NLG-Gen: Natural Language-Guided Generation of Long-Tail Critical Scenarios for Autonomous Driving

Tingting Lei[1], Yifan Zhu[2], Runxi Zhang[3], Feng Hu[1], Hong Yu[1], and Ye Wang[1]

[1] Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[2] Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
[3] Tongji University, Shanghai 200092, China
leitingting312@gmail.com, yifanzhu946@gmail.com,
zhangrunxi@tongji.edu.cn,
hufeng@cqupt.edu.cn, yuhong@cqupt.edu.cn, wangye@cqupt.edu.cn

Abstract. Long-tail critical scenarios are essential for evaluating the safety boundaries of autonomous driving systems, yet their rarity in real-world data makes them difficult to generate controllably and reliably. Existing scenario generation methods often struggle to impose fine-grained semantic control over hazardous interactions while preserving the intended risk semantics during generation. To address these limitations, we propose **NLG-Gen** (**N**atural **L**anguage-**G**uided Generation), a framework for long-tail critical scenario generation in autonomous driving. The framework first parses natural language instructions into structured scenario constraints, and then explicitly injects target hazard semantics by editing key entities and their interactions in vectorized scenario space. To improve realism while maintaining the edited hazard semantics, we introduce a constraint-preserving diffusion inpainting module that repairs locally unnatural regions under semantic protection. Experiments on three representative long-tail critical scenarios, including pedestrian crossing, hard braking of the lead vehicle, and forced cut-in, demonstrate that the proposed method achieves strong semantic alignment, high scenario feasibility, and good structural quality. Closed-loop evaluations on the nuPlan benchmark across three mainstream planners show that the generated scenarios more effectively challenge the planners and reveal their safety weaknesses. These results indicate that the proposed framework provides an effective and controllable paradigm for long-tail critical scenario generation in autonomous driving safety evaluation.

Keywords: Autonomous Driving · Scenario Generation · Natural Language Guidance · Long-Tail Critical Scenarios · Safety Evaluation

1 Introduction

Safety evaluation is a fundamental prerequisite for the deployment of autonomous driving systems. Although current systems perform well in routine traffic, their safety boundaries are often challenged by rare but high-risk events, such as

sudden pedestrian crossing, abrupt braking of a lead vehicle, and forced cut-in maneuvers [6]. These long-tail critical scenarios are scarce in real-world data, yet they are essential for exposing the weaknesses of planning systems [3]. Therefore, efficiently constructing such scenarios has become an important problem in autonomous driving research.

Existing studies on safety-critical scenario generation mainly fall into three categories: rule-based construction, adversarial search, and data-driven generative modeling. Rule-based methods are interpretable but often inefficient and limited in diversity [8]. Adversarial methods can discover challenging interactions by optimizing surrounding agents, but they may produce behaviors that deviate from realistic traffic patterns [7]. Data-driven generative models, especially recent diffusion-based frameworks such as SLEDGE [2], have shown strong capability in synthesizing realistic driving scenarios [11]. However, when applied to long-tail safety evaluation, these methods still struggle to reliably generate rare hazardous interactions under fine-grained semantic control.

Existing generative simulation methods still face three main limitations in autonomous driving safety evaluation. First, the inherent long-tail distribution of real-world data causes generative models to favor common traffic scenarios, making it difficult to stably generate rare but safety-critical interactions [9]. Second, existing conditional guidance is often too coarse-grained to enforce fine-grained constraints on critical participants and conflict relations [14]. Third, because diffusion models are strongly regularized by the prior of the learned real-world distribution, even explicitly injected hazard semantics may be gradually weakened during denoising, causing the final generated results to drift away from the intended control target [13].

To address these limitations, we propose **NLG-Gen** (**N**atural **L**anguage-**G**uided Generation), a framework for long-tail critical scenario generation in autonomous driving. Building upon SLEDGE [2], the framework first converts natural language instructions into structured semantic constraints, then explicitly edits key entities and interaction relationships in vectorized scenario space to inject target hazard semantics. To improve realism while preserving these semantics, we further introduce a constraint-preserving diffusion inpainting mechanism that repairs locally unnatural regions under semantic protection. In this way, the proposed method balances semantic controllability and scenario realism without modifying the original diffusion backbone. The main contributions of this paper are summarized as follows:

(1) We propose a natural language-guided framework for generating long-tail critical scenarios in autonomous driving, which transforms language instructions into structured constraints for fine-grained semantic control.
(2) We develop a semantic editing and constraint-preserving diffusion inpainting pipeline to explicitly inject target hazard semantics while maintaining scenario realism and structural plausibility.
(3) We conduct extensive experiments on representative critical scenarios and closed-loop simulation, showing that the proposed method can generate controllable and challenging scenarios for autonomous driving safety evaluation.

