2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

PCASim: Promptable Closed-loop Adversarial Simulation for Urban Traffic Environment

Chuancheng Zhang[1†], Zhenhao Wang[2†], Kaizheng Li[1], Yaran Lin[1], Qiang Guo[3*], Bin Jiang[1*]

Abstract— **Real-world autonomous driving, particularly in urban environments with numerous corner cases, requires rigorous testing to ensure product safety and robustness. However, few studies have explored integrating adversarial scenario generation with the training of safety agents in closed-loop testing, enabling efficient co-evolution and mutual enhancement of both. To address this challenge, an adversarial behavior knowledge repository is constructed by applying rule-based filtering to an open-source dataset, combined with knowledge retrieval modules tailored for simulation environments. A large language model (LLM) is employed to integrate knowledge-, data-, and adversarial-driven approaches, generating safety-critical traffic scenarios customized to user needs. Additionally, while evaluating the generated scenarios, we employ reinforcement learning models to train the behaviors of different types of vehicles, thereby enriching scenario diversity beyond existing datasets while preserving realism. Experimental results demonstrate that the proposed framework improves the accuracy of domain-specific language generation by 12%. Moreover, the success rate of newly generated scenario transformations increases by 8%, while obstacle-avoidance capability is enhanced by 30%. For the complete manuscript, please refer to: https://zhenhaooo.github.io/PCASim.github.io/**

I. Introduction

Despite significant advancements in autonomous driving technology, high-level autonomous driving accidents, exemplified by Waymo [1] and Tesla [2], continue to occur frequently. This underscores the persistent challenges faced in achieving a fully reliable and safe system. In contrast to traditional vehicle safety assessments, which typically rely on controlled collision experiments and behavioral tests within predictable environments, autonomous driving systems (ADS) are designed to operate in far more open and uncontrolled settings. Therefore, prior to large-scale deployment, it is essential to consider a wide range of testing scenarios to include the inherent range of challenges [3].

The immense diversity of real-world traffic scenarios indicates that relying solely on physical testing is insufficient to address all potential risks [4]. To mitigate this issue, the traditional paradigm involves collecting traffic data from real-world scenes through onboard sensors, which naturally reflect the true distribution of driving data [5]. This data is then used to construct training datasets for developing the driving capabilities of ADS. However, this approach is inefficient and contains many redundant scenarios, which, if used directly for training, could hinder the model's ability to learn the safety-critical scenarios that are more likely to cause higher collision rates [6]. Consequently, increasing attention has been paid to generating adversarial scenarios, such as more hazardous scenarios of lane change [7] or car follow-up situations [8]. Furthermore, current research has demonstrated that training autonomous vehicles with generated adversarial scenarios can significantly reduce safety issues in ADS [9], [10].

Currently, safety-critical scenario generation methods, represented by data-, knowledge- and adversarial-driven approaches — excel in different stages such as scene understanding, rule creation and encoding. These methods often fail to integrate effectively and fully capitalize on their respective strengths. However, with the advent of LLMs trained on massive internet-scale data and billions of parameters, remarkable capabilities in commonsense reasoning, planning, interaction and decision-making have been demonstrated [11], highlighting their immense potential to address the aforementioned challenges. This paper proposes a novel closed-loop traffic simulation framework, aimed at bridging the gap between data-driven and knowledge-driven approaches, with a focus on the design of the data end and the enhancement of LLM's adversarial scenario generation capabilities. By requiring only natural language descriptions from the user, it utilizes specialized designs for LLM to generate a Domain-Specific Language (DSL) that guides the step-by-step execution of each dynamic object within the scene, thereby creating complex adversarial scenarios for the robust and safe training of autonomous vehicles. To fully utilize these scenarios, a reinforcement learning agent is trained to avoid adversarial behaviors embedded within the environment, resulting in complete scenarios encompassing both adversarial actions and successful avoidance maneuvers.

As shown in Fig. 1, we have developed a rule-based refinement method to extract potential hazardous scenarios from an open-source urban intersection dataset [12], which further contributes to the construction of a scenario corpus. This corpus consists of primary scenes centered on the ego vehicle and secondary scenes centered on the adversarial vehicle. Based on this corpus and leveraging current mainstream large language models [13]–[15], we employ Retrieval-Augmented Generation (RAG) to assist in retrieval, further enhanced by prompt engineering and reinforcement learning-based optimization to construct a comprehensive adversarial scenario repository. Moreover, a lightweight middleware translates the generated DSL into Python code, enabling integration with

†Both authors contributed equally to this research.
[1]Chuancheng Zhang, Kaizheng Li, Yaran Lin and Bin Jiang are with school of Airspace Science and Engineering, Shandong University, 264209 Weihai, China.
[2]Zhenhao Wang is with the School of Mathematics and Statistics, Shandong University, 264209 Weihai, China.
[3]Qiang Guo is with the School of Computer Science and Technology, Shandong University of Finance and Economics, 250014 Jinan, China.
*Corresponding author email: jiangbin@sdu.edu.cn

