Adversarial Generation and Collaborative Evolution of Safety-Critical Scenarios for Autonomous Vehicles

Jiangfan Liu[1], Yongkang Guo[1], Fangzhi Zhong[1], Tianyuan Zhang[1], Zonglei Jing[1], Siyuan Liang[2], Jiakai Wang[3], Mingchuan Zhang[4], Aishan Liu[1*], Xianglong Liu[1,3,5]

[1]SKLCCSE, Beihang University
[2]Nanyang Technological University
[3]Zhongguancun Laboratory, Beijing
[4]Henan University of Science and Technology
[5]Institute of Dataspace, Hefei
{liujiangfan, yongkangguo, zfz0411, zhangtianyuan, raykr, liuaishan, xlliu}@buaa.edu.cn,
pandaliang521@gmail.com, wangjk@mail.zgclab.edu.cn, zhang_mch@haust.edu.cn

Abstract

The generation of safety-critical scenarios in simulation has become increasingly crucial for safety evaluation in autonomous vehicles (AV) prior to road deployment in society. However, current approaches largely rely on predefined threat patterns or rule-based strategies, which limit their ability to expose diverse and unforeseen failure modes. To overcome these, we propose SCENGE, a framework that can generate plentiful safety-critical scenarios by reasoning novel adversarial cases and then amplifying them with complex traffic flows. Given a simple prompt of a benign scene, it first performs *Meta-Scenario Generation*, where a large language model (LLM), grounded in structured driving knowledge (*e.g.*, traffic regulations, real-world accident records), infers an adversarial agent whose behavior poses a threat that is both plausible and deliberately challenging. This meta-scenario is then specified in executable code for precise in-simulator control. Subsequently, *Complex Scenario Evolution* uses background vehicles to amplify the core threat introduced by Meta-Scenario. It builds an adversarial collaborator graph to identify key agent trajectories for optimization. These perturbations are designed to simultaneously reduce the ego vehicle's maneuvering space and create critical occlusions. Extensive experiments conducted on multiple reinforcement learning (RL) based AV models show that SCENGE uncovers more severe collision cases (+31.96%) on average than SoTA baselines. Additionally, our SCENGE can be applied to large model based AV systems and deployed on different simulators; we further observe that adversarial training on our scenarios improves the model robustness. We hope our paper can build up a critical step towards building public trust and ensuring their safe deployment.

Code — https://github.com/JoFrc/ScenGE

1 Introduction

As autonomous vehicles (AVs) (Hu et al. 2023; Shao et al. 2024; Ma et al. 2024; Li et al. 2022) approach widespread deployment, ensuring their safety (Wei et al. 2018; Liang et al. 2020, 2024; Liu et al. 2025; Liang et al. 2023, 2025b,a; Liu et al. 2023; Kong, Liang, and Ren 2024) and reliability (Chen et al. 2024b,a; Ho et al. 2024) to earn public trust has become critical. However, this endeavor is challenging due to the scarcity of real-world data on rare yet critical incidents. *Simulation-based testing* (Cai et al. 2025; Lu et al. 2024) offers a controlled, reusable, and cost-effective way of evaluating behavior under various conditions, particularly safety-critical scenarios that probe their safety capacity. Current scenario generation approaches, which largely rely on predefined threat templates (Zhang, Xu, and Li 2024; Wang et al. 2024a) or rule-based strategies (Ding et al. 2020; Wang et al. 2021; Zhang et al. 2022), struggle to reveal the full spectrum of safety flaws. Due to these weak *risk exposure* abilities, AVs retain undiscovered vulnerabilities from inadequate validation.

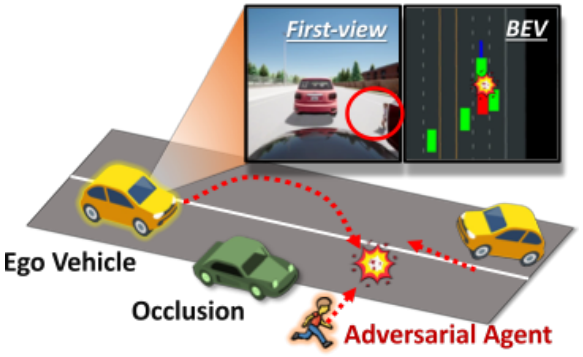

Figure 1: The illustration of the safety-critical scenario: a pedestrian emerging from behind a parked truck, with the threat amplified by a vehicle obstructing the line-of-sight.

To address these limitations, we propose SCENGE, a two-stage framework that exposes safety vulnerabilities in AV by performing adversarial threat generation and collaborative trajectory evolution. Our first stage, *Meta-Scenario Generation*, uses an LLM to creatively generate a core adversarial threat from a benign text prompt. To ensure this threat is

*Corresponding Author

both plausible and challenging, a retrieval augmented generation RAG (Wu et al. 2024; Jiao et al. 2025) framework grounds the LLM's reasoning in a structured knowledge base of traffic regulations, driver qualification standards, and realistic pre-crash scenarios. The generated meta-scenario is then expressed as executable Scenic code (Fremont et al. 2019, 2022), allowing for diverse and scalable instantiation in the CARLA simulator (Dosovitskiy et al. 2017). However, the threat created by a single adversarial agent is often predictable and insufficient to create a truly critical dilemma for the AV. Our second stage, *Complex Scenario Evolution*, therefore crafts more complex threats by coordinating the surrounding background traffic. It first builds an adversarial collaborator graph to identify the most influential vehicles. The trajectories of these key agents are then carefully optimized. Rather than causing direct collisions or simple chaos, these optimizations intensify interaction complexity by strategically limiting the ego's escape paths and obstructing its line-of-sight, ultimately increasing the likelihood of a critical incident.

Extensive experiments on multiple RL-based AV models demonstrate that SCENGE uncovers more severe collision cases (+31.96%) on average than state-of-the-art baselines. We also confirm its broad generalizability: the scenarios effectively challenge advanced VLM models like LMDrive, are transferable to the MetaDrive simulator. Beyond testing, we demonstrate the practical value of our data through adversarial training. Models trained on our scenarios exhibit substantially improved robustness. This improvement is validated on real-world nuScenes data, where the enhanced models make demonstrably safer decisions. To bridge the sim-to-real gap, we further validate our approach with real-world vehicle tests and human driver surveys, which confirm our generated scenarios represent plausible and critical real-world risks. Our main **contributions** are:

- We propose SCENGE, a two-stage framework that generates safety-critical scenarios by seamlessly combining knowledge-grounded LLM reasoning with multi-agent trajectory optimization.
- We introduce two core components: Meta-Scenario Generation, which generates richly detailed meta-scenarios by grounding an LLM's reasoning in knowledge priors; Complex Scenario Evolution, which enhances the resulting threats by optimizing the trajectory of key background vehicles identified via an Adversarial Collaborator Graph.
- Extensive experiments conducted on RL-based AV models show the effectiveness of SCENGE (+31.96% collision rate on average) compared to state-of-the-art baselines.

