SCP: Scalable and Customizable Generation of Planning-specific Corner Cases in Autonomous Driving

Lingfeng Zhou[1], Jin Gao[1], Mohan Jiang[1], Yufeng Liu[1], Yuankai Li[2], and Dequan Wang[1,3*]

***Abstract*— Advances in autonomous driving have improved algorithmic performance, but handling rare corner cases remains challenging. Existing methods for generating corner cases often lack scalability, particularly in planning-specific scenarios. In this paper, we propose SCP, a scalable and customizable approach for planning corner case generation that uses Large Language Models (LLMs). We introduce SCP-NuPlan, a benchmark derived from the NuPlan dataset, and SCP-LimSim, a simulator for creating corner cases via natural language. Our experiments show significant performance drops in planning algorithms tested on SCP-NuPlan, underscoring the difficulty of corner cases and the need for more robust solutions. We also demonstrate the customization of agents in SCP-LimSim by extensive experiments. We will release SCP-NuPlan and SCP-LimSim soon. Visit our website for more details https://maple-zhou.github.io/SCP.**

I. INTRODUCTION

Recent progress in autonomous driving research has been propelled by advancements in algorithmic development and system integration. On the algorithmic side, significant strides have been made in perception, planning, and control, supported by datasets and benchmarks such as NuPlan [1], Waymo [2], and Lyft Level-5 [3]. Furthermore, system integration ensures that various modules operate cohesively in both real-world and simulated environments, exemplified by platforms like CARLA [4] and HighwayEnv [5]. While typical driving environments are well-handled by existing algorithms, real-world deployment requires driving systems to handle unpredictable and extreme situations, which are called corner cases.

To bridge this gap, researchers propose to generate synthetic corner cases instead of collecting corner case data which is costly, and time-consuming. While considerable progress has been made in generating diverse corner cases for perception modules [6], [7], [8], [9], [10], [11], studies in synthetic planning-specific corner cases remain limited. Specifically, random sampling [12], [13] often lacks authenticity due to its disregard for environmental context; model-based methods [14], [15] are labor-intensive and lack scalability; data-driven techniques [16] are challenging to control and lack explicit modeling capabilities. Furthermore, these methods primarily focus on fixed datasets or benchmarks and have yet to explore flexible and dynamic simulation environments.

In this paper, we propose a novel method for **S**calable and **C**ustomizable generation of **P**lanning-specific corner cases, which we refer to as **SCP**. SCP leverages Large Language Models (LLMs) to scale across diverse scenarios and customize agent behaviors and environmental variables based on nuanced language descriptions for flexible extension. SCP provides a benchmark consisting of corner cases (**SCP-NuPlan**) for planning algorithm evaluation, and a simulator (**SCP-LimSim**) where users can create corner cases easily for flexible corner case simulation.

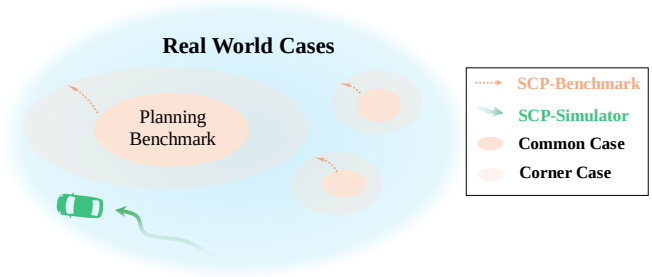

Fig. 1. **Our SCP facilitates the corner case generation.** SCP-Benchmark helps expand the case boundary of existing benchmarks in a *scalable* manner. SCP-Simulator supports *customizable* corner case simulation.

SCP-NuPlan enables scalable evaluation by converting the NuPlan [1] into a corner case benchmark, leveraging LLMs to design agent trajectories which are rare in the real world based on global environmental information. We modify 6382 scenarios from the NuPlan dataset, covering 14 distinct types from the NuPlan closed-loop challenge [1], each lasting over five seconds. Such an automated process is extensible, allowing for rapid expansion to other datasets and facilitating the development of a more diverse and comprehensive benchmark. Our extensive experiments on SCP-NuPlan, involving three different types of planning algorithms, reveal a significant performance decline compared to the original NuPlan dataset, underscoring the need for more robust planning algorithms.

While SCP-NuPlan focuses on global scenario design, our SCP-LimSim utilizes LLMs to generate the rare activity of each agent by fine-grained language description. SCP-LimSim builds upon the LimSim++ [18], extending the benchmark by offering users an open platform to design customized corner case scenarios and environments. It enables more intuitive and efficient simulation, such as defining *an ambulance in a hurry* and setting a driver's character to *angry*. We validate SCP-LimSim's effectiveness by exhibiting diverse driving behaviors of agents in response to natural language inputs. Moreover, we develop several complex corner case scenarios to showcase the simulator's capability to create intricate environments. Additionally, SCP-LimSim

[1]Shanghai Jiao Tong University, [2]Fudan University, [3]Shanghai Artificial Intelligence Laboratory, *Corresponding Author.