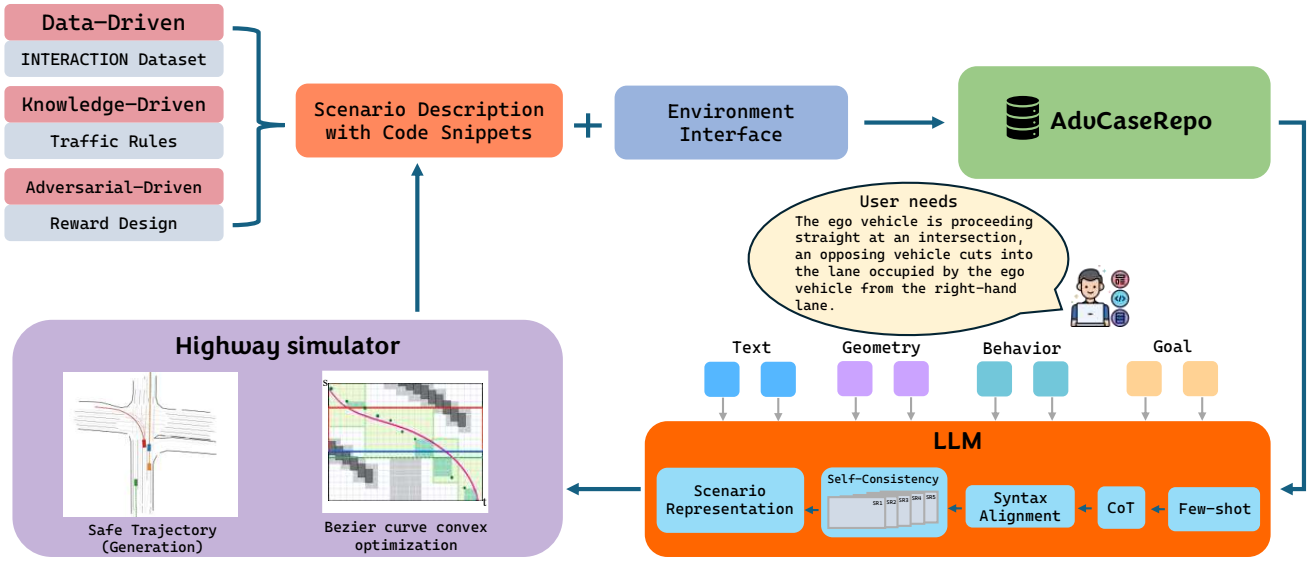


Fig. 1. The overall framework of the closed-loop adversarial simulation

the simulation environment. To the best of our knowledge, this represents the first notable attempt in closed-loop ADS testing to achieve on-demand scenario generation and leverage the extended scenarios for training autonomous agents. In summary, the contribution of this paper is three-fold:

1) An urban traffic adversarial scenario repository was constructed, and a RAG+LLM-based prompt engineering paradigm was developed for scenario generation. This framework enables efficient, diverse, and interpretable generation of realistic and strongly adversarial scenarios tailored to specified prompts.
2) An RL-based traffic-flow model is proposed to control the ego and adversarial vehicles, simultaneously validating the quality of the generated adversarial scenarios and producing new safety-critical scenario data.
3) A closed-loop scenario generation and evaluation framework is introduced to assess and filter high-quality scenarios, iteratively enriching the scenario repository and alleviating the limitations of small-sample distributions.

II. Related Works

A. Generation of Safety-Critical Scenarios

Safety-critical scenario generation for autonomous vehicles (AVs) typically falls into three categories: data-, adversarial-, and knowledge-based approaches. Data-driven approaches [16]–[18] advocates the utilization of real-world data to guide vehicle behavior. [19] proposed extracting static and dynamic variables from the CIMSS-TA dataset and employing conditional tabular generative adversarial networks (CTGANs) to increase diversity of generated scenarios. Although realistic, this approach is limited by long-tail sparsity of critical scenarios and costly data collection. Traditional adversarial-driven approaches often utilize Monte Carlo Tree Search [20], Bayesian optimization [21], and reinforcement learning within the framework of structured domain randomization (SDR) [22]. Grounded in uncertainty-aware, sampling-based policy optimization, these techniques balance exploration and exploitation in sparse-reward settings and generate realistic adversarial scenarios. However, these methods are often inefficient and generate adversarial environments with limited diversity. To address these limitations, A generative adversarial network (GAN)-based model combined with reinforcement learning, incorporating human driving priors to generate naturalistic adversarial scenarios, was proposed by [23]. [24] introduced kinematics gradients to generate robust, safety-critical scenarios for imitation learning. [25] introduced a closed-loop adversarial scenario generation framework based on LLM. This framework iteratively refines scenarios, ultimately producing high-quality adversarial environments with significantly enhanced generation efficiency. [26] proposed feasibility-guided generation, balancing adversariality and realism in safety-critical scenarios. Knowledge-based generation produces scenarios based on predefined traffic rules and physical constraints [27]–[29]. Previous studies rely on complex rule encoding, yet manually crafted rules could not cover all safety-critical cases. With the advancement of end-to-end approaches in autonomous driving, [30] introduced TARGET — an end-to-end architecture that automatically generates ADS test scenarios based on traffic rules, effectively addressing the complexity of rule-to-simulator mapping and minimizing manual intervention.

B. LLM for Autonomous Driving Simulation

LLM exhibits human-like understanding and reasoning capabilities in complex scenarios. For instance, [30] validated LLMs' capabilities in knowledge extraction, verification, and syntactic alignment, enabling them to comprehend contextual information and generate structured outputs, effectively simulating human reasoning in complex scene interpretation. [31] introduced a large multimodal model (LMM) capable of reasoning, abstraction and scenario construction from

traffic videos using a multimodal few-shot chain-of-thought (CoT) approach. Through leveraging these capabilities, LLMs are increasingly used for generating and optimizing safety-critical scenarios. [32] further improved ADS training by incorporating reinforcement learning with human feedback into the LLM-based framework. ChatScene utilizes LLM to automatically extract data, generate and retrieve code, render and evaluate scenes and iteratively optimize collision-related parameters, thereby producing safety-critical scenarios for ADS development [33]. Further, Text2Scenario presented in [34] is the most closely related to our method, which employs LLMs to autonomously generate simulation test scenario frameworks from user-provided natural language descriptions. This approach enables the creation of segmented scenarios within simulation environments and significantly enhances the efficiency of autonomous driving testing. Moreover, the closed-loop ProSim framework together with its self-constructed driving dataset exhibits strong performance [35], but it lacks a systematic quality evaluation of the generated scenarios.