2 Related Work

Autonomous driving scenario generation is closely tied to the underlying scenario representation. Early studies often adopted rasterized representations, which are intuitive for grid-based modeling but may introduce substantial redundancy when describing structured traffic scenes [10]. In contrast, vectorized representations have become the mainstream choice because they provide a more compact and structured description of lane geometry and agent states [10]. Building on this idea, SLEDGE further introduces a raster-to-vector autoencoding framework and latent diffusion generation pipeline, enabling realistic and scalable closed-loop scenario synthesis [2]. However, although latent generative frameworks effectively model overall traffic distributions, their abstract latent space still makes fine-grained semantic control over critical interactions difficult.

Existing approaches to safety-critical scenario generation can be broadly grouped into rule-based methods, adversarial search, and data-driven generative models. Rule-based methods are interpretable but often limited in diversity and scalability [8]. Adversarial methods can expose planner weaknesses by optimizing surrounding-agent behaviors, but they may produce unrealistic interactions [7]. Data-driven generative models, especially diffusion-based frameworks such as SLEDGE [2], have shown strong potential for synthesizing realistic driving scenarios [11]. Yet, their limitations become evident when the goal is to generate rare critical scenarios with both realism and precise semantic control.

Recently, diffusion models have shown potential for scenario synthesis due to strong distribution modeling and generation fidelity. Several studies have explored controllable generation through conditional guidance, guided sampling, or local inpainting [5, 12]. Nevertheless, these methods remain limited for long-tail critical scenario generation. Rare hazardous interactions are severely underrepresented in real-world data, causing pre-trained models to favor common traffic patterns. Furthermore, existing guidance mechanisms are soft and coarse-grained, making it difficult to explicitly preserve key hazard semantics and conflict relations. These limitations necessitate a more explicit and stable control paradigm for long-tail critical scenario generation.

While natural language offers an intuitive interface for scenario generation, translating abstract prompts into precise spatial-temporal interactions remains challenging [1]. Existing text-conditioned frameworks rely on cross-attention mechanisms that provide loose semantic alignment. This soft guidance is inadequate for safety-critical evaluations, where minor kinematic deviations drastically alter the risk profile. Consequently, language instructions must be transformed into explicit structural constraints rather than implicit network mappings [16, 19].

Overall, existing methods provide useful foundations for autonomous driving scenario generation, yet they remain insufficient for long-tail critical scenario generation that requires explicit hazard injection, stable semantic preservation, and realistic scenario restoration. This observation motivates the method proposed in this paper.

3 Method

3.1 Problem Formulation and Overview

Our goal is to generate long-tail critical driving scenarios from natural language instructions while preserving both hazard semantics and scenario realism. Following SLEDGE [2], a local driving scene is represented in an ego-centric coordinate system as a structured scenario

$$S = \{M, A\}, \tag{1}$$

where M denotes the vectorized road topology and A denotes the set of traffic participants. Specifically, M includes structured map elements such as lane centerlines, road boundaries, and crosswalks, while each agent in A is described by its geometric and kinematic attributes, including position, heading, size, velocity, and category. This representation provides a compact and structured basis for scenario modeling and editing [10].

Following the latent generative pipeline of SLEDGE [2], the vectorized scenario is rasterized into a multi-channel bird's-eye-view tensor and encoded into a latent feature representation through a raster-to-vector autoencoder. Scenario generation is then performed in the latent space using a diffusion model. Given the latent representation z_0 of a scenario, the noisy latent at step t is defined as

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \tag{2}$$

where $\epsilon \sim \mathcal{N}(0, I)$. The reverse denoising process progressively removes the injected noise to recover a plausible latent scenario representation.

Built upon this backbone, we propose **NLG-Gen**, a natural language-guided framework for controllable long-tail critical scenario generation. As illustrated in Fig. 1, NLG-Gen consists of three tightly coupled stages: natural language intent parsing, semantic editing in vectorized scenario space, and constraint-preserving diffusion inpainting. The framework first converts language instructions into structured semantic constraints, then explicitly edits key entities and interaction relations to inject target hazard semantics, and finally restores local realism through constrained diffusion inpainting while preserving the edited semantics. During generation, the original and edited scenarios are both mapped into rasterized and latent representations, from which differential and semantic protection masks are constructed. Multiple restoration candidates are then generated from different low-noise starting points and random seeds, and the final output is selected through semantic alignment, compliance, and fidelity-based filtering. In this way, NLG-Gen unifies explicit semantic control and realistic scenario generation within a single pipeline.