2 Related Work

Simulation-Based Testing for AV. Simulation-based testing, facilitated by platforms such as CARLA (Dosovitskiy et al. 2017), MetaDrive (Li et al. 2021), and LimSim (Wen et al. 2023), offers a cost-effective and controlled environment for evaluating AVs under diverse driving conditions. A key advantage is the ability to replicate rare yet critical scenarios that are impractical or hazardous to test in the real world. This capability for controlled, repeatable testing is critical for systematically identifying performance vulnerabilities and ultimately validating the safety of AVs.

Safety-Critical Scenario Generation. The generation of safety-critical scenarios is crucial for evaluating AV safety, with existing approaches including generative models that learn from real-world data (Ding et al. 2020; Suo et al. 2021; Rempe et al. 2022; Feng et al. 2023), optimization-based methods that synthesize scenarios via tailored objectives (Chen et al. 2021; Wang et al. 2021; Cao et al. 2022), and semantic-driven methods that incorporate high-level context (Zhong et al. 2023; Wang et al. 2024a,b; Zheng et al. 2025). More recent works leverage language models for generation from text or logs (Li, Azfar, and Ke 2024; Zhao et al. 2024), or enhance coverage through semantic replay and trajectory compression (Tian et al. 2024).

However, these approaches primarily focus on single-agent rule violations or replaying observed patterns, thus failing to reveal novel, compound failure modes. Even recent LLM-based methods, while improving semantic coverage, still lack fine-grained control over multi-agent interactions. These **limitations** motivate SCENGE, our framework that addresses semantic novelty and emergent risk through structured generation and collaborative perturbation.

3 SCENGE Approach

SCENGE is designed to target AV system failures stemming from explicit rule violations and subtle multi-agent interactions. An overview is illustrated in Fig. 2.

3.1 Problem Definition

Let $\mathcal{S}_{\text{meta}} = \{\mathbf{a}_{\text{ego}}, \mathbf{a}_{\text{adv}} \mid \text{R}, \text{L}\}$ denote the **meta scenario**, which includes an ego vehicle $\mathbf{a}_{\text{ego}}$ and a single adversarial agent $\mathbf{a}_{\text{adv}}$ operating within an environmental context defined by road type R and traffic light state L. The adversarial agent is further specified by semantic properties $(\text{c}, \text{p}, \text{b})$, denoting its type, position, and behavior. To construct such scenarios, we start from a simple prompt of a benign scene Φ_{base}, augmented by a fixed instruction prompt Φ_{inst} to induce safety-violating behavior. A retrieval function f_R selects relevant entries from a knowledge base $\mathbb{D}$, and the resulting context is used by an LLM f_{LLM} to produce semantic descriptions $\langle \Phi_{\text{c}}, \Phi_{\text{p}}, \Phi_{\text{b}}, \Phi_{\text{R}}, \Phi_{\text{L}} \rangle$ for adversarial agent and environment properties. These are subsequently parsed into structured values $(\text{c}, \text{p}, \text{b}, \text{R}, \text{L})$ and instantiated to define $\mathcal{S}_{\text{meta}}$.

We define the **adversarial scenario** as $\mathcal{S}_{adv} = \mathcal{S}_{meta} \cup \{\mathbf{a}_1, \ldots, \mathbf{a}_N\}$, where N denotes the number of background vehicles. Each background vehicle $\mathbf{a}_i$ follows a trajectory $\tau_i = \{(x_t, y_t)\}_{t=0}^{T}$, representing its simulated coordinates over T frames, where $i \in \{1, \ldots, N\}$. A subset of background vehicles, indexed by $K \subset \{1, \ldots, N\}$, is selected for perturbation. Their trajectory segments are optimized to induce collaborative risky behaviors that increase the threat level of $\mathcal{S}_{\text{meta}}$. Specifically, for each selected vehicle $\mathbf{a}_{i^\star}$ with $i^\star \in K$, we identify a keyframe $t_{i^\star}^\star$ as the most influential frame, and define the corresponding perturbable segment $\tilde{\tau}_{i^\star} \subset \tau_{i^\star}$ as a temporal window centered at this keyframe. These segments are perturbed by optimizing the objective function $\mathcal{L}$, yielding the final adversarial scenario $\mathcal{S}_{adv}$ with optimized segments $\{\tilde{\tau}_{i^\star}^\star\}_{i^\star \in K}$.

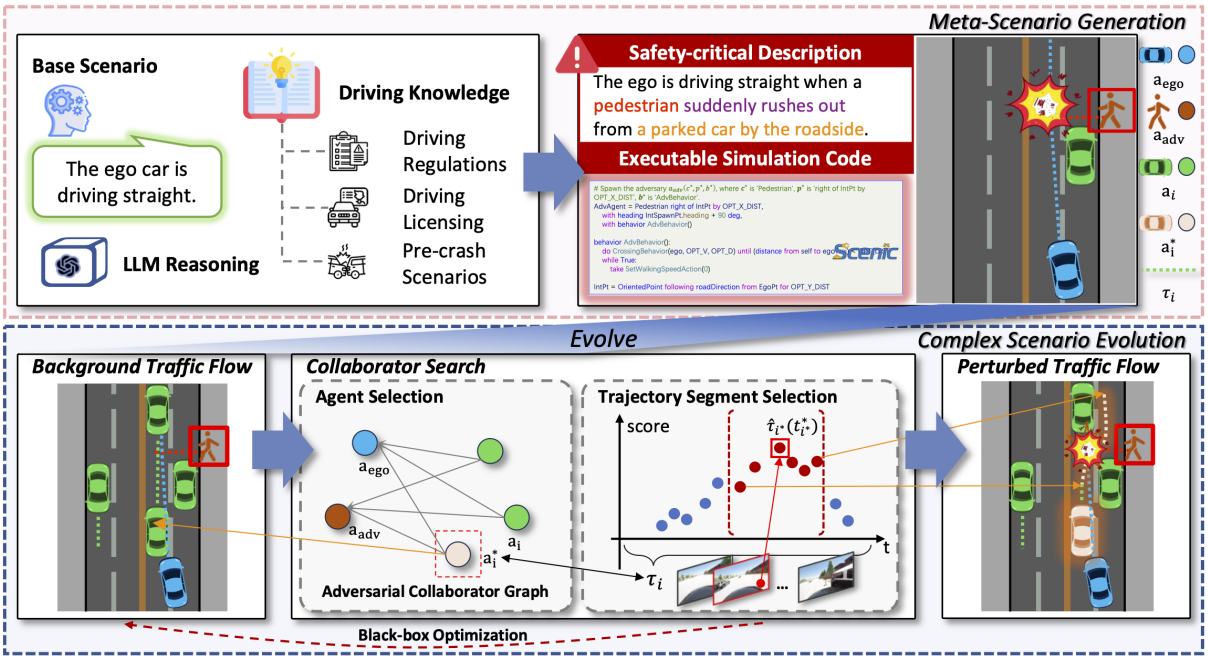

Figure 2: *Framework overview*. Given a simple description of base scenario, SCENGE first generates an meta-scenario, grounded in violations of established driving knowledge prior. SCENGE then perturbs the trajectories of key agents within the traffic.