III. Methodology

To overcome the limitations of isolated scenario generation methods, we propose a unified framework for scenario construction and trajectory optimization in closed-loop simulation under LLM guidance. As shown in Figure 1, the system transforms descriptions into structured DSL via a RAG process, compiles them into simulation code and refines ego trajectories with convex optimization. To enhance scenario diversity, adversarial agents are trained in the generated scenes. After training, the adversarial agent's weights are frozen, and the ego vehicle is subsequently trained against these fixed behaviors. Successful avoidance cases are appended to the corpus, creating an iterative training loop.

A. Scenario Corpus Construction via Multi-Driven Fusion

Our corpus construction begins with the INTERACTION dataset [12], containing detailed vehicle trajectories at urban intersections. Rule-based extraction methods identify first-level ego vehicle scenarios, such as driving straight, turning, braking, following, and lane changes. Considering interactions between the ego vehicle and background vehicles, adversarial scenario categories are further refined based on background vehicle behaviors, e.g., a straight-moving ego vehicle encountering a left-turning background vehicle. Scene classification uses spatiotemporal features like relative heading, time-to-collision (TTC), post-encroachment time (PET), and lane index transitions. Symbolic traffic knowledge from road network data adds lane properties such as line types, driving directions, and temporary control signals. These features are translated into descriptions (e.g., "temporary road closure," "slow-speed lane with dashed markings") that serve as a knowledge-driven scaffold for understanding the environment. In parallel, adversarial conditions are synthesized by adding behaviors like sudden braking, tailgating, unsafe lane changes, and excessive speeding. Each adversarial vehicle is positioned relative to the ego vehicle with behavior parameters reflecting challenging conditions, e.g., tailgating at 0.5m distance from the rear, or lateral cut-in at 1.2m from the left. This generation expands the corpus' behavioral diversity, enhancing its relevance for testing robustness.

The outputs from these three sources in Fig.1 are unified into a structured intermediate representation. Each scenario consists of five elements: a scene type label, a data-driven behavioral summary, a knowledge-driven road description, an adversarial condition descriptor, and a natural language scene description. This representation forms the basis for prompt-driven generation, allowing the LLM to condition its output on multiple semantic dimensions. The final corpus is serialized as an Excel file, with trajectory snippets and metadata preserved for further parsing. This corpus serves as the input for the retrieval-augmented DSL generation process and forms the primary semantic layer of the adversarial scenario repository.

B. DSL Generation with RAG-Augmented LLM

Generating executable DSL from natural descriptions provided by users based on their specific scenario requirements presents a dual challenge: capturing fine-grained semantic context and preserving structural correctness. RAG is adopted to address these challenges. Unlike traditional vector-space models (VSMs) that rely on surface-level similarity, RAG leverages dense embeddings and example-grounded generative reasoning, thereby improving retrieval fidelity and inference accuracy. This capability is particularly valuable in setting, where scenario fidelity depends not only on matching relevant scene elements, but also on composing structurally valid and contextually appropriate DSL code for downstream execution.

A DSL retrieval corpus is built from the processed scenarios. Each entry contains a structured scene description and its corresponding DSL representation, divided into geometry, generation, and behavior. They are mapped to the road network, positions and adversarial behaviors. The records are embedded using Sentence-BERT [36] and indexed via FAISS [37]. During inference, a new description is transformed into an embedding to retrieve the top-k relevant entries. The generation process employs a CoT prompt, facilitating step-by-step reasoning to generate well-structured DSL. The prompt template incorporates few-shot examples and a scenario component dictionary to guide vocabulary selection and syntax usage. DeepSeek-V3 is employed as the base model due to its reasoning and interactive performance. A semantic alignment verification step is introduced to ensure the semantic fidelity of the generated DSL. The generated elements are first verified against the key attributes and behaviors described in the input natural language scene. If inconsistencies are detected, the model revises the output accordingly. Subsequently, the generated DSL elements are cross-checked against the predefined scenario components in the adversarial scenario repository. If no close match is found, the output is retained as long as it remains consistent with the original scene description. This alignment process guarantees both semantic accuracy and syntactic validity before code generation. To enhance robustness, a self-consistency mechanism [38] samples the LLM multiple

