

AED: Automatic Discovery of Effective and Diverse Vulnerabilities for Autonomous Driving Policy with Large Language Models

Le Qiu^{1*}, Zelai Xu^{1*}, Qixin Tan^{2*}, Wenhao Tang^{1,3}, Chao Yu^{1†}, Yu Wang^{1†}

¹Department of Electronic Engineering, Tsinghua University

² Institute for Interdisciplinary Information Sciences, Tsinghua University

³ Tsinghua Shenzhen International Graduate School, Tsinghua University

* Equal contribution, qiu121@mails.tsinghua.edu.cn

†Corresponding authors: zoeyuchao@gmail.com, yu-wang@tsinghua.edu.cn

Abstract

Assessing the safety of autonomous driving policy is of great importance, and reinforcement learning (RL) has emerged as a powerful method for discovering critical vulnerabilities in driving policies. However, existing RL-based approaches often struggle to identify vulnerabilities that are both effective—meaning the autonomous vehicle is genuinely responsible for the accidents—and diverse—meaning they span various failure types. To address these challenges, we propose **AED**, a framework that uses large language models (LLMs) to Automatically discover Effective and Diverse vulnerabilities in autonomous driving policies. We first utilize an LLM to automatically design reward functions for RL training. Then we let the LLM consider a diverse set of accident types and train adversarial policies for different accident types in parallel. Finally, we use preference-based learning to filter ineffective accidents and enhance the effectiveness of each vulnerability. Experiments across multiple simulated traffic scenarios and tested policies show that AED uncovers a broader range of vulnerabilities and achieves higher attack success rates compared with expert-designed rewards, thereby reducing the need for manual reward engineering and improving the diversity and effectiveness of vulnerability discovery. The implementation can be found on: <https://github.com/thunics/AED>.

INTRODUCTION

Autonomous driving is a safety-critical domain where even minor decision-making flaws can lead to severe accidents (Guo, Kurup, and Shah 2019). One approach to assessing the safety of autonomous driving policies is vulnerability discovery, which systematically uncovers scenarios where autonomous driving policy fails or behaves undesirably (Rastogi and Nygard 2020; Wan et al. 2022). By controlling the behaviors of other traffic participants, vulnerability discovery creates adversarial driving scenarios where the tested driving policy is exposed to hazardous conditions. Due to the rarity of such failure cases in real-world road tests, automated vulnerability discovery methods have been developed in simulation environments, where reinforcement learning (RL) emerges as an efficient approach that trains adversarial agents to induce accidents of the tested policy (Kutti, Fallah, and Bowden 2020; Feng et al. 2021; Chen et al. 2021; Mu et al. 2024).

However, there are two key challenges in leveraging RL

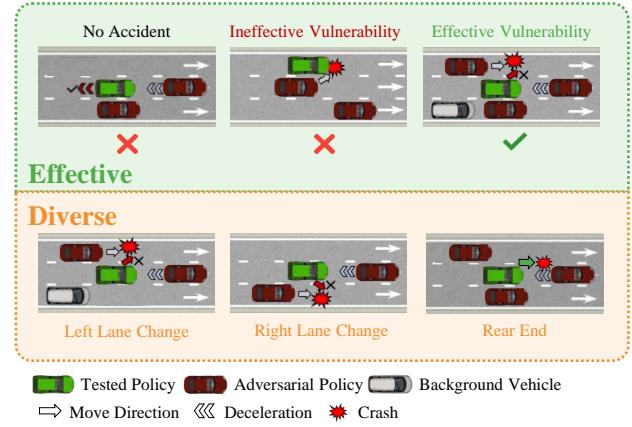


Figure 1: Illustration of effective and diverse vulnerabilities. An effective vulnerability is an accident that is caused by the false decision-making of the tested policy. Diverse vulnerabilities refer to different accident types that are consistent with human understanding and regulations.

for vulnerability discovery. First, it is hard to automatically identify **effective** vulnerabilities that require determining accident responsibility. When an accident occurs, it is essential to attribute the failure to the autonomous vehicle’s (AV) decision-making or those of other traffic participants. An effective vulnerability is one where the tested policy is responsible for the accident, revealing a genuine deficiency in its driving policy (Schulman et al. 2015; Corso et al. 2019; Mu et al. 2024). Accident responsibility, however, is complex to assess, often requiring human experts with domain knowledge in real-world settings (Shalev-Shwartz 2017; Sun et al. 2024). Similarly, in simulation-based RL training, designing a suitable reward function for responsibility attribution also demands substantial domain expertise (Kwon et al. 2023; Ma et al. 2023; Yu et al. 2025), posing a challenge to the automated discovery of effective vulnerabilities.

The second challenge in RL-based vulnerability discovery is ensuring **diverse** failure cases. Since a safety-critical flaw can arise from various accident types, it is crucial to comprehensively explore the space of possible failures (Corso et al. 2019; Koren and Kochenderfer 2020; Mu et al. 2024). Traditional RL methods, however, tend to converge on the most

probable failure case, leading to a lack of diversity in discovered vulnerabilities (Eysenbach et al. 2018). Existing approaches to diversity-aware RL generally fall into two types. The first type is optimizing a general diversity metric alongside task objectives. However, these methods often lead to vulnerabilities with uninterpretable differences that are not consistent with human understanding and regulations. The second type is incorporating heuristic-driven diversity measures tailored to specific tasks, which rely heavily on expert knowledge and struggle to generalize across different scenarios. These limitations hinder the interpretability and automation of diverse vulnerability discovery, leaving potential safety risks unsolved.