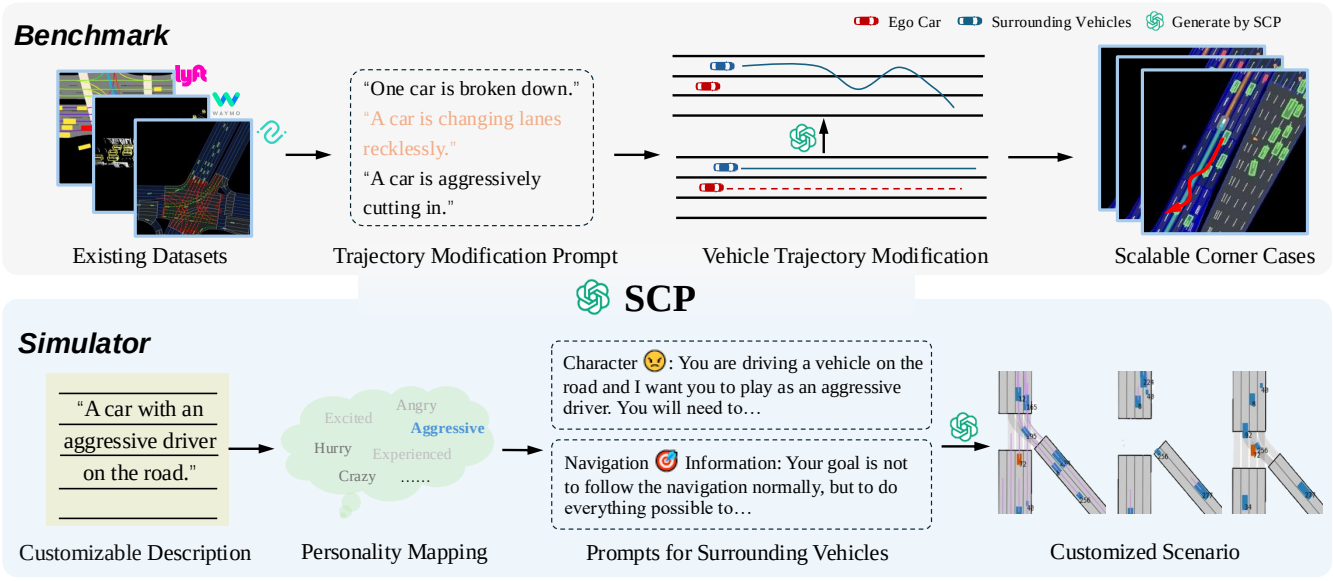

Fig. 2. **SCP provides a scalable and customizable approach utilizing LLMs for generating planning-specific corner cases.** *Top*: SCP-NuPlan builds upon existing dataset, NuPlan [1], and modifies vehicle trajectories using global information for scalability. *Bottom*: SCP-LimSim is established based on LimSim platform [17] and obtains high-level customization by individual control using natural language.

supports co-simulation with SUMO [19] and CARLA [4].

Our main contributions are summarized as follows:

- We present a scalable and customizable method, SCP, for planning-specific corner case generation using LLMs.
- We introduce a scalable corner case benchmark SCP-NuPlan that transforms existing datasets into diverse and challenging evaluation scenarios.
- We provide a flexible simulator SCP-LimSim that enables users to create customized corner cases via natural language commands.

II. RELATED WORK

A. Autonomous Driving Planning Benchmark & Simulator

Benchmarks have been instrumental in advancing autonomous driving research, particularly in developing and testing planning algorithms [20], [21]. Several benchmarks and datasets have been developed to support the evaluation of planning algorithms. NuPlan [1], the first benchmark explicitly designed for autonomous vehicle planning, allows researchers to assess their algorithms under real-world conditions. The Waymo Open Dataset [2] provides extensive real-world driving data, particularly in complex urban environments, facilitating a broad range of research applications, including planning. Similarly, the Lyft Level-5 dataset [3] offers high-resolution sensor data on a large scale, further contributing to research in autonomous driving planning. Other datasets, such as nuScenes [22], though primarily aimed at perception tasks, are also utilized for planning research due to their detailed scene annotations. These benchmarks vary in scope, data volume, and specific use cases, offering diverse options for evaluating the effectiveness of various planning methods.

In addition to benchmarks, numerous simulators have been introduced to support planning research. CARLA [4], an open-source simulator, is widely adopted for its high-fidelity simulations of diverse environments, weather conditions, and vehicle dynamics, making it well-suited for developing and testing planning algorithms. AirSim [23] is another simulator tailored for autonomous driving research, while LGSVL [24] provides customizable scenarios and integrates with platforms like Apollo [25], supporting real-time simulation of vehicle dynamics and sensor models. Waymax [26] focuses on large-scale simulation and testing of decision-making and planning modules. MetaDrive [27] emphasizes photorealistic scenarios, particularly beneficial for end-to-end learning approaches, while SMARTS [28] targets multi-agent interactions, focusing on social driving behavior. Tools such as SUMO [19] and Prescan [29] allow for customizable, reproducible traffic scenarios using real-world trajectory data for robust evaluations. Lightweight simulators like LimSim [17] and HighwayEnv [5] provide convenient, rapid deployment environments for simulation testing. These simulators offer flexibility and scalability, serving both academic and industrial research in autonomous driving.