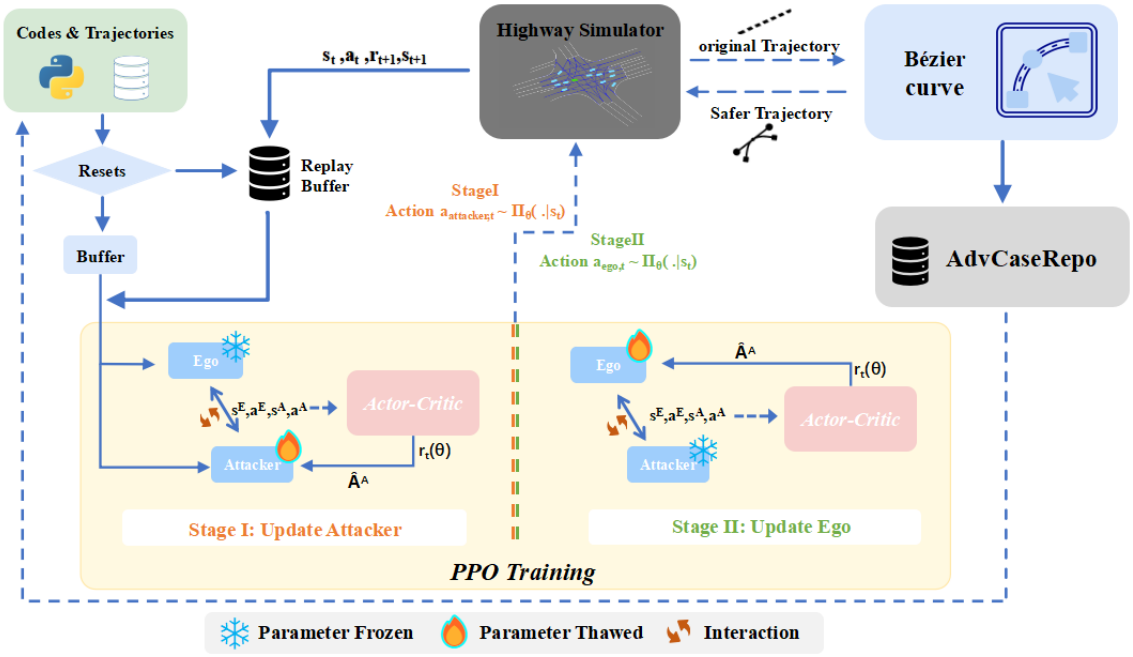


Fig. 2. The framework of newly generated adversarial scenarios.

times per query and uses embedding-based voting to select the most consistent DSL, reducing syntax drift. The finalized DSL, enriched with retrieved scenario elements and code snippets, is input to a code-generation module that produces Python compatible with modified `highway-env`.

C. Adversarial Scenario Repository and Code Generation

To bridge the gap between DSL generation and executable code, an adversarial scenario repository is constructed. It functions as a capability layer consisting of a semantic retrieval base and a code generation interface. Unlike static corpora, the repository is designed as a dynamic and executable resource collection that integrates domain-specific scenario templates, high-level behavior taxonomies, and vectorized code anchors, thereby supporting multi-stage reasoning and simulation integration. The repository consists of three major parts: (i) a scenario corpus containing natural language descriptions paired with manually or semi-automatically generated DSL representations, segmented into geometry, spawn and behavior modules; (ii) a semantic dictionary and scenario taxonomy that define hierarchical component types, enabling controlled generation and alignment verification; and (iii) a FAISS-indexed database of Python code fragments extracted from the modified `highway-env` environment. Each entry in the repository is embedded using Sentence-BERT and annotated with metadata, including scenario intent, vehicle roles, and interaction patterns.

During DSL generation, the repository supports few-shot prompting and component grounding. Given a new scenario description, the top-k semantically similar scene pairs are retrieved to guide CoT reasoning. The hierarchical dictionary constrains vocabulary usage, ensuring syntactic correctness and semantic alignment.

By combining scene abstraction with code-level representation, the repository provides a reusable, extendable, and executable knowledge base that supports high-quality scenario generation and simulation. This design ensures semantic consistency, structural completeness, and enables scalable integration of new scenario types and simulator features. The repository is periodically augmented with successful adversarial-training cases (see Section III-D).

D. Adversarial Agent Training and Corpus Augmentation

To improve the diversity, difficulty and realism of traffic scenarios, we establish a closed-loop simulation cycle that alternates between adversarial training, ego vehicle adaptation, trajectory refinement, and corpus augmentation (as shown in Fig. 2). This cycle refines agent behaviors and incrementally expands the scenario repository with new data from simulation feedback, enhancing scenario diversity beyond existing datasets while maintaining realism.

a) Adversarial and Ego Agent Training.: Based on adversarial vehicle selections from the original trajectory segments, a PPO-based reinforcement learning model is employed to train adversarial behaviors, as shown in Fig 2. An adversarial reward function maintains vehicle aggressiveness and a reset technique [39] reinitializes planner parameters to prevent convergence to local optima. After adversarial agent training is completed, the adversarial vehicle model is fixed and the ego vehicle is trained using a specifically designed reward function to ensure safe decision-making.

b) Trajectory optimization and scenario recollection.: Bézier curve convex optimization is applied to smooth trajectories, reducing excessive lateral acceleration while preserving the original heading constraints. Collision avoidance is prioritized, followed by trajectory planning and execution of the strategy. Successful interactions between

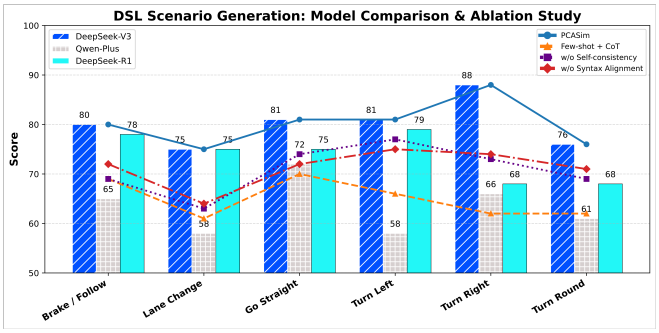


Fig. 3. **Comparison** of DSL scenario generation scores across different LLMs and **ablation** study of core components in the DSL generation pipeline.

ego and adversarial vehicles are retained as closed-loop simulation outputs, thus expanding the repository's coverage of adversarial intentions and response patterns.

Our realism evaluation aligns with the metrics used in the Waymo Sim Agent Challenge, emphasizing kinematic feasibility via Bézier smoothing and interaction fidelity through distribution matching of TTC and PET metrics. This process establishes a self-improving simulation system, where agent robustness and scenario repository are enhanced through two coupled feedback loops: a behavioral loop strengthening robustness and a data loop enriching the corpus with new high-quality examples.