3.2 Meta-Scenario Generation

Given a simple prompt Φ_{base}, which typically specifies a normal traffic situation without threats (*e.g.*, `the ego car is driving across the corner`), our goal is to construct a meta-scenario $\mathcal{S}_{\text{meta}}$ in which $\mathbf{a}_{\text{adv}}$ introduces a safety-critical threat. The process comprises two main components: (1) constructing a structured driving knowledge base via RAG, and (2) generating an executable scenario description using an LLM informed by that prior.

Safety Driving Knowledge Construction. The knowledge base $\mathbb{D} = \{\mathbf{D}_r, \mathbf{D}_l, \mathbf{D}_c\}$ consists of three components, each representing a distinct aspect of driving knowledge essential for simulating normative and adversarial traffic behavior. (1) $\mathbf{D}_r$ contains 27 driving regulations segmented from official manuals in the USA, Germany, and China, covering behaviors such as lane merging, overtaking, and other maneuvers. (2) $\mathbf{D}_l$ includes 100 standardized driver's license test questions and answers that assess traffic rule knowledge, situational awareness, and safe behavior selection. Together, $\mathbf{D}_r$ and $\mathbf{D}_l$ provide normative behavioral priors intentionally violated to construct safety-critical adversarial behaviors. In contrast, (3) $\mathbf{D}_c$ comprises 14 pre-crash scenarios drawn from taxonomies in the NHTSA Pre-Crash Typology Report (Najm et al. 2007) (*e.g.*, unprotected left turns, red-light violations), providing concrete adversarial patterns for scenario construction. Collectively, these components inform the synthesis of plausible threat scenarios and support the generation of critical adversarial conditions.

LLM-Driven Scenario Generation. Given a simple prompt Φ_{base} of a benign scene, the LLM is prompted to generate a detailed, safety-critical scenario by introducing one main adversarial agent $\mathbf{a}_{\text{adv}}$ into the scenario. It infers the agent's properties and the associated environmental context through in-context learning (Dong et al. 2024). However, simply adopting an LLM may lead to unsafe or unrealistic critical scenarios; thus, we ground the reasoning process in structured driving knowledge. To this end, relevant knowledge is retrieved from the database $\mathbb{D}$ and combined with the instruction prompt Φ_{inst} to form the input to the LLM f_{LLM}. The generation process is formalized as:

$$\begin{aligned}\langle \Phi_{\text{c}}, \Phi_{\text{p}}, \Phi_{\text{b}}, \Phi_{\text{R}}, \Phi_{\text{L}} \rangle = \\ f_{\text{LLM}}\left(\mathbf{a}_{\text{ego}}, \Phi_{\text{base}}, f_R(\mathbb{D}, \Phi_{\text{base}}) \mid \Phi_{\text{inst}}\right),\end{aligned} \tag{3.1}$$

where each Φ_* represents a text description of a scenario element, including the adversarial agent's properties and environmental context. The instruction prompt Φ_{inst} explicitly guides the model to generate rule-violating yet plausible actions, grounded in retrieved safety knowledge. Although expressed in textual form, the generation is controlled through few-shot prompting and slot-based templates, ensuring the outputs remain structured and scenario-compatible.

The generated descriptions are then parsed into structured values $(\text{c}, \text{p}, \text{b}, \text{R}, \text{L})$ and populated into a predefined Scenic template. This template encodes scenario-level semantics while enforcing syntactic and physical constraints, bridging language-driven generation and executable simulation. The resulting program is run in the simulator to instantiate $\mathcal{S}_{\text{meta}}$.

3.3 Complex Scenario Evolution

Building on the generated meta-scenario, Complex Scenario Evolution enhances its complexity by introducing back-

ground vehicles $\{\mathbf{a}_1, \ldots, \mathbf{a}_N\}$ with collaborative risky trajectories. To that end, their interactions with $\mathbf{a}_{\text{adv}}$ and $\mathbf{a}_{\text{ego}}$ are adjusted to create a more challenging scenario for the AV. This process comprises two main components: (1) *Collaborator Search*, which identifies the background vehicles that can most amplify the adversarial nature of the scenario, and (2) *Trajectory Perturbation*, which adjusts the selected vehicles to maximize the adversarial impact.

Collaborator Search. To identify influential background vehicles, we construct an Adversarial Collaborator Graph $\boldsymbol{G}$, where each node corresponds to an agent in the scenario, and the edges reflect directional behavioral relevance, particularly emphasizing the impact of background vehicles on the ego vehicle and adversarial agent. This graph is derived from a frame-wise attention matrix $\boldsymbol{M}_{\text{a}}$ that models trajectory-level dependencies using ego and adversarial trajectories as queries and background trajectories as keys. Specifically:

$$\boldsymbol{M}_{\text{a}} = \frac{(\tau_{\text{ego}}, \tau_{\text{adv}}) \cdot (\tau_1, \ldots, \tau_N)^\top}{\sqrt{d}} + \boldsymbol{M}_{\text{m}} + \log \boldsymbol{M}_{\text{d}}, \quad (3.2)$$

where d is the dimension of τ, $\boldsymbol{M}_{\text{m}}$ enforces causality by preventing attention to future frames, and $\boldsymbol{M}_{\text{d}}$ introduces a temporal decay bias to emphasize recent interactions.

Based on $\boldsymbol{M}_{\text{a}}$, we perform *Collaborator Search* in two stages. First, we aggregate attention scores across frames to estimate each background vehicle's relevance to the ego vehicle and the adversarial agent. This process identifies the Top-k most influential vehicles, which form the collaborator set K. Then, for each $i^\star \in K$, we locate the keyframe $t^\star_{i^\star}$ receiving the highest attention score for vehicle $\mathbf{a}_{i^\star}$, and extract a local temporal window $\tilde{\tau}_{i^\star}$ centered at keyframe as its perturbable trajectory segment. These segments serve as the input to the subsequent trajectory perturbation module.