3.2 Natural Language Intent Parsing

The intent parsing module converts natural language instructions into structured control conditions that can directly guide scenario editing. Given an input text

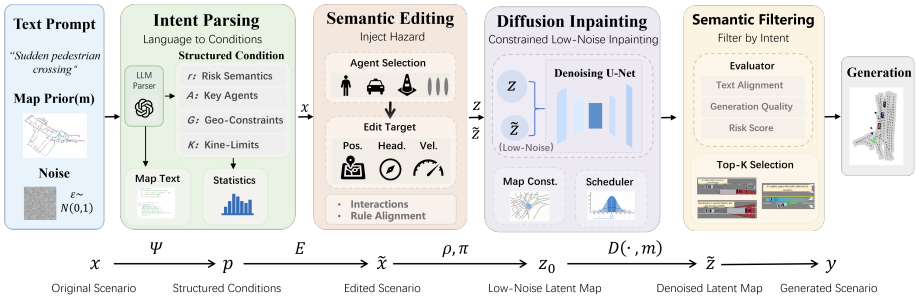


Fig. 1. Overview of the proposed NLG-Gen framework. Given a natural language instruction and an original driving scenario, NLG-Gen first performs intent parsing to obtain structured semantic constraints, then explicitly edits key entities and interaction relations in vectorized scenario space, and finally restores local realism through constraint-preserving diffusion inpainting with mask-guided low-noise sampling and candidate filtering.

$t \in T$, the parser outputs a structured representation

$$p = \Psi(t) = \{c, r, A, G, K\}, \tag{3}$$

where c denotes the scenario category, r denotes the target hazard semantics, A denotes the key participants, G denotes the geometric relation constraints, and K denotes the risk level.

In this work, the scenario categories mainly include pedestrian crossing, hard braking of the lead vehicle, and forced cut-in, while the hazard intensity is divided into three levels: mild, moderate, and aggressive. Different categories determine the edited entities and interaction patterns, while different risk levels influence parameters such as forward distance, lateral intrusion extent, speed settings, and conflict urgency. Thus, this module serves as the interface that maps natural language descriptions into executable scenario constraints.

The parsing module also supports batch scenario construction. For cached source scenarios, target scenario types and risk levels can be automatically assigned, after which corresponding language prompts are generated and parsed into unified structured constraints. This design supports both interactive control for single-scenario editing and automated construction of large-scale long-tail critical scenario sets.

3.3 Semantic Editing in Vectorized Scenario Space

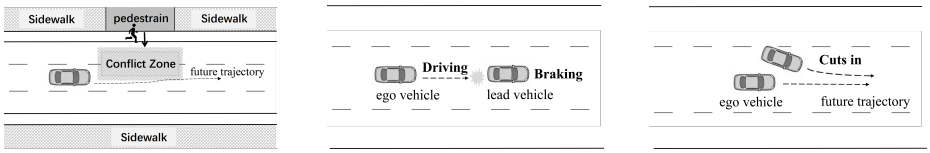
The semantic editing module is the core component for explicit hazard injection. Given an original scenario x and structured constraints p, it directly modifies the attributes, kinematics, and relative relations of key traffic participants in vectorized scenario space to obtain an edited scenario

$$\tilde{x} = E(x, p). \tag{4}$$

Unlike pixel-level editing, NLG-Gen operates on structured traffic entities, such as pedestrians, vehicles, occlusions, and their geometric relations [15]. This enables target hazard semantics to be explicitly written into the scenario before generation, rather than relying on the generative model to implicitly produce the desired interaction.

In addition to the edited scenario, the module outputs an editing report that records key participants, modified regions, and semantically critical regions to be protected in the subsequent restoration stage. Therefore, this module acts as a hard semantic injection stage, while the following diffusion inpainting module further improves realism under semantic constraints.

Scenario-Specific Editing Strategies To make the injected hazard semantics executable and stable in closed-loop simulation, NLG-Gen adopts scenario-specific editing strategies for different long-tail critical cases, as shown in Fig. 2.



(a) Pedestrian Crossing (b) Hard Braking (c) Forced Cut-in

Fig. 2. Examples of three long-tail critical scenarios.