While there are numerous benchmarks and simulators available for planning research, none are specifically designed to address corner case scenarios, leading to a critical gap between testing and development of planning algorithms. Our proposed SCP-NuPlan fills this gap by providing a robust evaluation framework that challenges algorithms to handle rare and unpredictable situations that are difficult to capture in standard datasets. Meanwhile, SCP-LimSim serves as a versatile platform where developers can efficiently create and customize personalized scenarios, offering greater

flexibility in simulating complex driving environments.

B. Corner Case Generation in Autonomous Driving

The long tail effect in autonomous driving has driven researchers to explore efficient methods for generating corner cases. Effective corner case generation requires accurate identification and a sufficient volume of relevant data [30]. Recent works have explored the generation of diverse scenarios for vehicle perception, including corner cases. DriveDreamer [6] enables controllable video generation to simulate real-world traffic scenes. Other research leverages diffusion models, such as GAIA-1 [7] and MagicDrive [8], while Neural Radiance Field (NeRF)-based approaches [31], [32], [33], [34] focus on building static backgrounds and active vehicles. Notable contributions also include UniSim [9] and MARS [10], with ChatSim [11] enabling automatic simulation editing via natural language commands.

While significant advances have been made in generating corner cases for vehicle perception, progress in planning-level corner case generation remains limited. Current methods can be categorized into random sampling, model-based, data-driven, and scenario-based approaches [14]. Random sampling methods [12], [13] often lack authenticity due to their disregard for environmental context. Model-based approaches [14], [15], use covering arrays and falsification techniques to model critical corner cases but are labor-intensive and lack scalability. Data-driven approaches [16], learn latent representations from real-world data to generate plausible out-of-domain scenarios. However, they are challenging to control and lack explicit modeling capabilities. Scenario-based methods typically rely on expert input or accident reports; for example, Pretschner et al.[35] simulate critical scenarios, while Kluck et al. [36] employ ontologies to sample corner cases based on vehicle-environment interactions.

Current methods are limited by their reliance on manual labor, lack of flexibility, and rigid frameworks. However, our proposed SCP addresses these issues by integrating Large Language Models (LLMs) to generate corner cases from global and individual perspectives. By leveraging LLMs' ability to interact with users through natural language, SCP enables scalable and customizable corner case generation, tailored to specific research needs.

III. METHOD

We introduce the SCP, which offers a novel framework leveraging Large Language Models (LLMs) for generating planning-specific corner cases in autonomous driving, depicted in Sec. III-A. SCP can be adopted to corner case benchmarks and simulators. We establish a benchmark SCP-NuPlan using LLMs to integrate global information for scalability, shown in Sec. III-B. As for simulators, we build the SCP-LimSim by controlling individuals using LLMs for customizability. SCP creates corner cases from both global and individual perspectives, offering a controllable and efficient method. These two approaches form the basis for the development of SCP-NuPlan and SCP-LimSim.

A. SCP

We propose a novel method for scalable and customizable generation of planning-specific corner cases, which we refer to SCP. Leveraging the capabilities of large language models (LLMs), our approach is scalable, as it can efficiently generate vast amounts of diverse corner cases with minimal human intervention, accommodating a wide range of driving scenarios. Additionally, it is customizable, allowing users to tailor the complexity and type of corner cases based on specific parameters such as vehicle behavior, road conditions, or environmental variables, ensuring that the generated scenarios are aligned with their unique needs. Finally, the method is designed to be plug-and-play, seamlessly integrating with existing planning algorithms and simulation environments, enabling researchers and engineers to easily incorporate these corner cases into their workflows without extensive reconfiguration or setup.

Leveraging SCP, we build a benchmark, SCP-NuPlan, and a simulator, SCP-LimSim. SCP-NuPlan is constructed by modifying the existing dataset, NuPlan [1], utilizing the LLMs' knowledge concerning corner cases from a global perspective, shown in Sec. III-B. This framework is highly scalable, enabling seamless expansion to larger and more comprehensive benchmarks. On the other hand, SCP-LimSim focuses on generating corner cases from an individual driver perspective, shown in Sec. III-C. It utilizes LLMs to independently simulate each agent and control their actions, allowing for customized behaviors through natural language commands. This approach provides a high level of control and flexibility in generating specific agent behaviors needed for corner case simulation.

B. SCP-NuPlan: Global Information for Scalability

Here we introduce our proposed benchmark, SCP-NuPlan, shown in Fig. 2. We leverage LLMs to integrate the positions of all vehicles and timestamps from the global context based on the real-world planning dataset, NuPlan [1]. However, given the limitations of the LLMs' context window and the risk of diluted attention from excessive timestamp data, we avoid supplying the model with the full sequence of timestamps in the scenario. Instead, we extract a set of keyframes that effectively represent the changes in both ego and agent trajectories. This enables the identification of the most suitable corner case scenarios and the agents best positioned to enact such behaviors. To preserve the coherence of the agent trajectories, we pre-identify key agents based on their proximity to the ego vehicle. By focusing on agents that are consistently near the ego vehicle, LLMs can apply targeted modifications to their trajectories. Moreover, modifications to the agent trajectories are governed by a set of customizable constraints, allowing for flexible and controlled corner case generation.