Trajectory Perturbation. We optimize the perturbable segments $\{\tilde{\tau}_{i^\star}\}_{i^\star \in K}$ of selected collaborators (indexed by K) to maximize the adversarial impact on the ego vehicle. This is formulated as the following objective:

$$\{\tilde{\tau}^\star_{i^\star}\}_{i^\star \in K} = \underset{\{\tilde{\tau}_{i^\star}\}_{i^\star \in K}}{\arg\max}\ \mathcal{L}(\tilde{\tau}_{\text{ego}}, \tilde{\tau}_{\text{adv}}, \{\tilde{\tau}_{i^\star}\}_{i^\star \in K}), \quad (3.3)$$

The optimization follows an iterative, gradient-based procedure. Specifically, for each perturbable segment, we compute the gradient of $\mathcal{L}$ *w.r.t* the trajectory coordinates and update them in the direction that increases the loss. Each update step uses a small, fixed step size and is projected back to the feasible space to ensure realism. The process continues until convergence or a predefined number of steps is reached.

$$\begin{aligned} \mathcal{L} = \lambda_1 \underbrace{\|\tilde{\tau}_{i^\star} - \tilde{\tau}_{\text{ego}}\|_2}_{\mathcal{L}_{\text{ego}}} + \lambda_2 \underbrace{\|\Delta^2 \tilde{\tau}_{i^\star}\|_2^2}_{\mathcal{L}_{\text{smooth}}} \\ + \lambda_3 \underbrace{\|(\tilde{\tau}_{i^\star} - \tilde{\tau}_{\text{ego}}) \times (\tilde{\tau}_{\text{adv}} - \tilde{\tau}_{\text{ego}})\|_\perp}_{\mathcal{L}_{\text{occ}}}. \end{aligned} \quad (3.4)$$

The objective function $\mathcal{L}$ comprises three components, as shown in Eq. (3.4). (1) $\mathcal{L}_{\text{ego}}$ minimizes the Euclidean distance between the perturbed background trajectory $\tilde{\tau}_{i^\star}$ and the ego trajectory $\tilde{\tau}_{\text{ego}}$ within a temporal window. (2) $\mathcal{L}_{\text{smooth}}$ penalizes second-order differences $\Delta^2 \tilde{\tau}_{i^\star}$ to reduce abrupt motion changes. (3) $\mathcal{L}_{\text{occ}}$ minimizes the normalized perpendicular distance via a 2D cross product, promoting alignment along the ego–adversary line-of-sight. From a behavioral modeling perspective, $\mathcal{L}_{\text{ego}}$ encourages spatial proximity to induce planning hesitation, $\mathcal{L}_{\text{smooth}}$ ensures kinematic feasibility via smoothness constraints, collectively balancing adversarial strength with physical plausibility, and $\mathcal{L}_{\text{occ}}$ amplifies perceptual ambiguity through occlusion. Finally, the optimized perturbations $\{\tilde{\tau}^\star_{i^\star}\}_{i^\star \in K}$ replace the corresponding segments of the original trajectories, yielding the final adversarial scenario $\mathcal{S}_{adv}$, which poses a significant threat to the ego vehicle's safe driving.

3.4 Overall Generation Workflow

Our workflow begins with the Meta-Scenario Generation module. This module takes a benign text description as input. First, an LLM generates a detailed textual description of a safety-critical threat. This generation process is grounded in a structured driving knowledge base. The LLM then translates this textual description into a parameterized Scenic script. Finally, this script is instantiated within the CARLA simulator to create the executable meta-scenario.

Next, we process the background traffic for the meta-scenario. We use CARLA's Traffic Manager to generate a flow of background vehicles and record their baseline trajectories. The Complex Scenario Evolution module then analyzes these trajectories offline. It builds an Adversarial Collaborator Graph to identify the most influential background vehicles to act as collaborators. The module selects critical segments of their trajectories and optimizes them to maximize the overall threat. This optimization aims to restrict the ego vehicle's maneuvering space and create critical occlusions. For the final evaluation, we replay the complete scenario in a closed-loop simulation. The optimized background vehicles execute their new adversarial trajectories as scripted events, forcing the ego vehicle to react to the coordinated, high-risk situation.

4 Experiment and Evaluation

4.1 Experimental Setup

Simulation environment and benchmark. We utilise the CARLA simulator (Dosovitskiy et al. 2017), an open-source and highly customizable urban driving simulator, to create a closed-loop simulation environment. We adopt SafeBench (Xu et al. 2022) as the benchmarking framework, which supports diverse RL-based AV agents and standardized evaluation. Following (Zhang, Xu, and Li 2024), we use 8 base traffic scenarios (*e.g.*, `Straight Obstacle`, `Lane Changing`) curated from the NHTSA Pre-Crash Typology Report (Najm et al. 2007), each containing 10 diverse driving routes. For each route, 10 adversarial scenarios are generated, resulting in 800 challenging scenarios for evaluation and comparison per method.

AV algorithms. Following (Zhang, Xu, and Li 2024), we mainly employ 3 representative RL-based AV algorithms as testing agents, including Proximal Policy Optimization (PPO) (Schulman et al. 2017), Soft Actor-Critic (SAC) (Haarnoja

Metric	Algo.	Base Traffic Scenarios								Avg.
		Straight Obstacle	Turning Obstacle	Lane Changing	Vehicle Passing	Red-light Running	Unprotected Left-turn	Right-turn	Crossing Negotiation	
CR ↑	LC	0.241	0.159	0.736	0.792	0.317	0.325	0.321	0.313	0.401
	AS	0.451	0.399	0.726	0.832	0.177	0.335	0.115	0.303	0.417
	CS	0.391	0.679	0.756	0.812	0.237	0.325	0.411	0.333	0.493
	AT	0.441	0.379	0.646	0.782	0.317	0.315	0.321	0.353	0.440
	ChatS	0.750	0.647	0.660	**0.907**	**0.833**	0.620	0.743	0.850	0.751
	Ours	**0.860**	**0.773**	**0.837**	0.897	0.823	**0.747**	**0.763**	**0.863**	**0.820**
OS ↓	LC	0.789	0.816	0.566	0.530	0.799	0.790	0.692	0.717	0.712
	AS	0.694	0.687	0.561	0.506	0.866	0.775	0.841	0.721	0.706
	CS	0.726	0.552	0.549	0.513	0.839	0.787	0.649	0.708	0.665
	AT	0.696	0.706	0.599	0.528	0.805	0.795	0.689	0.698	0.690
	ChatS	0.559	0.572	0.607	0.472	0.544	0.656	0.511	**0.459**	0.548
	Ours	**0.503**	**0.526**	**0.504**	**0.457**	**0.507**	**0.519**	**0.498**	0.477	**0.499**

Table 1: Evaluation of adversarial scenario generation methods across CR and OS metrics. ChatS is the abbreviation of Chatscene.

et al. 2018), and Twin Delayed Deep Deterministic Policy Gradient (TD3) (Fujimoto, van Hoof, and Meger 2018).

