVs may become distracted by pedestrian actions, potentially resulting in accidents — mirroring real-world adversarial conditions.

3.3 Combined Complexity

Our goal is to enhance scenario realism and complexity while maintaining computational feasibility. Aiming to guide the practical resolution of the optimization problem, we decompose C into distinct components based on traffic and pedestrian behavior complexity, as follows:

$$C = C_{traffic} + C_{pedestrian} + \lim_{l \to \infty} \frac{1}{l} + q \cdot \sum_{j=1}^{m} \left(H_p{}^{(j)} + H_f{}^{(j)} \right) \tag{1}$$

where: l denotes the network latency in socket communication within the simulator, assumed to remain within a controlled range for consistent performance in this study. q represents the functional capacity of the simulation, encompassing the breadth of AD-related features it supports.

Our strategy seeks to reach the complexity threshold $C\ (P, E) \leq\ \epsilon$ by iteratively generating and evaluating pedestrian behaviors across simulated scenarios, ensuring feasibility within computational constraints. By progressively increasing both scenario diversity D and complexity C within this threshold, we optimize each scenario's realism and diversity to create conditions that challenge the AVs.

As the complexity C increases, it proportionally raises the likelihood of generating "corner cases" — rare, high-stress scenarios that are vital for testing AV robustness. The probability $P\ (corner\ case \mid C)$ of encountering a corner case is given by the conditional probability: $\mathrm{P}\ (corner\ case \mid C) = \frac{\mathrm{P}(corner\ case \cap C)}{\mathrm{P}(C)}$.

3.4 Evaluation Criteria

The configurable traffic editor integrates modules for simulating diverse pedestrian behavior, controlling traffic flow, and facilitating communication with LLMs. It also includes an AD Module for activating autopilot vehicles and a Metahuman Module for human-like pedestrian behavior, all of which collaborate to generate dynamic traffic simulations. The proposed approach is evaluated through key metrics that assess the complexity of simulated environments and the collision rates of AVs under challenging conditions, aligning with the complexity measures and probability definitions outlined in previous sections.

Corner Case Probability: The probability of encountering a corner case quantifies the likelihood of rare, high-stress scenarios occurring within complex traffic configurations, directly assessing the AV's robustness in responding to unexpected situations, as defined in the previous section.

Collision Rate of the Ego-AV: The collision rate of the ego-autonomous vehicle (ego-AV) measures the ego-AV's safety performance by tracking the frequency of collisions involving the AV within the simulation environment. This rate is defined as follows: $CR_{ego} = \frac{\text{Total Collisions of egoAV}}{\text{Total Simulation Time}}$.

The configurable traffic editor, supported by the five core modules outlined above, enables testers to generate varied, high-complexity traffic scenarios that effectively facilitate the identification and evaluation of corner cases. By leveraging LLM-based agent behavior simulation for content generation and command execution, our approach enhances the realism and diversity of traffic scenes, allowing the simulation to closely reflect the dynamic nature of real-world environments. These evaluation metrics provide a quantitative assessment of the AV's ability to avoid collisions and navigate complex, high-variability conditions, capturing the ego-AV's decision-making effectiveness in interactions with diverse pedestrian and vehicle behaviors. Together, these criteria establish a comprehensive approach for assessing AV resilience and operational performance, ensuring that the AV can reliably navigate realistic, dynamic scenarios within a controlled simulation environment, advancing the overall robustness and adaptability of AV technology in real-world conditions.