Pedestrian crossing. For pedestrian crossing scenarios, the goal is to construct a lateral conflict between the pedestrian and the future path of the ego vehicle. Let L_{ego} denote the future driving region of the ego vehicle and R_{ped} denote the candidate region of the pedestrian. The editing objective is to place or relocate the pedestrian near a plausible roadside or crossing corridor while satisfying

$$R_{\text{ped}} \cap L_{\text{ego}} \neq \emptyset. \tag{5}$$

This ensures that the initial state already contains recognizable crossing-related hazard semantics. In practice, the edited pedestrian is constrained not only by location but also by movement direction and conflict tendency with the ego trajectory. Since pedestrian behavior in the simulator is relatively simple, once this crossing relation is explicitly established in the initial state, the corresponding hazard semantics can usually be maintained during subsequent simulation.

Hard braking of the lead vehicle. For hard-brake scenarios, NLG-Gen constructs a state proxy in the form of a short-headway, low-speed lead vehicle ahead of the ego vehicle. Let the set of candidate lead vehicles be

$$V_{\text{lead}} = \{v_i \mid x_i > 0,\ |y_i - y_{\text{lane}}| < \delta\}, \tag{6}$$

where x_i is the longitudinal distance from the ego vehicle, y_i is the lateral position of the vehicle, y_{lane} is the ego-lane center, and δ is a lane-width-related threshold. The editor selects or constructs a lead vehicle in the ego lane and adjusts its forward distance, lateral offset, and speed to establish a strong longitudinal braking conflict. The core of this scenario is not to simulate the full braking process, but to construct an initial state with clear deceleration pressure. Because SLEDGE uses lane-following and IDM-based longitudinal control for non-ego vehicles [18], such a state proxy can effectively persist during closed-loop simulation.

Forced cut-in. Forced cut-in is the most challenging case under the current simulator mechanism, because surrounding vehicles are gradually regularized toward nearby lane centerlines during rollout. As a result, directly placing a neighboring vehicle with a lane-changing tendency may not stably preserve cut-in semantics. To address this issue, NLG-Gen adopts a state-proxy strategy that emphasizes the hazardous cut-in state rather than the full continuous lane-change process. Specifically, the target vehicle is placed ahead of the ego vehicle, close to the ego-lane boundary, or partially intruding into the ego lane, such that

$$x_c > 0, \ |y_c - y_{\text{lane}}| < w_{\text{lane}}, \tag{7}$$

where w_{lane} denotes the lane width. Different severity levels are defined by forward distance, lateral intrusion depth, heading deviation, and pseudo-TTC relative to the ego vehicle. This design better matches the simulator constraints and improves the stability of cut-in semantics during closed-loop evaluation.

Overall, the role of semantic editing is to construct an intermediate hazardous scenario state with explicit semantics, plausible geometry, and readiness for subsequent restoration. It explicitly writes the target hazard interaction into the scenario, upon which diffusion inpainting further improves naturalness and distribution consistency.

3.4 Constraint-Preserving Diffusion Inpainting

Although semantic editing can explicitly inject target hazard semantics, rule-based modifications often disturb local realism and produce unnatural layouts. To address this issue, NLG-Gen employs a constraint-preserving diffusion inpainting module to restore edited scenarios under semantic protection. Instead of generating scenarios from pure Gaussian noise, it starts from the edited latent representation and performs low-noise restoration, allowing the diffusion model to act as a targeted restorer rather than a from-scratch generator.

Differential Region and Semantic Preservation Masks To reduce the risk that the diffusion prior weakens the edited hazard semantics, we construct two complementary protection masks. The first is a differential mask M_{diff}, which is obtained by comparing the rasterized original scenario and the rasterized edited scenario to identify the regions that have been explicitly modified. The second is a semantic region-of-interest mask M_{roi}, which is derived from the editing

report and marks semantically critical regions, such as key participants, conflict corridors, and ego-lane boundary anchors.

The final protection mask is defined as

$$M_{\text{mask}} = \max(M_{\text{diff}}, M_{\text{roi}}), \tag{8}$$

where the point-wise maximum merges the edited and semantically important regions into a unified constraint map. This mask is then used to restrict undesired changes in critical regions during the restoration process.

Low-Noise Inpainting Let z_e denote the latent representation of the edited scenario. Instead of starting reverse denoising from the maximum noise level, NLG-Gen performs restoration from a low-noise step t_s, so that the edited hazard semantics can be preserved as much as possible while still allowing the diffusion model to repair locally unnatural regions. A single restoration attempt is formulated as

$$\hat{z}^{(k)} = D(z_e, t_s^{(k)}, M_{\text{mask}}), \tag{9}$$

where $D(\cdot)$ denotes the constrained reverse denoising process and $t_s^{(k)}$ denotes the starting step of the k-th restoration attempt.