IV. Experiment

A. Experiment Setup

Our experimental framework evaluates the fidelity of generated DSL scenes and assesses the safety-critical performance of the autonomous agent through three key tasks. Firstly, conduct cross-model comparisons among different LLMs to identify their reliability in generating accurate and executable DSL scenarios. An ablation study further investigates the impact of prompt engineering and voting mechanisms, highlighting their contributions to scenario accuracy. Next, generated DSL scenarios are compiled into executable Python scripts and evaluated using a PPO-trained agent within our customized highway-env, focusing on behaviors such as going straight, braking, lane changing, and turning. Performance metrics, including collision rates, timeout occurrences, and conversion success rates, are systematically logged and analyzed across multiple runs to ensure robustness.

B. Experimental Results

1) Comparison of LLM Models under DSL Generation Metrics: We first evaluate how different LLMs perform in generating structured DSL scenes from natural language descriptions. For this, we apply the DSL scoring criteria introduced in Section 4.5, which includes semantic fidelity, executable validity, structural completeness, modularity, behavioral richness and voting centrality. Each LLM (e.g., DeepSeek-V3, Qwen2.5-Plus, DeepSeek-R1) is prompted with identical scenario descriptions using our full pipeline (few-shot + CoT + alignment + voting) and the average scores are reported over a held-out test set of 6 diverse scenes.

As shown in Fig. 3, DeepSeek-V3 consistently achieves the highest DSL generation scores across all ego behavior categories, with an average score of 80.17, outperforming both DeepSeek-R1 (73.83) and Qwen2.5-Plus (63.33). Its advantage is particularly pronounced in more complex behaviors such as Turn Right and Turn Left, where reasoning about interaction logic and scene structure is critical. This suggests that DeepSeek-V3 is better equipped to handle fine-grained semantic composition and structural alignment under our retrieval-augmented and prompt-guided framework. Notably, while DeepSeek-R1 performs comparably in scenarios like Brake/Following and Lane Change, its performance fluctuates more across behaviors, reflecting less stability under diverse traffic configurations. Qwen2.5-Plus, on the other hand, consistently lags behind, indicating limitations in maintaining semantic and syntactic coherence despite being exposed to the same few-shot CoT prompt format.

These results highlight that the choice of LLM substantially influences the quality and fidelity of the driving scenarios generated. This underscores the importance of LLM selection for scenario generation tasks in closed-loop autonomous driving simulations.

2) Comparison with baseline: Since the target is not limited to a single platform and the datasets differ, experiments were not conducted on cases such as KING [24] and FREA [26]. Instead, based on the same simulator and dataset, we selected the method described in Hao's work [23] as the benchmark. Comparisons of generated scenarios in both Fig. 4 and Table I with respect to adversarial metrics such as TTC, PET, and collision rate (CR) clearly demonstrate that our method achieves superior distributions. Notably, PCASim achieves an increase of about 8% in collision rate over baseline method, indicating its enhanced effectiveness in generating challenging adversarial scenarios (Phase I). Moreover, as shown in Fig. 5, the maximum heading angles of RL-validated trajectories enhanced by Bézier smoothing are significantly improved, indicating that our approach produces more natural trajectories and achieves enhanced sim-to-real transferability in scenario resampling (Phase II).

TABLE I
COLLISION RATE COMPARISON

Method	Phase I↑	Phase II↓
Baseline	62.84%	42.17%
Ours	**70.78%**	**39.93%**

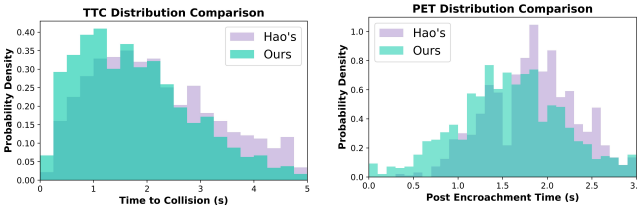


Fig. 4. Evaluation under adversarial criteria

3) Ablation Study on Prompting and Alignment Mechanisms: To quantify the contribution of semantic alignment and self-consistency, we perform ablation experiments using

TABLE II

ABLATION STUDY OF CORE COMPONENTS IN THE DSL GENERATION PIPELINE.

Configuration	Brake/Follow	Lane Change	Go Straight	Turn Left	Turn Right	Turn Round	Average
PCASim	80	75	81	81	88	76	**80.17**
Few-shot + CoT	69	61	70	66	62	62	**64.83**
w/o Self-consistency	69	63	74	77	73	69	**70.83**
w/o Syntax Alignment	72	64	72	75	74	71	**71.33**

* "w/o" stands for "without", indicating the removal of the respective component.

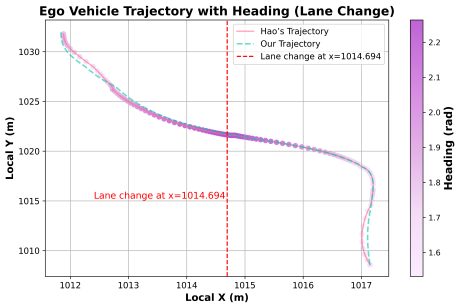


Fig. 5. Comparison of the greatest heading trajectories

DeepSeek-V3. Each configuration's performance is measured by the DSL score across a standardized set of scenarios.

Fig. 3 presents an ablation study that evaluates the contribution of each core component in our DSL generation pipeline. It is clear to observe that the complete pipeline, which includes few-shot prompting, CoT reasoning, semantic alignment and self-consistency voting, achieves the highest overall performance with an average DSL score of 80.17, showing a 12% improvement in the accuracy of domain-specific language generation compared to configurations without semantic alignment, as shown in Table II. This indicates that while few-shot prompting and reasoning are foundational for structured generation, they are insufficient alone to ensure high semantic fidelity and executable validity. The moderate score drop (70.83) following the removal of self-consistency suggests its primary contribution lies in stability. In contrast, the absence of semantic alignment (71.33) reveals its critical impact on DSL-to-description fidelity, especially for intricate behaviors like left and U-turns.