4. EXPERIMENT

The experiment was structured to align with the modular and two-phase framework outlined in the methodology. We utilized three simulation platforms — CARLA (v0.9.13) and Udacity (v1.45) — to implement and evaluate the AV testing scenarios under varied environmental complexities and pedestrian behaviors. Each platform facilitated scenario creation, autonomous navigation, and real-time pedestrian control, providing a comprehensive testing ground for the AV's performance across controlled yet dynamically challenging setups. In CARLA and Udacity, the vehicle was set to autopilot mode, allowing the AV's navigation to proceed without manual intervention once connected with LLMs responsible for pedestrian behavior control. The LLMs selected for this experiment included ChatGPT 3.5-turbo, LLaMA2-7B, Qwen7B, LLaMA2-13B, and Qwen-14B. For all platforms, these LLMs delivered directional commands — "-2" (right), "-1" (forward), "0" (stay still), "1" (left), and "2" (backward) — along with an additional value specifying turning orientation. Each command initiated specific pedestrian responses in real-time, emulating the unpredictability of real-world pedestrian interactions. In the event of a collision, the pedestrian would modify their behavior by changing direction, jumping, or even running, adding another layer of complexity to the simulation.

4.1 Phase 1: Traffic and Population Control

In the initial phase, the traffic editing and population control modules were engaged to set diverse vehicle and pedestrian patterns across each simulation platform. Large crowds of pedestrians were introduced to simulate complex, realistic behaviors within intersections and along sidewalks. Vehicle density and flow were dynamically adjusted to create high variability, testing the AV's responsiveness to dynamic traffic patterns. This phase evaluated the AV's decision-making capability when presented with a high degree of pedestrian and vehicular activity in a controlled environment.

4.2 Phase 2: LLM-Controlled Pedestrian Interaction

In the second phase, LLM-controlled pedestrian behavior was integrated, introducing a more unpredictable element into the simulation. The configurable traffic editor allowed for seamless connection between the LLMs and the pedestrian agents, which adapted their responses according to real-time commands issued by the LLMs. This setup enabled the simulation of spontaneous and complex pedestrian movements, such as sudden directional changes or unexpected behaviors following a collision. The LLM-generated pedestrian actions created a heterogeneous behavior set, challenging the AV's collision-avoidance capabilities and highlighting its resilience in handling diverse real-world interactions.

4.3 Mixed Scenario Evaluation

Following the phase-specific evaluations, a mixed-scenario test was conducted, combining both high-density pedestrian environments and complex traffic flows with LLM-directed pedestrian control. This scenario presented the AVs with an intricate blend of traffic and population elements, aiming to assess their robustness in realistic, high-complexity conditions. The AV's performance was measured based on its ability to maintain lane adherence, avoid collisions, and respond effectively to unpredictable pedestrian behaviors. By combining modules in tandem, the simulation provided a rigorous test of the AV's decision-making and adaptability.

Such experimental setup, using LLM-based pedestrian control and configurable traffic conditions, enabled a thorough quantitative evaluation of the AV's collision-avoidance and decision-making in dynamic and diverse scenarios.

5. EVALUATION

The evaluation of the AV's performance under different LLMs and simulation platforms involved a detailed comparison of various metrics, including collision rates, corner case probabilities, and the variability of pedestrian behaviors. The following section presents a thorough analysis based on the experimental setup outlined in the previous section, with a particular focus on the differences between LLMs in generating pedestrian behaviors and the performance of different simulators in autonomously navigating complex traffic scenarios.

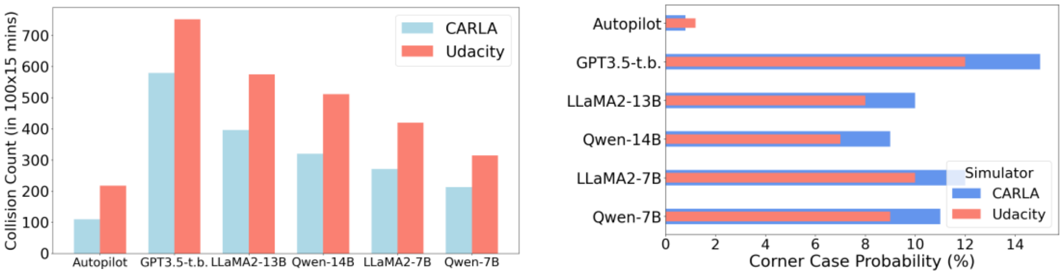

Figure 2. Comparison of collision rates (left) and corner case probabilities (right) under various pedestrian behaviors controlled by different LLMs in 2 different simulators. "GPT3.5t.b." represents "GPT 3.5-turbo" model.