Compared baselines. We compare our SCENGE with several existing scenario generation methods, including Learning-to-Collide (LC) (Ding et al. 2020), AdvSim (AS) (Wang et al. 2021), Carla Scenario Generator (CS) (Dosovitskiy et al. 2017), Adversarial Trajectory Optimization (AT) (Zhang et al. 2022), and ChatScene (Zhang, Xu, and Li 2024). For fair comparisons, each method is applied on the same 8 base scenarios and routes to generate 800 challenging scenarios under consistent generation logic and evaluation settings.

Metrics. Following SafeBench (Xu et al. 2022), we adopt a set of key metrics to evaluate AV performance in generated scenarios. Two core indicators are used: the **collision rate** (CR ↑) measures the frequency of collisions and reflects safety risk, and the **overall score** (OS ↓) aggregates system-level performance. In addition, we evaluate three additional dimensions: the **safety level** (*frequency of running red lights* (RR ↑), *frequency of running stop signs* (SS ↑), and *average distance driven out of road* (OR ↑), the **functionality level** (*route following stability* (RF ↓), *average percentage of route completion* (Comp ↓), and *average time spent to complete the route* (TS ↑)), and the **etiquette level** (*average acceleration* (ACC ↑), *average yaw velocity* (YV ↑), and *frequency of lane invasion* (LI ↑). Higher (↑) values indicate worse performance, while ↓ indicates the contrary.

Implementation details. All experiments were conducted on a server with an Intel(R) Core(TM) i9-14900K CPU and two NVIDIA GeForce RTX 4090 GPUs with 24GB memory. The LLM used for Meta-Scenario Generationis qwq-32b (Team 2025), the reasoning model from the Qwen series. We then construct 10 background vehicles and perturb the trajectories of 4 selected ones, each over 60% of their trajectory. The 4 perturbed vehicles are selected based on the highest attention relevance to ego and adversarial agents. The 60% perturbation window is centered around each vehicle's most relevant keyframe. We set γ in the decay matrix to 0.8. In the loss calculation, we set $\lambda_1 = 0.3$, $\lambda_2 = 0.2$, and $\lambda_3 = 0.5$.

4.2 Main Results

Our main results, presented in Tab. 1 and Tab. 2, compare SCENGE against baseline methods across eight base scenarios. To ensure a robust and generalizable evaluation, all metrics are averaged over three distinct RL agents, as detailed in Sec. 4.1. This approach validates that our scenarios pose a universal challenge to a range of modern driving policies, rather than merely exploiting agent-specific flaws. The tables offer complementary views: Tab. 1 details the CR and OS for each individual scenario, whereas Tab. 2 assesses the overall impact on AV behavior from the three key aspects of *safety and risk exposure*, *functionality under stress*, and *driving etiquette*, with all results averaged across the scenarios.

Safety and Risk Exposure. As shown in Tab. 1, SCENGE achieves the highest average CR. Notably, this high collision rate is not achieved by forcing simplistic rule violations. In fact, the scores for RR, SS, and OR are not the highest, as seen in Tab. 2. This outcome is a direct result of our design philosophy. We avoid creating simplistic, predictable setups that rely on obvious rule violations. Instead, SCENGE focuses on generating high-pressure situations from complex multi-agent interactions. These plausible scenarios challenge the predictive and planning capabilities under pressure. They force the vehicle to navigate moments where no single, simple rule applies. Consequently, the resulting collisions expose more profound and subtle vulnerabilities in the AV's core logic, rather than surface-level failures in rule compliance.

Functionality Challenges. As shown in Tab. 1, SCENGE achieves the lowest average OS. Additionally, Tab. 2 shows a **4.96%** drop in RF and a **29.16%** reduction in Comp. The moderate TS is caused by early collisions, which demonstrates that SCENGE induces rapid and decisive failures by persistently disrupting the AV's planning.

Driving Etiquette. As shown in Tab. 2, SCENGE increases ACC, YV, and LI by **16.5%**, **12.7%**, and **8.48%** respectively. These results suggest that SCENGE causes AV to behave less smoothly and more erratically in ways that remain socially plausible. Introducing temporally coordinated perturbations across multiple agents disrupts fine-grained control and social driving compliance, revealing limitations that simpler, single-

agent or rule-based methods fail to expose.

Algo.	Safety			Functionality			Etiquette		
	RR ↑	SS ↑	OR ↑	RF ↓	Comp ↓	TS ↑	ACC ↑	YV ↑	LI ↑
LC	**0.325**	0.165	0.039	0.884	0.807	0.224	0.225	0.231	0.087
AS	0.299	0.167	0.032	0.901	0.821	0.269	0.217	0.233	0.102
CS	0.312	**0.168**	**0.043**	0.880	0.817	0.252	0.229	0.235	0.106
AT	0.311	0.167	0.035	0.883	0.802	**0.287**	0.233	0.236	0.112
ChatS	0.228	0.145	0.018	0.890	0.571	0.074	0.281	0.225	0.064
Ours	0.231	0.125	0.009	**0.838**	**0.472**	0.124	**0.402**	**0.359**	**0.179**

Table 2: Aggregated evaluation results across three dimensions. ChatS is the abbreviation of Chatscene.

4.3 Ablation Studies

We perform ablation experiments by selectively disabling key modules and observing the effect on performance. Otherwise specified, this part keeps the same setting as the main experiment. Fig. 3 reports the results under different settings.

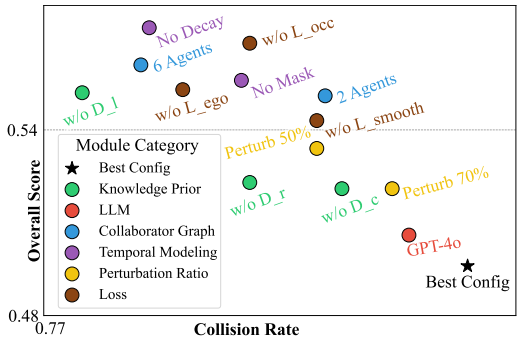

Figure 3: Scatter-plot visualization of ablation studies.