In practice, multiple low-noise starting points and random seeds are used to generate candidate scenarios. A lower starting noise level is more favorable for retaining edited hazard semantics, while multi-start restoration increases the probability of obtaining candidates that are both semantically valid and structurally realistic. In this way, the diffusion model serves as a constrained local restorer that improves realism without overriding the injected critical semantics.

Candidate Filtering Because diffusion inpainting cannot guarantee semantic validity for every restored attempt, NLG-Gen further applies semantic evaluation and compliance checking to all generated candidates. Let $\hat{x}^{(k)}$ denote the k-th restored scenario, and let $A^{(k)}$ and A_e denote the semantic alignment scores of the restored and edited scenarios. The semantic fidelity ratio is defined as

$$r^{(k)} = \frac{A^{(k)}}{A_e}. \tag{10}$$

A higher value of $r^{(k)}$ indicates that the restored scenario preserves more of the originally injected hazard semantics.

In addition to semantic fidelity, each candidate is checked for structural and numerical compliance, such as valid state values and simulatable scenario structure. Only candidates that satisfy both semantic and compliance requirements are retained for ranking. To select the final output, we define the ranking score as

$$R^{(k)} = 50 \cdot 1(\text{semantic\_ok}^{(k)}) + 30 \cdot 1(\text{compliance\_ok}^{(k)}) + 10 \cdot r^{(k)} + A^{(k)}, \tag{11}$$

where $1(\cdot)$ is the indicator function. This design prioritizes semantic validity and scenario feasibility, while further favoring candidates with stronger semantic preservation and higher alignment quality.

Overall, the diffusion inpainting module complements semantic editing by restoring local realism under semantic constraints, enabling NLG-Gen to balance hazard controllability and scenario naturalness in a unified generation pipeline.

4 Experiments

4.1 Datasets and Experimental Setup

Experiments are conducted on the public nuPlan dataset using the vectorized scenario representation and closed-loop simulation pipeline of SLEDGE. nuPlan contains approximately 1,300 hours of real-world driving data collected from Las Vegas, Boston, Pittsburgh, and Singapore. Following the standard SLEDGE setting, each scenario is represented within a $64\,\mathrm{m} \times 64\,\mathrm{m}$ ego-centric local field and converted into a unified vectorized format including lane centerlines, road boundaries, crosswalks, and dynamic traffic participants. Since Boston better matches the right-hand traffic setting considered in this work, we primarily use the Boston subset for scenario construction and evaluation.

We evaluate NLG-Gen on three representative long-tail critical scenarios: pedestrian crossing (*Ped-Crossing*), hard braking of the lead vehicle (*Hard-Brake*), and forced cut-in (*Cut-in*). For each scenario type, three hazard severity levels are considered: *Mild*, *Moderate*, and *Aggressive*. Given an original scenario and a natural language instruction, the framework performs intent parsing, semantic editing, and constraint-preserving diffusion inpainting to generate the final critical scenario.

To further analyze the contribution of different stages in the pipeline, we compare three variants: *None*, which denotes the original scenario without semantic control; *+edited*, which denotes the semantically edited scenario before restoration; and *+repaint*, which denotes the final restored scenario after diffusion inpainting.

4.2 Evaluation Metrics

We evaluate the proposed framework from three complementary perspectives: closed-loop planner challenge, generation quality and controllability, and stage-wise diagnostic analysis. Specifically, we report Mean Semantic Alignment Score (MPA), Compliance Rate (CR), Scenario Quality Score (SQS), Drivable Route Length (DRL), High-Risk Interaction Score (Inter.), Contextual Change (CTC), Repaint Success Rate (RSR), and the standard nuPlan closed-loop metrics. Except for DRL, all metrics are reported in percentage.

MPA measures how well the generated scenario matches the target hazard semantics. CR evaluates whether the generated scenario satisfies basic geometric, structural, and interaction constraints. SQS reflects the overall plausibility

of the scenario in terms of layout consistency and traffic quality. Inter. measures whether the target hazardous interaction is effectively established. RSR evaluates the success of diffusion inpainting in preserving semantics while improving local quality. Finally, nuPlan closed-loop metrics are used to assess whether the generated scenarios truly increase the challenge for planners in closed-loop simulation.

For the i-th scenario, the drivable route length is computed by first locating the closest current lane of the ego vehicle and then recursively accumulating the reachable lane length L^i_{reach} along its successor lanes. To avoid extreme values, an upper bound $L_{\max}$ is applied:

$$\text{DRL}_i = \min\left(L^i_{\text{reach}}, L_{\max}\right). \tag{12}$$

The overall average DRL reflects how much usable road structure is retained ahead of the ego vehicle.