Since we aim to generate corner cases, we do not have the ground truth trajectories of the ego vehicle, which is collected from the real world in NuPlan. Therefore, we only focus on the close-loop challenge of NuPlan, which

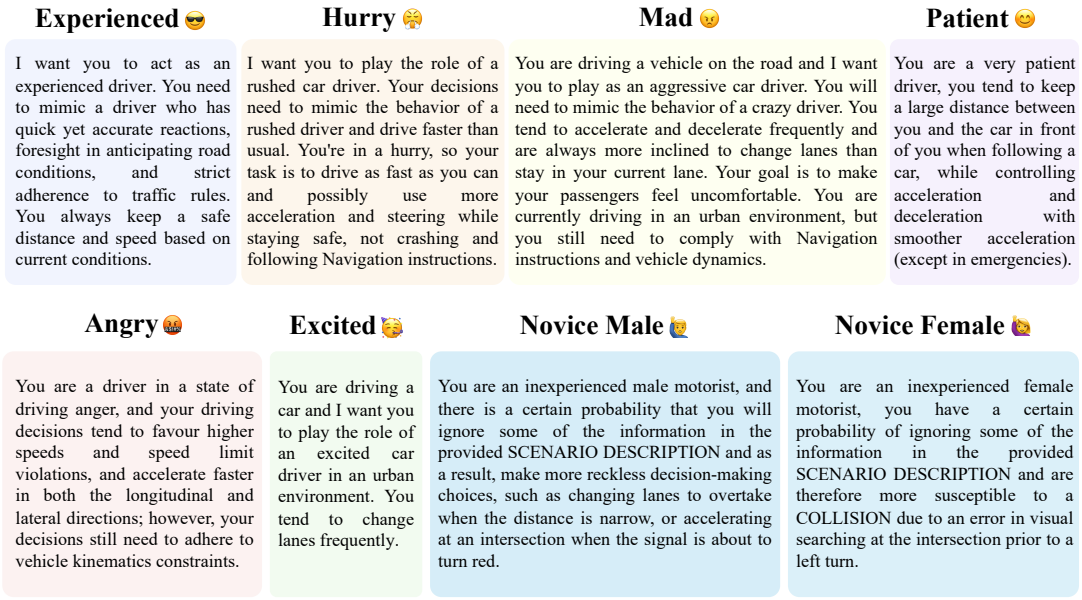

Fig. 3. **Our SCP-LimSim accommodates a wide spectrum of driving personas, including but not limited to *Experienced*, *Hurry*, *Mad*, *Patient*, *Angry*, *Excited*, *Novice Male*, and *Novice Female*.** These characters embody unique driving behaviors, strategies, and reactions, precisely modeled within a simulated environment. Customizing these characters is both practical and simple, facilitating their adaptation to specific research or application requirements. Moreover, our framework accommodates a wide range of character profiles, thereby enhancing the simulation's realism and applicability.

cares about 14 types of scenarios including high-magnitude-speed, changing-lane, etc. As for the close-loop challenge, NuPlan provides two simulation types, non-reactive and reactive agents. Since reactive agents are controlled by a simple controller Intelligent Driver Model (IDM) [37], and they may actively avoid other vehicles, which may make the corner case less difficult, thus we only take the close-loop non-reactive challenge as our pre-modified benchmark. Specifically, we extract the scenarios lasting over 5 seconds, considering that the NuPlan close-loop challenge only covers the scenarios more than 8 seconds. We eventually processed all 1403 logs in both test split and mini split of NuPlan, and modified 6382 scenarios.

C. SCP-LimSim: Individual Control for Customization

In this section, we introduce SCP-LimSim, an advanced framework that facilitates users to customize agent driving styles by controlling individual agents, shown in Fig. 2. While SCP in Sec. III-B concentrates on tackling global information for scalability of corner case generation, SCP in this section focuses on flexibility and customization of corner case simulator. The agents in SCP-LimSim are controlled by Large Language Models (LLMs), corresponding to the individual perspective of corner case generation. Users can control agents at a fine granularity, such as vehicle category and driver personality. Additionally, SCP-LimSim integrates SUMO [19], CARLA [4], and LimSim++ [18] for perception-planning connection.

To integrate Large Language Models (LLMs) into the planning module of agents, it is necessary to contextualize the driving environment for LLMs by describing the driving context in a way that LLMs can comprehend the current state. For example, past states are especially important for agents to keep the movements consistent, avoiding other actions during a lane turning for instance. These parameters enable LLMs to choose the most appropriate action from a predefined set, outlined in Tab. I. The actions are categorized into five main types, focusing primarily on speed adjustments and lane-changing maneuvers. Then, LLMs are required to return acceleration range and steering angle based on the chosen action. LLMs need to return the action and the required values according to the set character or other instructions, which is the key to performing customizable driving behaviors. We have implemented 8 distinct characters (shown in Fig. 3 to show the convenience of customizing agents in SCP-LimSim, with the experiment results validating the distinction in Sec. IV-B. Furthermore, LLMs provide explanations to clarify their planning process, greatly improving the rationality of agents' decisions in SCP-LimSim. Given that LLMs are not very precise with actual numerical values, we will only have LLMs return the range of acceleration and steering angles, and employ the Intelligent Driver Model (IDM) [37] to calculate the most suitable control signals within these ranges. Compared to directly selecting actions,

TABLE I
PREDEFINED AVAILABLE ACTIONS AND DESCRIPTIONS.