Knowledge Prior. Removing $\mathbf{D}_r$ yields CR 79.4% and OS 52.3%, reflecting its role in guiding rule-focused violations. Removing $\mathbf{D}_l$ gives CR 77.4% and OS 55.2%, showing its effect on enhancing logical consistency in behavior. Removing $\mathbf{D}_c$ results in CR 80.5% and OS 52.1%, confirming its importance in producing realistic and high-risk scenarios.

LLM. Both GPT-4o and qwq-32b demonstrate strong capabilities, yielding similar CRs of 81.3% and 82.0%, respectively. This suggests that the performance is not critically sensitive to the specific choice of a powerful LLM.

Collaborator Graph. We ablate the number of perturbed agents by selecting 2, 4, and 6 collaborators. Perturbing 4 agents performs best with CR 82% and OS 49.9%, balancing adversarial strength and scenario plausibility. Interestingly, perturbing 6 agents (CR 78.1%) is less effective than perturbing 2 (CR 80.3%). We hypothesize that this occurs because an excessive number of agents transforms our intended coordinated, precise threat into easily avoidable chaos. Their mutual interference and abnormal behavior likely prompt the AV to adopt a conservative policy. This hypothesis is also supported by OS, where the 6-agent setting yields a higher score (56.1%) compared to the 2-agent setting (55.1%).

Temporal Modeling. The full setting (with both mask and decay) yields the best result with CR 82% and OS 49.9%. Removing the temporal mask reduces temporal causality in collaborator selection, leads to CR 79.3% and OS 55.6%, while removing the temporal decay results in CR 78.2% and OS 57.3%. These results highlight the complementary role of both components in capturing temporally coherent influence.

Perturbation Ratio. We compare three perturbation ratios centered around the selected keyframe: 50%, 60%, and 70%. Perturbing 60% of the segment achieves the best result with CR 82% and OS 49.9%. 50% leads to CR 80.2% and OS 53.4%, indicating insufficient behavioral deviation, while 70% causes CR 81.1% and OS 52.1% due to overmodification and reduced plausibility. These results shows moderate ratio balances realism and adversarial effect.

Loss. We ablate each component in $\mathcal{L}$ to assess its contribution. Removing $\mathcal{L}_{\text{ego}}$ leads to CR 78.6% and OS 55.3%, reflecting reduced collision targeting. Excluding $\mathcal{L}_{\text{smooth}}$ yields CR 80.2% and OS 54.3%, with trajectories becoming visibly unstable. Removing $\mathcal{L}_{\text{occ}}$ results in CR 79.4% and OS 56.8%, indicating weaker alignment between adversary and ego. The full loss yields the best trade-off, and ablating any term consistently reduces CR and increases OS.

4.4 Generalization Ability Analysis

We study the generalizability of SCENGE: (1) effectiveness on other AV models; and (2) applicability on other simulators.

Model Generalization. Beyond RL-based AV models, we further evaluate our generated scenarios on LMDrive (Shao et al. 2024), a large vision-language model for AV deployed on the CARLA Leaderboard (Dosovitskiy et al. 2017). LMDrive navigates by following natural language instructions sequentially, using the multi-view camera and Lidar perception for scene understanding and planning. To accommodate its instruction-driven execution mode, we redesign the test routes into multi-instruction sequences. Evaluation follows LMDrive's original metrics: Route Completion (RC), Infraction Score (IS), and Driving Score (DS). We evaluate LMDrive under three increasingly challenging settings: (1) ego-only benign routes as a baseline, (2) meta-scenarios with a single adversarial agent, and (3) full adversarial scenarios generated by our framework, including the perturbed background vehicle. As shown in Tab. 3, LMDrive's performance drops from **87.7** DS in the benign case to **83.7** in meta-scenarios and further to **80.4** under full adversarial conditions. These results demonstrate that our generated scenarios significantly stress LMDrive's planning capability, especially under rare or occluded interactions.

Algo.	Benign Scenario	Meta Scenario	Adversarial Scenario
RC	92.2 $\pm$ 2.9	92.9 $\pm$ 2.7	89.9 $\pm$ 5.1
IS	0.97 $\pm$ 0.01	0.9 $\pm$ 0.05	0.89 $\pm$ 0.04
DS	87.7 $\pm$ 2.4	83.7 $\pm$ 4.7	80.4 $\pm$ 5.5

Table 3: Performance of LMDrive.

4.5 Training on the Generated Scenarios

This section evaluates the utility of our SCENGE in enhancing AV robustness by adversarial training on the generated scenarios.

Robustness evaluation in simulation. We adversarially train the SAC-based ego vehicle across eight base traffic scenarios using scenes from the first eight routes per scenario, and evaluate on unseen scenes from the remaining two routes. The training process uses 500 epochs with a learning rate of 0.0001. As shown in Tab. 4, adversarial training with SCENGE-generated scenarios yields the best results among methods, reducing the CR by **3.1%** while increasing the OS by **94.7%**. These results demonstrate that our method generates scenarios missing from standard AV training. Adversarial training on these data remedies the AV vulnerabilities, leading to a significant improvement in robustness.

Metric	LC	AS	CS	AT	ChatScene	**Ours**
CR ↑	0.210	0.216	0.176	0.135	0.043	**0.031**
OS ↓	0.813	0.806	0.825	0.864	0.905	**0.947**

Table 4: Evaluation of adversarially trained ego vehicle.

Robustness evaluation on real-world data. We further evaluated our adversarially trained model on real-world data from the nuScenes dataset (Caesar et al. 2020) to address the visual domain gap between simulation and reality. In particular, we first manually select 140 scenario segments (2700 images) with latent risks, such as a pedestrian standing by the roadside; subsequently, we evaluate both the original RL model and the enhanced model on these real-world image sequences. The enhanced model achieved a **21.7% lower** Euclidean distance to the ground truth trajectory than the original model, indicating safer and more stable decisions. This result demonstrates that the robustness gained from our scenarios transfers effectively to real-world perception data.

4.6 Real-world Experiments

Figure 4: Real-World Experiments.

Here, we conducted real-world experiments in a closed road, where we arranged a layout that is similar to our generated scenario and tested a real-world vehicle (Fig. 4). Due to commercial confidentiality, the appearance of our test vehicle has been obscured. Under strict safety protocols, we recreated and repeated two challenging scenarios 15 times each: a pedestrian suddenly emerging from an occlusion, and an unprotected left-turn challenged by an accelerating scooter. Despite safety measures designed to lower the scenarios' difficulty (*e.g.*, limiting vehicle speed), the AV exhibited critical failures. In **73.3%** of the pedestrian tests, the vehicle failed to react to the pedestrian's emergence in a timely manner. Similarly, in **60%** of the left-turn tests, it failed to alter its trajectory to avoid the conflicting scooter, exposing a decisive vulnerability. These results provide definitive, physical-world evidence that our generated scenarios identify physical risks for autonomous systems, not just simulation artifacts.