To quantify how much editing and restoration perturb the surrounding context outside the target region, we define the Contextual Change (CTC) metric. For the i-th sample, let the target editing region be denoted as Ω_i, and let the expanded contextual annulus be

$$\Omega^e_i = \text{Expand}(\Omega_i, m) \setminus \Omega_i, \tag{13}$$

where m is the context expansion boundary. For each contextual object category c, the position sets in the original and target scenarios are denoted as $P^{\text{orig}}_{i,c}$ and $P^{\text{tar}}_{i,c}$, respectively. Their bidirectional Chamfer distance is defined as

$$D_{\text{ch}}(P,Q) = \frac{1}{2}\left(\frac{1}{|P|}\sum_{p\in P}\min_{q\in Q}\|p-q\|_2 + \frac{1}{|Q|}\sum_{q\in Q}\min_{p\in P}\|p-q\|_2\right). \tag{14}$$

Based on this distance, the position change term $S_{i,c}$ is obtained by normalized truncation, and the quantity change term $N_{i,c}$ is obtained from the relative difference in object counts. The contextual change for category c is then defined as

$$C_{i,c} = 0.7S_{i,c} + 0.3N_{i,c}. \tag{15}$$

Finally, the single-sample contextual change metric is given by

$$\text{CTC}_i = \frac{1}{|c_i|}\sum_{c\in C_i} C_{i,c}, \tag{16}$$

where c_i denotes the set of valid object categories in the i-th sample. A larger CTC indicates a greater disturbance to the surrounding context.

4.3 Closed-loop Challenge on the nuPlan Benchmark

To verify whether the generated scenarios truly increase the challenge for autonomous driving systems, Table 1 reports the closed-loop results on the nuPlan

benchmark across three planners, namely PDM-Closed, Diffusion Planner, and Flow Planner [4, 20, 17]. Here, nuPlan denotes the official validation split of the nuPlan benchmark, which is commonly used to assess the general closed-loop planning performance of autonomous driving systems.

Table 1. Closed-loop Safety Evaluation on nuPlan Benchmark.

Planner	Dataset	Collision-free ↑	Off-road comp. ↑	Direction comp. ↑	Progress success ↑	Route progress ↑	TTC ↑	Score ↑
PDM-Closed CoRL 2023	nuPlan	100.00	100.00	100.00	100.00	99.03	100.00	97.61
	Ours	88.75	99.40	93.15	92.50	77.81	76.50	27.86
Diffusion Planner ICLR 2025	nuPlan	100.00	100.00	100.00	100.00	89.59	100.00	95.70
	Ours	69.78	53.41	86.95	86.04	72.59	63.45	18.59
Flow Planner NeurIPS 2025	nuPlan	100.00	100.00	100.00	100.00	90.89	100.00	97.13
	Ours	58.81	34.94	85.10	87.51	77.55	51.86	10.05

Note: All metric values are reported in percentage (%).

As shown in Table 1, the scenarios generated by NLG-Gen consistently impose substantial challenge across all three planners. Compared with the official validation split, all planners experience dramatic performance degradation on our generated scenario set. Specifically, the closed-loop score drops from 97.61 to 27.86 for PDM-Closed, from 95.70 to 18.59 for Diffusion Planner, and from 97.13 to 10.05 for Flow Planner. Clear declines are also observed in several key safety- and task-related metrics, including collision-free rate, progress success, route progress, and TTC. For example, the collision-free rate decreases to 88.75, 69.78, and 58.81 for the three planners, respectively, while TTC falls to 76.50, 63.45, and 51.86. These results indicate that the generated scenarios are substantially more challenging than the standard benchmark split and can more effectively expose planner weaknesses under closed-loop evaluation.

The consistent degradation across different planner architectures also suggests that the challenge introduced by NLG-Gen is not tied to a single planner design, but reflects the intrinsic difficulty of the generated critical scenarios. This result directly supports the practical value of the proposed framework for autonomous driving safety evaluation.

4.4 Generation Quality Evaluation

After validating the closed-loop challenge of the generated scenarios, we further analyze whether NLG-Gen maintains stable generation quality across different categories of long-tail critical scenarios and different hazard intensity levels.

Performance across Scenario Types Table 2 reports the performance of NLG-Gen on three representative long-tail critical scenario categories, namely Ped-Crossing, Hard-Brake, and Cut-in. The results show that NLG-Gen achieves

Table 2. Performance across Scenario Types.