Action	Description
Turn-left	Change lane to the left of the current lane
Turn-right	Change lane to the right of the current lane
IDLE	Remain in the current lane with current speed
Acceleration	Accelerate the vehicle in the current lane
Deceleration	Decelerate the vehicle in the current lane

this approach allows LLMs to control the agent's behavior with finer granularity while avoiding significant issues caused by extreme outlier values.

IV. EXPERIMENT

In this section, we provide experiments of SCP-NuPlan and SCP-LimSim. In Sec. IV-A, we provide a detailed comparison between SCP-NuPlan and NuPlan among three algorithms. In Sec. IV-B, we demonstrate the high-level customization of SCP-LimSim on agent driving styles and corner cases

A. SCP-NuPlan

We first test some planning models on SCP-NuPlan to show the effectiveness of SCP. We utilize planTF [39] to evaluate SCP-NuPlan given that it outperforms PDM [40], the winner of the 2023 NuPlan Challenge, and represents the latest open-source state-of-the-art (SOTA) approach. For comparative analysis, we select two additional models: the classic UrbanDriver [38] and the CNN-based approach, PlanCNN, offered in NuPlan [1]. In terms of methodology, we adopt the same testing protocols used in the planTF evaluation. PlanTF employs two distinct scenario selection schemes in its testing: Test14-random and Test14-hard. Test14-random approach randomly samples 20 scenarios from each scenario type in NuPlan. Test14-hard leverages PDM-close [40] as the baseline algorithm. Specifically, Test14-hard involves running 100 scenarios for each type of NuPlan Challenge and selecting the 20 least-performing scenarios to form the final set.

We notice that NuPlan provides a special mini split for simple tests or development, and thus we also applied SCP to the mini split to obtain SCP-NuPlan-Mini. However, since the mini split only contains 10 logs from the test split, SCP-NuPlan-Mini only has 186 scenarios of Test14-random and 7 scenarios of Test14-hard, leading to a lack of representativeness and great variance. Therefore, we conduct a pre-experiment to see whether the mini-split is valid. We directly run the three models on the mini split under Test14-random and Test14-hard, as shown in Tab. III. We use the non-reactive close-loop score (NR-CLS). The results indicate that planTF's performance on the Mini split is consistent with the reported results, while UrbanDriver and PlanCNN both demonstrate improved performance, with PlanCNN even surpassing planTF's replicated performance. The results confirm our suspicion that the small number of scenarios will lead to a large variance. Therefore, to cater to NuPlan's official idea of providing small data sets for rapid testing, we implement a medium benchmark SCP-NuPlan-Air, which includes SCP-NuPlan-Mini and 336 additional logs from the Test split, with 400 logs and 2,000 modified scenarios in total, including 249 Test14-random scenarios and 89 Test14-hard scenarios. The full SCP-NuPlan consists of 1403 logs from the test split and the mini split, with 6382 scenarios.

Using three benchmark scales, we evaluated the performance of PlanCNN [1], UrbanDriver [38], and planTF [39], with the results presented in Tab. II. We can tell from the results that SCP-NuPlan indeed brings some difficulties to the planning models and the bigger benchmark results in the worse performance of all three models. Under Test14-random, SCP-NuPlan brings a huge performance drop to the models, while SCP-NuPlan-Mini and SCP-NuPlan-Air also make the models decline, which ensures that the two small splits are valid corner case benchmarks when satisfying the rapid test requirement. Furthermore, considering the inherent difficulty of some scenarios in the NuPlan dataset, particularly those resembling corner cases, it is likely that Test14-hard already contains challenging scenarios. Comparing the results of SCP-NuPlan with the original results under Test14-hard, it can be found that whether under Test14-random or Test14-hard, the experimental results of SCP-NuPlan are equivalent to or lower than the original results. This indicates that SCP-NuPlan has scaled the NuPlan to a more difficult level, at least to the Test14-hard level. And the scalability of difficult cases may be important to researchers.

B. SCP-LimSim

SCP-LimSim leverages the capabilities of LimSim++ [18], an advanced platform that seamlessly integrates SUMO's robust traffic flow modeling [19] with CARLA's photorealistic rendering [4]. We employ Qwen-turbo [41] as agents with distinct driving characters for customizable planning. We assess SCP-LimSim's customizable driving ability in busy urban settings using the following behavioral metrics. We refer to CARLA leaderboard metrics and some from LimSim++ [18].