5.1 Collision Rate Analysis

The integration of LLM-controlled pedestrian behaviors introduced substantial variability in simulated environments, impacting the ego-AV's ability to navigate and avoid collisions. Different LLMs exhibited distinct capabilities in generating pedestrian behaviors: ChatGPT 3.5-turbo[19] produced the most unpredictable actions, such as abrupt lane crossings and sudden road movements, intensifying the complexity for the ego-AV (Fig. 2 - left). In contrast, LLaMA2-13B[20] and Qwen-14B[21]generally followed commands with predictable paths, though they occasionally required the ego-AV to adjust in high-traffic scenarios, while LLaMA2-7B and Qwen-7B introduced sporadic, moderate unpredictability with unexpected pauses and directional shifts. This unpredictability resulted in a notable increase in collision rates compared to baseline autopilot scenarios, where the ego-AV could avoid most collisions through conventional traffic flow simulation. The heightened collision rate with LLM-driven pedestrians highlights the AV's challenges in adapting to dynamic, unscripted pedestrian behaviors, emphasizing the need for advanced behavior modeling to realistically test AV resilience in high-variability environment.

5.2 Corner Case Probability

Corner cases are rare, extreme situations that test the limits of an AV's decision-making and control capabilities, crucial for assessing its resilience in unpredictable environments. In our experiments, corner case probability was evaluated based on the ego-AV's ability to respond to significant deviations in pedestrian behavior or traffic flow. The introduction of LLM-controlled pedestrians, especially with models like ChatGPT 3.5-turbo, led to a slight increase in corner case probability, with more frequent scenarios requiring split-second decisions to avoid pedestrians and vehicles (Fig. 2 - right). This increase was most noticeable in high-density pedestrian areas, where erratic or adversarial behaviors, such as darting into traffic or changing paths unexpectedly, occurred. Such increase in corner case occurrences further underscores the need for advanced simulation environments that account for a wide range of unpredictable behaviors, ensuring that the AV is rigorously tested against rare and extreme situations that might occur in real-world traffic scenarios.

5.3 Simulator Comparison

The experiment also compared the autopilot capabilities of different simulators, CARLA and Udacity, which provided environments of varying complexity and fidelity for simulating real-world driving scenarios and LLM-controlled pedestrian behavior. CARLA, known for its advanced simulation of dynamic vehicle motions, showed an increase in collision rates and corner case probabilities when paired with LLM-controlled pedestrians, especially with larger LLMs. In contrast, Udacity, a simpler platform, showed lower collision rates and corner case probabilities, though still showed an increase compared to the baseline autopilot, likely due to its less complex environment and lower physical realism.

6. CONCLUSION

We introduced an LLM-enhanced traffic editing strategy to improve the simulation of pedestrian behaviors in AV testing. By leveraging the capabilities of LLMs, our approach enables dynamic and context-sensitive configuration of pedestrian actions, enhancing the realism and complexity of simulated environments. Experimental results from CARLA and Udacity simulators show that AVs tested in LLM-driven scenarios experience higher collision rates under autopilot, indicating the added challenge of more realistic, pedestrian-dense conditions. However, there are several limitations to this approach.

Constructing in-context learning examples that leads the LLM to an accurate command for controlling the Metahuman and filtering the effective LLM commands requires vast manual effort, sometimes. Additionally, occasional system crashes occur when controlled pedestrians become stuck in certain areas of the map, highlighting the need for further refinement in scenario management. Future work could focus on improving the robustness of pedestrian behavior simulations, enhancing the system's ability to handle complex interactions without failure. Expanding the scope of pedestrian behaviors and integrating more dynamic environmental factors could also further increase the relevance and utility of this approach for real-world AV testing.