Type	MPA (%)	CR (%)	SQS (%)	DRL (m)
Ped-Crossing	92.66	100	75.51	35.98
Hard-Brake	86.93	100	73.64	40.09
Cut-in	83.48	100	75.57	43.57

consistently strong performance across all scenario types. In particular, the compliance rate remains 100% in all cases, indicating that the generated scenarios satisfy the required geometric and structural constraints.

In terms of semantic alignment, Ped-Crossing achieves the highest MPA of 92.66, followed by Hard-Brake at 86.93 and Cut-in at 83.48. This suggests that pedestrian crossing conflicts are relatively easier to inject and preserve, while Hard-Brake and especially Cut-in involve more complex interaction patterns. Nevertheless, all three scenario types maintain competitive scenario quality, with SQS values ranging from 73.64 to 75.57, and preserve reasonable drivable structure, with DRL values between 35.98 m and 43.57 m. These results demonstrate that NLG-Gen generalizes well across different types of long-tail critical scenarios without sacrificing plausibility.

Fine-grained Control across Severity Levels We further examine whether NLG-Gen supports fine-grained control over hazard intensity. Table 3 summarizes the results under three severity levels, namely Mild, Moderate, and Aggressive. As severity increases from Mild to Aggressive, the MPA improves from 88.42 to 94.42, while SQS rises from 73.69 to 77.45. At the same time, CR remains 100% for all settings, indicating that stronger hazard injection does not compromise structural validity. DRL also increases from 36.47 m to 40.83 m, suggesting that the generated scenarios still preserve usable road structure under more aggressive configurations.

Table 3. Performance across Severity Levels.

Severity	MPA (%)	CR (%)	SQS (%)	DRL (m)
Mild	88.42	100	73.69	36.47
Moderate	93.93	100	76.37	38.69
Aggressive	94.42	100	77.45	40.83

This trend indicates that NLG-Gen does not merely generate critical scenarios in a binary manner, but also provides controllable adjustment of hazard intensity. More aggressive settings introduce more explicit and recognizable conflict relations, which are more easily captured by semantic evaluation metrics.

Therefore, the proposed framework supports fine-grained language-guided scenario generation rather than only coarse category-level control.

4.5 Ablation Study

Table 4. Evaluation of Each Stage of NLG-Gen.

Model	MPA (%)	SQS (%)	Inter. (%)	CTC (/m)	RSR (%)
None	29.04	76.82	50.87	–	–
+edited	89.49	80.92	63.67	3.87	–
+repaint	91.65	75.37	69.91	41.03	91.12

To analyze the contribution of each stage in the pipeline, we compare three settings: the original scenario without semantic control (*None*), the semantically edited scenario (*+edited*), and the restored scenario after constraint-preserving diffusion inpainting (*+repaint*). As shown in Table 4, semantic editing is the main source of hazard injection. Compared with the uncontrolled baseline (*None*), the *+edited* stage markedly improves MPA from 29.04 to 89.49 and Inter. from 50.87 to 63.67, while also raising SQS from 76.82 to 80.92. This shows that explicit editing can effectively establish the target hazardous interaction. The *+repaint* stage further improves MPA to 91.65 and Inter. to 69.91, with an RSR of 91.12%. Although SQS drops slightly to 75.37 and CTC increases from 3.87 to 41.03, the results indicate that constrained repainting enhances semantic preservation and interaction strength, at the cost of broader contextual adjustment.

5 Conclusion and Limitations

This paper proposed **NLG-Gen**, a natural language-guided framework for long-tail critical scenario generation in autonomous driving. By combining semantic editing in vectorized scenario space with constraint-preserving diffusion inpainting, NLG-Gen enables explicit hazard semantic injection while maintaining scenario realism. Experiments on three representative scenarios, namely Ped-Crossing, Hard-Brake, and Cut-in, show that NLG-Gen can generate semantically aligned, structurally feasible, and high-quality critical scenarios under fine-grained language control. Stage-wise analysis further confirms that semantic editing is the key source of hazard injection, while diffusion inpainting improves local realism and interaction consistency under semantic constraints. Closed-loop evaluation on the nuPlan benchmark across three representative planners further demonstrates that the generated scenarios impose a greater challenge on the planners and more effectively expose their safety weaknesses. Overall, NLG-Gen provides an effective and practical solution for autonomous driving safety evaluation through controllable long-tail critical scenario generation.