These results underscore a modular synergy: few-shot prompting and CoT establish a robust generative foundation, while alignment and voting act as critical post-hoc safeguards to rectify inconsistencies and suppress variance. Altogether, the full pipeline ensures both structural integrity and semantic correctness in the generated scenarios, confirming the necessity of integrating reasoning, validation and aggregation in closed-loop DSL generation. Specifically, the RAG and semantic alignment modules ensure the 'direction' by grounding DSL to user prompts, while the Phase I RL training injects the 'adversarial' nature by optimizing conflict intensity.

4) Collision and Timeout Metrics in Adversarial Scenarios: To evaluate adversarial strategy learning in multi-vehicle scenarios, we designed a training and evaluation framework that tracks collision and timeout rates in four scenario types: braking, going straight, turning, and lane changing. Multiple scenes are sampled during each training session and run in parallel using a shared PPO model. Trained vehicles are initialized in their own environment and controlled by the PPO policy, while other vehicles follow a fixed behavior. After collecting trajectories for all scenarios, we aggregate the data and perform a policy update. This centralized update strategy ensures consistent learning while benefiting from multi-process rollout acceleration. This framework enables systematic large-scale adversarial training and provides metrics (mean performance) for evaluating the impact of adversarial behavior under various realistic traffic conditions. As shown in Fig. 6, by introducing PPO combined with a reset technique for training, the ego vehicle demonstrates safer obstacle avoidance capabilities against adversarial vehicles compared to the untrained baseline.

TABLE III

COLLISION RATE AND TIMEOUT STATISTICS UNDER ADVERSARIAL SETTINGS.

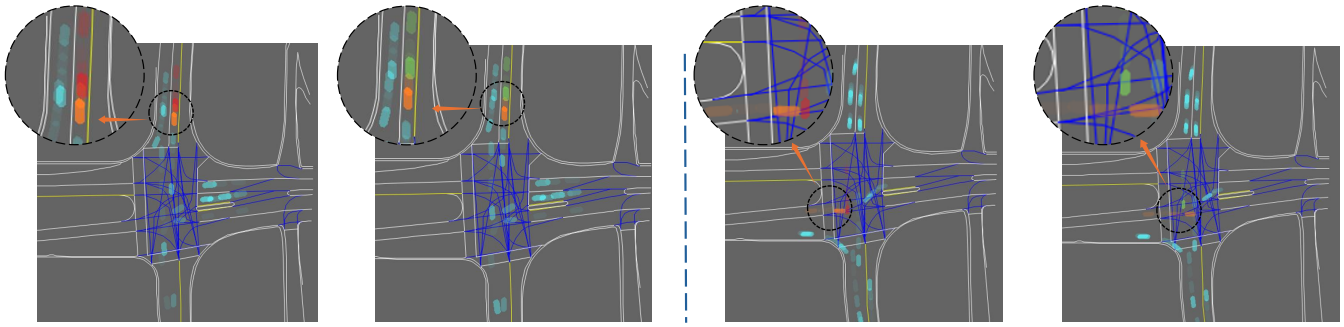
Ego Behavior	Adv against ego		Ego avoid adv	
	Collision Rate	Timeout	Collision Rate	Timeout
Brake	72.7%	3	46.1%	7
Lane Change	57.1%	3	24.2%	6
Go Straight	73.3%	4	46.6%	8
Turn	80.0%	2	42.8%	8

Note. Each experimental metric is taken from the average of multiple runs.

The trained ego vehicle exhibits markedly lower collision rates than the untrained baseline — demonstrating greater robustness and adaptability in adversarial scenarios of lane change, straight-driving, and turning. In Table III, the avoidance policy lowers collision rates by roughly 30 % on average while maintaining stable timeout counts, thus boosting safety without sacrificing efficiency.

5) Comparative Studies with Recent Related Work: To provide context for our framework, we conducted a comparative analysis with recent related works in autonomous-driving scenario generation, as summarized in Table IV.

While frameworks like Text2Scenario and ProSim enable prompting for scene creation, they operate without a closed-loop validation mechanism, which limits their reliability for safety-critical testing. Text2Scenario converts natural language to OpenScenario DSL in an open-loop manner, and ProSim focuses on multi-modal realism rather than stress testing. Conversely, CAT and VCAT are designed to generate adversarial scenarios to improve system robustness. For



(a) (b) (c) (d)

Fig. 6. **Generating Scenarios and Validation.** (a) (b) are the braking scenarios, (c) (d) are the going straight scenarios, the orange indicates adversarial vehicle, the red indicates ego vehicle without training and the green ego vehicle has been strengthened by the reinforcement learning.

TABLE IV
COMPARISON OF SCENARIO-GENERATION FRAMEWORKS FOR AUTONOMOUS-DRIVING SIMULATION.

Method	Promptable	Validation	Training Aug.	Scene Repo (self).
Text2Scenario [34]	✓	✗	✗	✓
ProSim [35]	✓	✗	✗	✓
CAT [40]	✗	✗	✓	✗
VCAT [41]	✗	✗	✓	✗
Ours	✓	✓	✓	✓

instance, CAT accelerates adversarial synthesis, while VCAT achieves a high attack success rate by using vulnerability-aware rewards. Our framework distinguishes itself by integrating all these capabilities: it combines natural-language prompting with closed-loop validation and automated scene exploration without needing a pre-built repository. This unique combination allows our system to explore both nominal and extreme traffic conditions, yielding a 12% gain in DSL accuracy and a 30% boost in executable scene yield. This integrated approach ensures both controllability and validation, making our framework effective for comprehensive testing.