4.7 Discussion and Analysis

Human Evaluation. We conducted a human study with 30 licensed drivers to assess our generated scenarios. In the study, participants viewed 40 unique videos and rated them using a 5-point scale. The survey covered all eight base scenarios used in our experiments, with five distinct variations selected for each. An analysis of 1,200 responses revealed high average scores for both plausibility (**4.765** out of 5) and perceived risk (**4.934** out of 5). This confirms that human drivers consider the scenarios to be both realistic and dangerous.

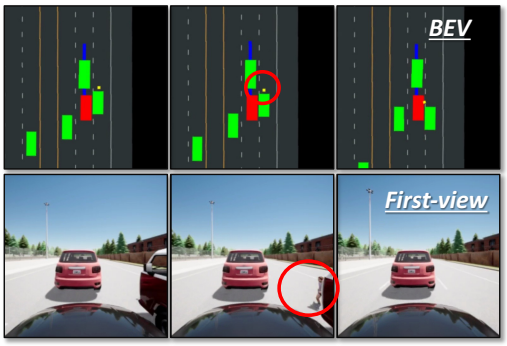

Figure 5: Case visualization. Frames of the collision.

Case Study. In Fig. 5, we show some frames from the final adversarial scenario. This scenario involves a pedestrian emerging from an occlusion. The threat is amplified by background vehicles: one occupies the adjacent lane to restrict maneuvering space (a behavior promoted by $\mathcal{L}_{\text{ego}}$), while another obstructs the line-of-sight to the adversary (guided by $\mathcal{L}_{\text{occ}}$). These compounded interactions prevent the ego vehicle from executing a safe evasive action. This case effectively illustrates how SCENGE uses subtle, coordinated behaviors of background traffic to expose critical AV vulnerabilities.

5 Conclusion

In this paper, we introduce SCENGE, a two-stage framework for generating safety-critical scenarios to expose vulnerabilities in AV. From a benign scene description, SCENGE introduces *Meta-Scenario Generation* and *Complex Scenario Evolution* to generate scenarios that are more likely to cause failures. Experiments on multiple RL-based AV models show that SCENGE reveals more severe collision cases.

Ethical Statement

SCENGE is designed for exposing vulnerabilities in AV systems, while acknowledging the ethical considerations. **(1) Safety Equity.** AV systems may inherit biases, potentially offering less protection to non-standard road agents, such as wheelchair users or animals. SCENGE can generate corresponding training data to impove AV models. **(2) Ethical Responsibility in Infrastructure.** Ethical responsibility demands that we build new urban infrastructure without creating foreseeable risks to the public. SCENGE provides a necessary tool for 'digital safety audits' before construction. **(3) Responsible Deployment and Potential for Misuse.** We recognize the potential for misuse of SCENGE. We advocate for regulatory oversight and responsible deployment to ensure that SCENGE are used for their intended purpose: **improving system robustness and safety.**

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62476018, 62206009), the Fundamental Research Funds for the Central Universities, the State Key Laboratory of Complex & Critical Software Environment (CCSE), and Aeronautical Science Fund (Grant. 20230017051001).

References

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.

Cai, X.; Bai, X.; Cui, Z.; Xie, D.; Fu, D.; Yu, H.; and Ren, Y. 2025. Text2Scenario: Text-Driven Scenario Generation for Autonomous Driving Test. arXiv:2503.02911.

Cao, Y.; Xiao, C.; Anandkumar, A.; Xu, D.; and Pavone, M. 2022. Advdo: Realistic adversarial attacks for trajectory prediction. In *European Conference on Computer Vision*, 36–52. Springer.

Chen, B.; Chen, X.; Wu, Q.; and Li, L. 2021. Adversarial evaluation of autonomous vehicles in lane-change scenarios. *IEEE transactions on intelligent transportation systems*, 23(8): 10333–10342.

Chen, R.; Liang, S.; Li, J.; Liu, S.; Li, M.; Huang, Z.; Zhang, H.; and Cao, X. 2024a. Interpreting object-level foundation models via visual precision search. arXiv:2411.16198.

Chen, R.; Zhang, H.; Liang, S.; Li, J.; and Cao, X. 2024b. Less is More: Fewer Interpretable Region via Submodular Subset Selection. arXiv:2402.09164.

Ding, W.; Chen, B.; Xu, M.; and Zhao, D. 2020. Learning to collide: An adaptive safety-critical scenarios generating method. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2243–2250. IEEE.

Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Ma, J.; Li, R.; Xia, H.; Xu, J.; Wu, Z.; Liu, T.; Chang, B.; Sun, X.; Li, L.; and Sui, Z. 2024. A Survey on In-context Learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 1107–1128. Miami, Florida, USA: Association for Computational Linguistics.

Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; and Koltun, V. 2017. CARLA: An Open Urban Driving Simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, 1–16.

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

Fremont, D. J.; Dreossi, T.; Ghosh, S.; Yue, X.; Sangiovanni-Vincentelli, A. L.; and Seshia, S. A. 2019. Scenic: a language for scenario specification and scene generation. In *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*, 63–78.

Fremont, D. J.; Kim, E.; Dreossi, T.; Ghosh, S.; Yue, X.; Sangiovanni-Vincentelli, A. L.; and Seshia, S. A. 2022. Scenic: A language for scenario specification and data generation. *Machine Learning*, 1–45.

Fujimoto, S.; van Hoof, H.; and Meger, D. 2018. Addressing Function Approximation Error in Actor-Critic Methods. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, 1587–1596.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, 1861–1870.

Ho, Z. Y.; Liang, S.; Zhang, S.; Zhan, Y.; and Tao, D. 2024. NoVo: Norm Voting off Hallucinations with Attention Heads in Large Language Models. arXiv:2410.08970.

Hu, Y.; Yang, J.; Chen, L.; Li, K.; Sima, C.; Zhu, X.; Chai, S.; Du, S.; Lin, T.; Wang, W.; Lu, L.; Jia, X.; Liu, Q.; Dai, J.; Qiao, Y.; and Li, H. 2023. Planning-oriented Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Jiao, Y.; Tan, Z.; Yang, D.; Sun, D.; Feng, J.; Shen, Y.; Wang, J.; and Wei, P. 2025. Hirag: Hierarchical-thought instruction-tuning retrieval-augmented generation. arXiv:2507.05714.