- **Completion Percentage**: The ratio of trips without deviations or collisions.
- **Driving Time**: Time taken for a vehicle to reach its target in seconds.
- **Comfort Score**: The comfort score during vehicle operation, calculated as $\frac{a_s+\dot{a}_s+a_d+\dot{a}_d}{4}$, with a_s as longitudinal acceleration of the vehicle, a_d as lateral acceleration and $\dot{a}$ as jerk. It can capture the effects of sudden speed changes on passengers.
- **Efficiency Score**: Calculated as $\frac{\bar{v}}{min(v_{limit},\bar{v}_{others})}$, with $\bar{v}$ as the average speed of the ego car over 10 frames, v_{limit} as the speed limit in this lane and $\bar{v}_{others}$ as the minimum value of the average speed of surrounding vehicles over 10 frames.
- **Safety Score**: Record the Time to Collision (TTC) between ego car and other vehicles when driving. The minimum record of TTC is the safety score.

The final score is determined by summing the comfort score, efficiency score, and safety score, with the penalties for red light violations, speed limit violations, and collisions. The resulting value reflects both the performance and adherence to traffic regulations. We tested different driving characters on three urban routes. It can be observed from the radar chart in Fig. 4 that the performance of each character in Route A, Route B, and Route C remains consistent, demonstrating the stability and controllability of our driver characters. However, due to the varying difficulty of each route, different degrees of distinction are shown in the three

TABLE II

RESULTS OF PLANNING MODELS ON SCP-NUPLAN. SNU REPRESENTS FOR SCP-NUPLAN.

Planner	Test14-random				Test14-hard			
	Original	SNuMini	SNuAir	SNu	Original	SNuMini	SNuAir	SNu
PlanCNN [1]	69.66	68.39	63.21	54.48	49.47	77.84	46.18	42.08
UrbanDriver [38]	63.27	62.10	61.00	51.26	51.54	68.19	45.59	44.22
PlanTF [39]	86.48	72.09	73.88	68.19	72.68	70.44	69.67	63.59

TABLE III

REPRODUCED RESULTS ON NUPLAN MINI SPLIT.

Planner	Test14-random		Test14-hard	
	Original	Reproduced	Original	Reproduced
PlanCNN [1]	69.66	79.92	49.47	82.38
UrbanDriver [38]	63.27	68.18	51.54	68.08
PlanTF [39]	86.48	85.11	72.68	74.91

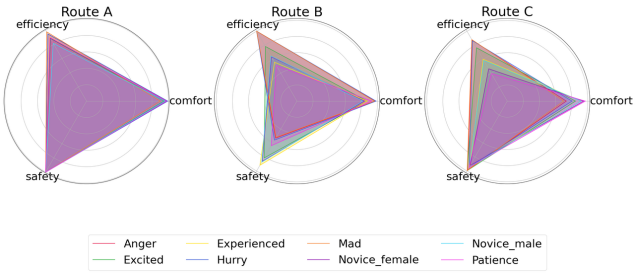

Fig. 4. **Performance variations are evident across distinct characters driven by Qwen-turbo on three urban routes.** The *Experienced* character consistently outperforms others across all routes. Conversely, aggressive characters such as *Mad*, *Hurry*, and *Excited* show poor safety and comfort performance but exceptional efficiency, indicating a higher tendency for risky driving and accidents. The moderate performance of *Novice_Female*, *Novice_Male*, and *Patient* characters suggests safer yet less efficient driving, likely due to their more cautious behavior.

radar charts. In the following analysis, we will focus on the performance of Route B.

The *Experienced* character consistently performs the best across all routes, indicating its proficiency in handling urban driving conditions and achieving a good balance between efficiency, safety, and comfort. Aggressive characters such as *Mad*, *Hurry*, and *Excited* tend to perform poorly in terms of safety and comfort but exhibit exceptional efficiency, suggesting a higher propensity for risky driving behaviors and accidents. The moderate performance of the *Novice_Female*, *Novice_Male*, and *Patient* characters indicates that they are generally safer but less efficient, likely due to more cautious driving behaviors. According to the research by Witt M et al. [42], *Novice_Female* tends to be more conservative, while *Novice_Male* tends to be more aggressive. This is also reflected in the radar charts, with *Novice_Male* showing higher efficiency and lower safety.

To illustrate the versatile capabilities of SCP-LimSim, we have created several multi-agent simulation corner cases, as depicted in 5. These scenarios are intended to demonstrate the customizable features of SCP's LLM-based agents,

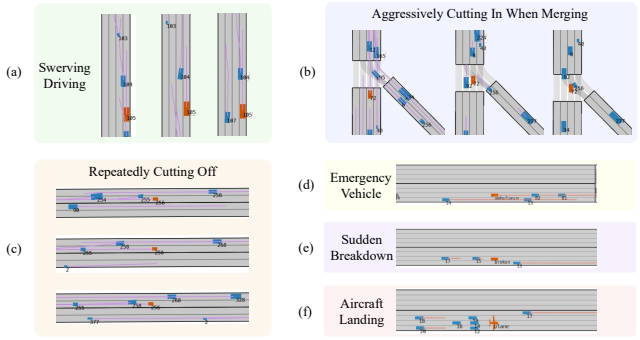

Fig. 5. **We build several multi-agent corner cases to demonstrate the flexibility of SCP in creating scenarios at will.** (a) Vehicle *104* is swerving on the road and affects other agents, which can be implemented by setting driver character or navigation information. (b) When the ego vehicle *72* merges the lanes, vehicles from the other lane are constantly cutting in aggressively. (c) There are vehicles constantly cutting off, hindering the movement of the ego vehicle *256*. (d) An emergency vehicle is calling for path clearance. (e) A sudden breakdown on the highway blocks other vehicles. (f) An aircraft needs an immediate landing on a busy highway, which is a rare but critical case for autonomous driving.