V. LIMITATIONS

Our framework is subject to several important limitations that constrain its generalizability and scalability. The scenario corpus is currently constructed from the INTERACTION dataset [12], which, despite its richness in urban interactions, provides limited geographical and behavioral diversity. The simulation environment relies on `highway-env`, which has a simplified kinematic model facilitates efficient experimentation but lacks high-fidelity dynamics and multimodal sensing capabilities. Furthermore, the pipeline depends on externally hosted LLM APIs (e.g., GPT-5 [42], Claude 4 [43]), which introduces uncertainties in availability, stability, and scalability due to external constraints.

Methodological limitations include DSL conversion errors due to map format mismatches and API-driven inconsistencies in LLM generation. Moreover, insufficient traffic-rule reasoning and the RL's narrow focus on basic maneuvers—omitting complex multi-stage dynamics—collectively diminish the diversity and robustness of the resulting system.

Future work will focus on mitigating these limitations by incorporating heterogeneous datasets (e.g., NGSIM, nuScenes, Waymo) to broaden scenario diversity, migrating to high-fidelity simulators such as CARLA [44], and developing domain-specific LLMs with enhanced traffic reasoning.

VI. CONCLUSION

This paper presents PCASim, a promptable closed-loop simulation framework for generating and evaluating safety-critical urban traffic scenarios. The framework incorporates an adversarial scenario repository together with a RAG+LLM-based prompt engineering paradigm, enabling efficient, diverse, and interpretable scenario generation tailored to specified prompts. An RL-based traffic-flow model is employed to jointly control the ego and adversarial vehicles, which not only validates the fidelity of the generated scenarios but also produces additional safety-critical data. Moreover, a closed-loop generation–evaluation pipeline is designed to iteratively filter and enrich the repository, alleviating the limitations of small-sample distributions and ensuring comprehensive coverage of challenging traffic events.

Compared with existing approaches, PCASim achieves a higher degree of directedness and controllability in adversarial scenario generation. By leveraging prompt-guided DSL translation with semantic alignment, the framework generates qualitatively more adversarial and diverse cases than rule-based or purely generative methods. In combination with PPO-based reinforcement learning and Bézier-curve convex optimization, the proposed system substantially improves the robustness of autonomous vehicles, equipping them with stronger capabilities to navigate complex and adversarial urban environments.

REFERENCES

[1] Waymo, "Waymo's incidents: As of the end of 2024, the national highway traffic safety administration (nhtsa) had received 835 reports documenting 696 incidents involving waymo vehicles," 2024, accessed: 2025-04-14. [Online]. Available: https://en.wikipedia.org/wiki/Waymo

[2] Tesla, "As of 2024, there have been 51 reported fatalities involving tesla's autopilot function, with 44 verified by nhtsa or expert testimony." 2024, accessed: 2025-04-14. [Online]. Available: https://en.wikipedia.org/wiki/Tesla_Autopilot

[3] F. Gao, J. Mu, X. Han, Y. Yang, and J. Zhou, "Performance limit evaluation strategy for automated driving systems," *Automotive Innovation*, vol. 5, no. 1, pp. 79–90, 2022.

[4] Z. Wei, H. Zhou, and R. Zhou, "Risk and complexity assessment of autonomous vehicle testing scenarios," *Applied Sciences*, vol. 14, no. 21, p. 9866, 2024.

[5] Y. Wang, Z. Han, Y. Xing, S. Xu, and J. Wang, "A survey on datasets for the decision making of autonomous vehicles," *IEEE Intelligent Transportation Systems Magazine*, vol. 16, no. 2, pp. 23–40, 2024.

[6] W. Ding, C. Xu, M. Arief, H. Lin, B. Li, and D. Zhao, "A survey on safety-critical driving scenario generation—a methodological perspective," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 7, pp. 6971–6988, 2023.

[7] C. Zhang, Z. Wang, J. Wang, K. Su, Q. Lv, B. Jiang, K. Hao, and W. Wang, "Generative modeling of adversarial lane-change scenario," *arXiv preprint arXiv:2503.12055*, 2025.

[8] Y.-H. Yin, X. Lü, R. Jiang, B. Jia, and Z. Gao, "Kinetic analysis and numerical tests of an adaptive car-following model for real-time traffic in its," *Physica A: Statistical Mechanics and its Applications*, vol. 635, p. 129494, 2024.

[9] J. Wang, A. Pun, J. Tu, S. Manivasagam, A. Sadat, S. Casas, M. Ren, and R. Urtasun, "Advsim: Generating safety-critical scenarios for self-driving vehicles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9909–9918.

[10] L. Feng, Q. Li, Z. Peng, S. Tan, and B. Zhou, "Trafficgen: Learning to generate diverse and realistic traffic scenarios," in *2023 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2023, pp. 3567–3575.

[11] 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.

[12] W. Zhan, L. Sun, D. Wang, H. Shi, A. Clausse, M. Naumann, J. Kummerle, H. Konigshof, C. Stiller, A. de La Fortelle *et al.*, "Interaction dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps," *arXiv preprint arXiv:1910.03088*, 2019.

[13] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[14] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan *et al.*, "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437*, 2024.

[15] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei *et al.*, "Qwen2. 5 technical report," *arXiv preprint arXiv:2412.15115*, 2024.

[16] S. Tan, B. Ivanovic, X. Weng, M. Pavone, and P. Kraehenbuehl, "Language conditioned traffic generation," in *7th Annual Conference on Robot Learning*, 2023.