Kong, D.; Liang, S.; and Ren, W. 2024. Environmental Matching Attack Against Unmanned Aerial Vehicles Object Detection. arXiv:2405.07595.

Li, Q.; Peng, Z.; Xue, Z.; Zhang, Q.; and Zhou, B. 2021. MetaDrive: Composing Diverse Driving Scenarios for Generalizable Reinforcement Learning. arXiv:2109.12674.

Li, S.; Azfar, T.; and Ke, R. 2024. Chatsumo: Large language model for automating traffic scenario generation in simulation of urban mobility. *IEEE Transactions on Intelligent Vehicles*.

Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Qiao, Y.; and Dai, J. 2022. BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers. In *European Conference on Computer Vision*, 36–52. Springer.

Liang, J.; Liang, S.; Liu, A.; and Cao, X. 2025a. Vl-trojan: Multimodal instruction backdoor attacks against autoregressive visual language models. *International Journal of Computer Vision*, 1–20.

Liang, S.; Liang, J.; Pang, T.; Du, C.; Liu, A.; Zhu, M.; Cao, X.; and Tao, D. 2025b. Revisiting Backdoor Attacks against Large Vision-Language Models from Domain Shift. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 9477–9486.

Liang, S.; Wang, W.; Chen, R.; Liu, A.; Wu, B.; Chang, E.-C.; Cao, X.; and Tao, D. 2024. Object detectors in the open environment: Challenges, solutions, and outlook. arXiv:2403.16271.

Liang, S.; Wei, X.; Yao, S.; and Cao, X. 2020. Efficient adversarial attacks for visual object tracking. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*.

Liang, S.; Zhu, M.; Liu, A.; Wu, B.; Cao, X.; and Chang, E.-C. 2023. Badclip: Dual-embedding guided backdoor attack on multimodal contrastive learning. arXiv:2311.12075.

Liu, A.; Guo, J.; Wang, J.; Liang, S.; Tao, R.; Zhou, W.; Liu, C.; Liu, X.; and Tao, D. 2023. {X-Adv}: Physical adversarial object attacks against x-ray prohibited item detection. In *32nd USENIX Security Symposium (USENIX Security 23)*.

Liu, M.; Liang, S.; Howlader, K.; Wang, L.; Tao, D.; and Zhang, W. 2025. Natural Reflection Backdoor Attack on Vision Language Model for Autonomous Driving. arXiv:2505.06413.

Lu, Q.; Wang, X.; Jiang, Y.; Zhao, G.; Ma, M.; and Feng, S. 2024. Multimodal large language model driven scenario testing for autonomous vehicles. arXiv:2409.06450.

Ma, Y.; Cao, Y.; Sun, J.; Pavone, M.; and Xiao, C. 2024. Dolphins: Multimodal language model for driving. In *European Conference on Computer Vision*, 403–420. Springer.

Najm, W. G.; Smith, J. D.; Yanagisawa, M.; et al. 2007. Pre-crash scenario typology for crash avoidance research. Technical report, United States. National Highway Traffic Safety Administration.

Rempe, D.; Philion, J.; Guibas, L. J.; Fidler, S.; and Litany, O. 2022. Generating useful accident-prone driving scenarios via a learned traffic prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17305–17315.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. arXiv:1707.06347.

Shao, H.; Hu, Y.; Wang, L.; Song, G.; Waslander, S. L.; Liu, Y.; and Li, H. 2024. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15120–15130.

Suo, S.; Regalado, S.; Casas, S.; and Urtasun, R. 2021. Trafficsim: Learning to simulate realistic multi-agent behaviors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10400–10409.

Team, Q. 2025. QwQ-32B: Embracing the Power of Reinforcement Learning.

Tian, H.; Han, X.; Wu, G.; Zhou, Y.; Li, S.; Wei, J.; Ye, D.; Wang, W.; and Zhang, T. 2024. LMM-enhanced Safety-Critical Scenario Generation for Autonomous Driving System Testing From Non-Accident Traffic Videos. arXiv:2406.10857.

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

Wang, X.; Zhu, Z.; Huang, G.; Chen, X.; Zhu, J.; and Lu, J. 2024a. DriveDreamer: Towards Real-World-Drive World Models for Autonomous Driving. In *European Conference on Computer Vision*, 55–72. Springer.

Wang, Y.; He, J.; Fan, L.; Li, H.; Chen, Y.; and Zhang, Z. 2024b. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14749–14759.

Wei, X.; Liang, S.; Chen, N.; and Cao, X. 2018. Transferable adversarial attacks for image and video object detection. arXiv:1811.12641.

Wen, L.; Fu, D.; Mao, S.; Cai, P.; Dou, M.; Li, Y.; and Qiao, Y. 2023. LimSim: A long-term interactive multi-scenario traffic simulator. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, 1255–1262. IEEE.

Wu, S.; Xiong, Y.; Cui, Y.; Wu, H.; Chen, C.; Yuan, Y.; Huang, L.; Liu, X.; Kuo, T.-W.; Guan, N.; et al. 2024. Retrieval-augmented generation for natural language processing: A survey. arXiv:2407.13193.

Xu, C.; Ding, W.; Lyu, W.; Liu, Z.; Wang, S.; He, Y.; Hu, H.; Zhao, D.; and Li, B. 2022. Safebench: A benchmarking platform for safety evaluation of autonomous vehicles. *Advances in Neural Information Processing Systems*, 35: 25667–25682.

Zhang, J.; Xu, C.; and Li, B. 2024. Chatscene: Knowledge-enabled safety-critical scenario generation for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15459–15469.

Zhang, Q.; Hu, S.; Sun, J.; Chen, Q. A.; and Mao, Z. M. 2022. On Adversarial Robustness of Trajectory Prediction for Autonomous Vehicles. arXiv:2201.05057.

Zhao, Y.; Xiao, W.; Mihalj, T.; Hu, J.; and Eichberger, A. 2024. Chat2Scenario: Scenario Extraction From Dataset Through Utilization of Large Language Model. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, 559–566. IEEE.

Zheng, W.; Chen, W.; Huang, Y.; Zhang, B.; Duan, Y.; and Lu, J. 2025. OccWorld: Learning a 3D Occupancy World Model for Autonomous Driving. In *European Conference on Computer Vision*, 55–72. Springer.

Zhong, Z.; Rempe, D.; Chen, Y.; Ivanovic, B.; Cao, Y.; Xu, D.; Pavone, M.; and Ray, B. 2023. Language-guided traffic simulation via scene-level diffusion. In *Conference on Robot Learning*, 144–177. PMLR.