thereby assisting users in developing additional simulations. The scenarios include swerving driving, aggressively cutting in, emergency aircraft landing, *etc*. We achieved the construction of these scenarios only by modifying the driver characters or navigation information. More details including videos of these corner cases can be found on our website https://maple-zhou.github.io/SCP.

The aforementioned scenarios merely exemplify SCP's functionalities. We invite more users to join our community to design various scenarios using SCP's customizable agents, which exhibit diverse human-like driving styles. Additionally, SCP can host challenges and competitions to enrich the dataset of corner cases.

V. CONCLUSION

In conclusion, we introduce SCP, a scalable and customizable approach for generating planning-specific corner cases using Large Language Models. We present SCP-NuPlan, a scalable benchmark that transforms real-world datasets into challenging corner case scenarios, and SCP-LimSim, a customizable platform allowing users to simulate complex driving environments through natural language inputs. Together, these contributions push forward the development of more robust and adaptable planning algorithms, addressing the critical need for handling rare and unpredictable driving situations in autonomous systems.

REFERENCES

[1] H. Caesar, J. Kabzan, K. S. Tan, W. K. Fong, E. Wolff, A. Lang, L. Fletcher, O. Beijbom, and S. Omari, "nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles," *arXiv preprint arXiv:2106.11810*, 2021.

[2] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2446–2454.

[3] J. Houston, G. Zuidhof, L. Bergamini, Y. Ye, L. Chen, A. Jain, S. Omari, V. Iglovikov, and P. Ondruska, "One thousand and one hours: Self-driving motion prediction dataset," in *Conference on Robot Learning*. PMLR, 2021, pp. 409–418.

[4] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16.

[5] E. Leurent, "An environment for autonomous driving decision-making," https://github.com/eleurent/highway-env, 2018.

[6] X. Wang, Z. Zhu, G. Huang, X. Chen, and J. Lu, "Drivedreamer: Towards real-world-driven world models for autonomous driving," *arXiv preprint arXiv:2309.09777*, 2023.

[7] A. Hu, L. Russell, H. Yeo, Z. Murez, G. Fedoseev, A. Kendall, J. Shotton, and G. Corrado, "Gaia-1: A generative world model for autonomous driving," *arXiv preprint arXiv:2309.17080*, 2023.

[8] R. Gao, K. Chen, E. Xie, L. Hong, Z. Li, D.-Y. Yeung, and Q. Xu, "Magicdrive: Street view generation with diverse 3d geometry control," *arXiv preprint arXiv:2310.02601*, 2023.

[9] Z. Yang, Y. Chen, J. Wang, S. Manivasagam, W.-C. Ma, A. J. Yang, and R. Urtasun, "Unisim: A neural closed-loop sensor simulator," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1389–1399.

[10] Z. Wu, T. Liu, L. Luo, Z. Zhong, J. Chen, H. Xiao, C. Hou, H. Lou, Y. Chen, R. Yang, *et al.*, "Mars: An instance-aware, modular and realistic simulator for autonomous driving," in *CAAI International Conference on Artificial Intelligence*. Springer, 2023, pp. 3–15.

[11] Y. Wei, Z. Wang, Y. Lu, C. Xu, C. Liu, H. Zhao, S. Chen, and Y. Wang, "Editable scene simulation for autonomous driving via collaborative llm-agents," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 077–15 087.

[12] F. Codevilla, M. Müller, A. López, V. Koltun, and A. Dosovitskiy, "End-to-end driving via conditional imitation learning," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 4693–4700.

[13] A. E. Sallab, M. Saeed, O. A. Tawab, and M. Abdou, "Meta learning framework for automated driving," *arXiv preprint arXiv:1706.04038*, 2017.

[14] C. E. Tuncali, G. Fainekos, D. Prokhorov, H. Ito, and J. Kapinski, "Requirements-driven test generation for autonomous vehicles with machine learning components," *IEEE Transactions on Intelligent Vehicles*, vol. 5, no. 2, pp. 265–280, 2019.

[15] G. Chou, Y. E. Sahin, L. Yang, K. J. Rutledge, P. Nilsson, and N. Ozay, "Using control synthesis to generate corner cases: A case study on autonomous driving," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, no. 11, pp. 2906–2917, 2018.

[16] F. Moller, D. Botache, D. Huseljic, F. Heidecker, M. Bieshaar, and B. Sick, "Out-of-distribution detection and generation using soft brownian offset sampling and autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 46–55.

[17] L. Wenl, D. Fu, S. Mao, P. Cai, M. Dou, Y. Li, and Y. Qiao, "Limsim: A long-term interactive multi-scenario traffic simulator," in *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2023, pp. 1255–1262.