[17] J. M. Scanlon, K. D. Kusano, T. Daniel, C. Alderson, A. Ogle, and T. Victor, "Waymo simulated driving behavior in reconstructed fatal crashes within an autonomous vehicle operating domain," *Accident Analysis & Prevention*, vol. 163, p. 106454, 2021.

[18] Z. Yang, Y. Chai, D. Anguelov, Y. Zhou, P. Sun, D. Erhan, S. Rafferty, and H. Kretzschmar, "Surfelgan: Synthesizing realistic sensor data for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 118–11 127.

[19] Z. Wei, H. Huang, G. Zhang, R. Zhou, X. Luo, S. Li, and H. Zhou, "Interactive critical scenario generation for autonomous vehicles testing based on in-depth crash data using reinforcement learning," *IEEE Transactions on Intelligent Vehicles*, 2024.

[20] R. Lee, M. J. Kochenderfer, O. J. Mengshoel, G. P. Brat, and M. P. Owen, "Adaptive stress testing of airborne collision avoidance systems," in *2015 IEEE/AIAA 34th Digital Avionics Systems Conference (DASC)*. IEEE, 2015, pp. 6C2–1.

[21] Y. Abeysirigoonawardena, F. Shkurti, and G. Dudek, "Generating adversarial driving scenarios in high-fidelity simulators," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 8271–8277.

[22] A. Prakash, S. Boochoon, M. Brophy, D. Acuna, E. Cameracci, G. State, O. Shapira, and S. Birchfield, "Structured domain randomization: Bridging the reality gap by context-aware synthetic data," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 7249–7255.

[23] K. Hao, W. Cui, Y. Luo, L. Xie, Y. Bai, J. Yang, S. Yan, Y. Pan, and Z. Yang, "Adversarial safety-critical scenario generation using naturalistic human driving priors," *IEEE Transactions on Intelligent Vehicles*, 2023.

[24] N. Hanselmann, K. Renz, K. Chitta, A. Bhattacharyya, and A. Geiger, "King: Generating safety-critical driving scenarios for robust imitation via kinematics gradients," in *European Conference on Computer Vision*. Springer, 2022, pp. 335–352.

[25] Y. Mei, T. Nie, J. Sun, and Y. Tian, "Llm-attacker: Enhancing closed-loop adversarial scenario generation for autonomous driving with large language models," *arXiv preprint arXiv:2501.15850*, 2025.

[26] K. Chen, Y. Lei, H. Cheng, H. Wu, W. Sun, and S. Zheng, "Frea: Feasibility-guided generation of safety-critical scenarios with reasonable adversariality," *arXiv preprint arXiv:2406.02983*, 2024.

[27] G. Bagschik, T. Menzel, and M. Maurer, "Ontology based scene creation for the development of automated vehicles," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 1813–1820.

[28] P. Cai, Y. Lee, Y. Luo, and D. Hsu, "Summit: A simulator for urban driving in massive mixed traffic," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 4023–4029.

[29] M. Klischat, E. I. Liu, F. Holtke, and M. Althoff, "Scenario factory: Creating safety-critical traffic scenarios for automated vehicles," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2020, pp. 1–7.

[30] Y. Deng, J. Yao, Z. Tu, X. Zheng, M. Zhang, and T. Zhang, "Target: Automated scenario generation from traffic rules for testing autonomous vehicles," *arXiv preprint arXiv:2305.06018*, 2023.

[31] H. Tian, X. Han, G. Wu, Y. Zhou, S. Li, J. Wei, D. Ye, W. Wang, and T. Zhang, "An llm-enhanced multi-objective evolutionary search for autonomous driving test scenario generation," *arXiv preprint arXiv:2406.10857*, 2024.

[32] Y. Sun, N. Salami Pargoo, P. Jin, and J. Ortiz, "Optimizing autonomous driving for safety: A human-centric approach with llm-enhanced rlhf," in *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2024, pp. 76–80.

[33] J. Zhang, C. Xu, and B. Li, "Chatscene: Knowledge-enabled safety-critical scenario generation for autonomous vehicles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 459–15 469.

[34] X. Cai, X. Bai, Z. Cui, D. Xie, D. Fu, H. Yu, and Y. Ren, "Text2scenario: Text-driven scenario generation for autonomous driving test," *arXiv preprint arXiv:2503.02911*, 2025.

[35] S. Tan, B. Ivanovic, Y. Chen, B. Li, X. Weng, Y. Cao, P. Krähenbühl, and M. Pavone, "Promptable closed-loop traffic simulation," *arXiv preprint arXiv:2409.05863*, 2024.

[36] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[37] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with gpus," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019.

[38] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," *arXiv preprint arXiv:2203.11171*, 2022.

[39] E. Nikishin, M. Schwarzer, P. D'Oro, P.-L. Bacon, and A. Courville, "The primacy bias in deep reinforcement learning," in *International conference on machine learning*. PMLR, 2022, pp. 16 828–16 847.

[40] L. Zhang, Z. Peng, Q. Li, and B. Zhou, "Cat: Closed-loop adversarial training for safe end-to-end driving," in *7th Annual Conference on Robot Learning*, 2023.

[41] X. B. R. K. Z. M. H. Y. Y. R. Xuan Cai, Zhiyong Cui*, "Vcat: Vulnerability-aware and curiosity-driven adversarial training for enhancing autonomous vehicle robustness," *arXiv preprint arXiv:2304.02391*, 2024.

[42] OpenAI, "Gpt-5 technical overview," https://openai.com, 2025.

[43] Anthropic, "Claude 4 model family," https://www.anthropic.com, 2025.

[44] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proceedings of the 1st Annual Conference on Robot Learning*, 2017, pp. 1–16.