REFERENCES

[1] Codevilla, Felipe, et al. "Exploring the limitations of behavior cloning for autonomous driving." Proceedings of the IEEE/CVF international conference on computer vision. 2019.

[2] Peintner, Jakob, et al. "Driving Behavior Analysis: A Human Factors Perspective on Automated Driving Styles." 2024 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2024.

[3] Shoman, Maged, et al. "Autonomous Vehicle–Pedestrian Interaction Modeling Platform: A Case Study in Four Major Cities." Journal of Transportation Engineering, Part A: Systems 150.9 (2024): 04024045.

[4] Rashid, Md Mobasshir, MohammadReza Seyedi, and Sungmoon Jung. "Simulation of pedestrian interaction with autonomous vehicles via social force model." Simulation modelling practice and theory 132 (2024): 102901.

[5] Rezwana, Saki, and Nicholas Lownes. "Interactions and Behaviors of Pedestrians with Autonomous Vehicles: A Synthesis." Future Transportation 4.3 (2024): 722-745.

[6] Dang, Meiting, et al. "Coupling intention and actions of vehicle–pedestrian interaction: A virtual reality experiment study." Accident Analysis & Prevention 203 (2024): 107639.

[7] Muktadir, Golam Md, and Jim Whitehead. "Adaptive pedestrian agent modeling for scenario-based testing of autonomous vehicles through behavior retargeting." IEEE Int. Conf. Robot. Automat.(ICRA). 2024.

[8] Wang, Jingbo, et al. "Learning human dynamics in autonomous driving scenarios." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023

[9] Zhang, Chi, and Christian Berger. "Pedestrian behavior prediction using deep learning methods for urban scenarios: A review." IEEE Transactions on Intelligent Transportation Systems 24.10 (2023): 10279-10301.

[10] Wang, Jinghua, et al. "Investigating heterogeneous car-following behaviors of different vehicle types, traffic densities and road types." Transportation research interdisciplinary perspectives 9 (2021): 100315.

[11] Morales-Alvarez, Walter, et al. "Vehicle automation field test: Impact on driver behavior and trust." 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2020.

[12] Li, Zirui, et al. "A hierarchical framework for interactive behaviour prediction of heterogeneous traffic participants based on graph neural network." IEEE Transactions on Intelligent Transportation Systems 23.7 (2021): 9102-9114.

[13] Peng, Xue Bin, et al. "Deepmimic: Example-guided deep reinforcement learning of physics-based character skills." ACM Transactions On Graphics (TOG) 37.4 (2018): 1-14.

[14] Mao, Wei, Miaomiao Liu, and Mathieu Salzmann. "Generating smooth pose sequences for diverse human motion prediction." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[15] Dosovitskiy, Alexey, et al. "CARLA: An open urban driving simulator." Conference on robot learning. PMLR, 2017.

[16] Smolyakov, M. V., et al. "Self-driving car steering angle prediction based on deep neural network an example of CarND udacity simulator." 2018 IEEE 12th international conference on application of information and communication technologies (AICT). IEEE, 2018.

[17] Xiaofei, Dong, et al. "A Survey of Research Progress and Theory Foundation in Large Model." 2022 IEEE 8th International Conference on Cloud Computing and Intelligent Systems (CCIS). IEEE, 2022.

[18] Zheng, Ou, et al. "ChatGPT is on the horizon: Could a large language model be all we need for Intelligent Transportation?." arXiv preprint arXiv:2303.05382 (2023).

[19] OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. https://openai.com/blog/chatgpt/.

[20] Touvron, Hugo, et al. "Llama 2: Open foundation and fine-tuned chat models." arXiv preprint arXiv:2307.09288 (2023).

[21] Bai, Jinze, et al. "Qwen technical report." arXiv preprint arXiv:2309.16609 (2023).