[18] D. Fu, W. Lei, L. Wen, P. Cai, S. Mao, M. Dou, B. Shi, and Y. Qiao, "Limsim++: A closed-loop platform for deploying multimodal llms in autonomous driving," *arXiv preprint arXiv:2402.01246*, 2024.

[19] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.-P. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner, and E. Wießner, "Microscopic traffic simulation using sumo," in *2018 21st international conference on intelligent transportation systems (ITSC)*. IEEE, 2018, pp. 2575–2582.

[20] T. Zhang, H. Liu, W. Wang, and X. Wang, "Virtual tools for testing autonomous driving: A survey and benchmark of simulators, datasets, and competitions," *Electronics*, vol. 13, no. 17, p. 3486, 2024.

[21] Y. Li, W. Yuan, S. Zhang, W. Yan, Q. Shen, C. Wang, and M. Yang, "Choose your simulator wisely: A review on open-source simulators for autonomous driving," *IEEE Transactions on Intelligent Vehicles*, 2024.

[22] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "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.

[23] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and Service Robotics: Results of the 11th International Conference*. Springer, 2018, pp. 621–635.

[24] G. Rong, B. H. Shin, H. Tabatabaee, Q. Lu, S. Lemke, M. Možeiko, E. Boise, G. Uhm, M. Gerow, S. Mehta, *et al.*, "Lgsvl simulator: A high fidelity simulator for autonomous driving," in *2020 IEEE 23rd International conference on intelligent transportation systems (ITSC)*. IEEE, 2020, pp. 1–6.

[25] "Baidu apollo," https://apollo.baidu.com/, accessed: 2024-06-02.

[26] C. Gulino, J. Fu, W. Luo, G. Tucker, E. Bronstein, Y. Lu, J. Harb, X. Pan, Y. Wang, X. Chen, *et al.*, "Waymax: An accelerated, data-driven simulator for large-scale autonomous driving research," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[27] Q. Li, Z. Peng, L. Feng, Q. Zhang, Z. Xue, and B. Zhou, "Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 3, pp. 3461–3475, 2022.

[28] M. Zhou, J. Luo, J. Villella, Y. Yang, D. Rusu, J. Miao, W. Zhang, M. Alban, I. Fadakar, Z. Chen, *et al.*, "Smarts: An open-source scalable multi-agent rl training school for autonomous driving," in *Conference on robot learning*. PMLR, 2021, pp. 264–285.

[29] (2024, Mar.) Simcenter prescan software. Siemens. [online] Available: https://plm.sw.siemens.com/en-US/simcenter/autonomous-vehicle-solutions/prescan.

[30] D. Bogdoll, J. Breitenstein, F. Heidecker, M. Bieshaar, B. Sick, T. Fingscheidt, and M. Zöllner, "Description of corner cases in automated driving: Goals and challenges," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1023–1028.

[31] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[32] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan, "Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 5855–5864.

[33] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, "Mip-nerf 360: Unbounded anti-aliased neural radiance fields," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5470–5479.

[34] P. Wang, Y. Liu, Z. Chen, L. Liu, Z. Liu, T. Komura, C. Theobalt, and W. Wang, "F2-nerf: Fast neural radiance field training with free camera trajectories," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4150–4159.

[35] A. Pretschner, F. Hauer, and T. Schmidt, "Tests für automatisierte und autonome fahrsysteme: Wiederverwendung aufgezeichneter fahrten ist nicht zu rechtfertigen," *Informatik Spektrum*, vol. 44, no. 3, pp. 214–218, 2021.

[36] F. Klück, Y. Li, M. Nica, J. Tao, and F. Wotawa, "Using ontologies for test suites generation for automated and autonomous driving functions," in *2018 IEEE International symposium on software reliability engineering workshops (ISSREW)*. IEEE, 2018, pp. 118–123.

[37] A. Kesting and M. Treiber, "Traffic flow dynamics: data, models and simulation," *no. Book, Whole)(Springer Berlin Heidelberg, Berlin, Heidelberg, 2013)*, 2013, see Chapter 11, Section 3.

[38] O. Scheel, L. Bergamini, M. Wolczyk, B. Osiński, and P. Ondruska, "Urban driver: Learning to drive from real-world demonstrations using policy gradients," in *Conference on Robot Learning*. PMLR, 2022, pp. 718–728.

[39] J. Cheng, Y. Chen, X. Mei, B. Yang, B. Li, and M. Liu, "Rethinking imitation-based planner for autonomous driving," 2023.

[40] D. Dauner, M. Hallgarten, A. Geiger, and K. Chitta, "Parting with misconceptions about learning-based vehicle motion planning," in *Conference on Robot Learning*. PMLR, 2023, pp. 1268–1281.

[41] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[42] M. Witt, K. Kompaß, L. Wang, R. Kates, M. Mai, and G. Prokop, "Driver profiling–data-based identification of driver behavior dimensions and affecting driver characteristics for multi-agent traffic simulation," *Transportation research part F: traffic psychology and behaviour*, vol. 64, pp. 361–376, 2019.