Despite these promising results, several limitations remain. First, the current implementation adopts scenario-specific editing strategies for the three studied critical cases. This design is intentional because safety-critical scenario generation requires explicit and verifiable control over key conflict relations. However, it also limits the current framework to predefined scenario categories, and extending NLG-Gen to more open-ended long-tail cases requires a more general library of editable semantic primitives. Second, a trade-off still exists between explicit semantic injection and overall scenario naturalness, since strong editing may perturb local geometric context. Third, the constraint mechanisms used to prevent semantic drift can restrict generation diversity. Fourth, the current evaluation protocol still relies partly on proxy-style metrics and does not fully cover realism, diversity, or long-term temporal consistency. Finally, the generalization of the framework to more complex multi-agent scenarios, richer critical scenario types, and more open-ended language instructions remains to be further validated. Future work will focus on better balancing controllability and realism, improving evaluation metrics, developing more general semantic editing primitives, and extending the framework to more flexible natural language-guided critical scenario generation.

Acknowledgments This work was partly supported by the National Natural Science Foundation of China (62136002, 62306056, and 62221005), the Natural Science Foundation of Chongqing (CSTB2023NSCQ-LZX0006), respectively.

References

1. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: CVPR (2023)
2. Chitta, K., Dauner, D., Geiger, A.: Sledge: Synthesizing driving environments with generative models and rule-based traffic. In: European Conference on Computer Vision. pp. 57–74 (2024)
3. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9268–9277 (2019)
4. Dauner, D., Hallgarten, M., Geiger, A., Chitta, K.: Parting with misconceptions about learning-based vehicle motion planning. In: Conference on Robot Learning. pp. 1268–1281 (2023)
5. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
6. Ding, W., Xu, C., Arief, M., Lin, H., Li, B., Zhao, D.: A survey on safety-critical driving scenario generation—a methodological perspective. IEEE Transactions on Intelligent Transportation Systems **24**(7), 6971–6988 (2023)
7. Ding, W., Xu, M., Zhao, D.: Cmts: A conditional multiple trajectory synthesizer for generating safety-critical driving scenarios. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 4314–4321 (2020)
8. Elhousni, M., Lyu, Y., Zhang, Z., Huang, X.: Automatic building and labeling of hd maps with deep learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 13255–13260 (2020)

9. Feng, S., Sun, H., Yan, X., Zhu, H., Zou, Z., Shen, S., Liu, H.X.: Dense reinforcement learning for safety validation of autonomous vehicles. Nature **615**(7953), 620–627 (2023)
10. Gao, J., Sun, C., Zhao, H., Shen, Y., Anguelov, D., Li, C., Schmid, C.: Vectornet: Encoding hd maps and agent dynamics from vectorized representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11525–11533 (2020)
11. Harshvardhan, G., Gourisaria, M.K., Pandey, M., Rautaray, S.S.: A comprehensive survey and analysis of generative models in machine learning. Computer Science Review **38**, 100285 (2020)
12. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11461–11471 (2022)
13. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. In: International Conference on Learning Representations (2022)
14. Pronovost, E., Ganesina, M.R., Hendy, N., Wang, Z., Morales, A., Wang, K., Roy, N.: Scenario diffusion: Controllable driving scenario generation with diffusion. Advances in Neural Information Processing Systems **36**, 68873–68894 (2023)
15. Suo, S., Regalado, S., Casas, S., Urtasun, R.: Trafficsim: Learning to simulate realistic multi-agent behaviors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10400–10409 (2021)
16. Tan, S., Ivanovic, B., Weng, X., Pavone, M., Kraehenbuehl, P.: Language conditioned traffic generation. In: Conference on Robot Learning. pp. 2714–2752 (2023)
17. Tan, T., Zheng, Y., Liang, R., Wang, Z., Zheng, K., Zheng, J., Li, J., Zhan, X., Liu, J.: Flow matching-based autonomous driving planning with advanced interactive behavior modeling. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
18. Treiber, M., Hennecke, A., Helbing, D.: Congested traffic states in empirical observations and microscopic simulations. Physical review E **62**(2), 1805 (2000)
19. Zhang, S., Tian, J., Zhu, Z., Huang, S., Yang, J., Zhang, W.: Drivegen: Towards infinite diverse traffic scenarios with large models. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 10100–10107 (2025)
20. Zheng, Y., Liang, R., ZHENG, K., Zheng, J., Mao, L., Li, J., Gu, W., Ai, R., Li, S.E., Zhan, X., Liu, J.: Diffusion-based planning for autonomous driving with flexible guidance. In: The Thirteenth International Conference on Learning Representations (2025)