To address the aforementioned challenges, we propose **AED**, a framework that leverages large language models (LLMs) to Automatically discover Effective and Diverse vulnerabilities in autonomous driving policies. Our framework consists of three key steps. First, an LLM is prompted with the environment and task descriptions to automatically design reward functions for vulnerability discovery. Then, to enhance the diversity of vulnerabilities, we use the LLM to consider a broad range of accident scenarios and train adversarial policies for different types in parallel. Finally, to further improve the effectiveness of the discovered vulnerabilities, we use preference-based reinforcement learning to train a reward model that filters ineffective accidents. By combining automatic reward design, diverse accident generation, and preference-based effectiveness enhancement, our framework automatically identifies a broad spectrum of effective and diverse vulnerabilities, improving the comprehensiveness of AV safety evaluation.

To evaluate the performance of our framework, we apply our method to discover vulnerabilities of both planning-based policies and learning-based policies across different traffic scenarios, involving different numbers of adversarial vehicle agents. Experiment results show that our method can successfully identify and distinguish a diverse set of different vulnerabilities. Moreover, our method is able to achieve higher attack success rates with automatically generated rewards than policies with expert-designed rewards. These results show that our proposed framework can discover a diverse set of effective vulnerabilities in various tested policies, significantly reducing the reliance on manually designed reward functions and diversity metrics.

In summary, the contributions of this paper are: (1) We propose to use LLMs to automatically discover a diverse set of potential vulnerabilities, enhancing the overall safety of a driving policy. (2) We propose a preference-based learning method to filter ineffective accidents and encourage the identification of effective vulnerabilities. (3) We demonstrate our approach through extensive experiments and analysis across various traffic environments and tested driving policies, demonstrating its robustness and generalizability.

RELATED WORK

Diverse Vulnerability Discovery

Adversarial RL has been widely used to generate safety-critical scenarios that reveal vulnerabilities in autonomous

driving policies. Prior works use Deep Q-Network (Mnih et al. 2015) to generate corner cases (Karunakaran, Worrall, and Nebot 2020; Feng et al. 2021; Sun et al. 2021), and apply DDPG (Lillicrap et al. 2015) to create challenging lane-change conditions(Chen et al. 2021). Adaptive Stress Testing (Corso et al. 2019) incorporates a Responsibility-Sensitive Safety (RSS) (Shalev-Shwartz 2017) reward into TRPO (Schulman et al. 2015) to classify AV failures. Multi-agent extensions such as MADDPG (Lowe et al. 2017) and MAPPO (Yu et al. 2022) have been applied to coordinate multiple adversarial vehicles (Wachi 2019; Mu et al. 2024). However, these methods rely on crafted reward functions, which require domain expertise to balance responsibility attribution and behavioral realism.

To improve the diversity of discovered vulnerabilities, existing methods can be categorized into general and heuristic methods. General methods optimize information-theoretic objectives (Eysenbach et al. 2018) or apply trajectory similarity metrics (Liu and Schneider 2012; Corso et al. 2019), but often lack interpretability in real-world scenarios. Heuristic methods incorporate expert knowledge, such as clustering (Chen et al. 2021) or intrinsic novelty reward (Mu et al. 2024) to promote rare failures, but remain dependent on handcraft designs and domain-specific assumptions. Our work addresses these limitations by using LLMs to automatically generate diverse reward functions tailored to distinct vulnerability types, improving adaptability while reducing the reliance on manual heuristics.

Reward Design with LLMs

Reward design is a fundamental challenge in reinforcement learning, often requiring extensive domain expertise and iterative trial-and-error (Sutton 2018; Singh, Lewis, and Barto 2009). Recent advances leverage LLMs to automate reward design by exploiting their capabilities in reasoning, planning, and code generation (Shinn et al. 2023; Du et al. 2023; Carta et al. 2024). Some studies use LLMs to interpret environment and agent behaviors and assign intrinsic rewards based on scenario understanding (Kwon et al. 2023; Wu et al. 2024; Chu et al. 2023; Wang 2024), while others focus on generating executable reward code (Ma et al. 2023; Yu et al. 2025, 2023; Li et al. 2024). Notably, Eureka (Ma et al. 2023) proposes an automatic self-reflective framework where LLMs generate and iteratively refine reward functions using environment code and sparse human-designed evaluation rewards. ICPL (Yu et al. 2025) extends this to tasks without predefined rewards, enabling preference-based learning from human feedback. Our work extends ICPL to multi-agent traffic scenarios, where reward design must consider agent interactions and responsibility attribution.

Preference-Based Reinforcement Learning

Preference-based RL learns directly from expert preference rather than handcrafted numeric rewards, reducing the reliance on domain-specific knowledge. Prior works (Akrour, Schoenauer, and Sebag 2012; Wilson, Fern, and Tadepalli 2012; Wirth, Fürnkranz, and Neumann 2016; Mathewson and Pilarski 2017; Lee, Smith, and Abbeel 2021; Myers et al. 2022; Christiano et al. 2017) explore RL from human

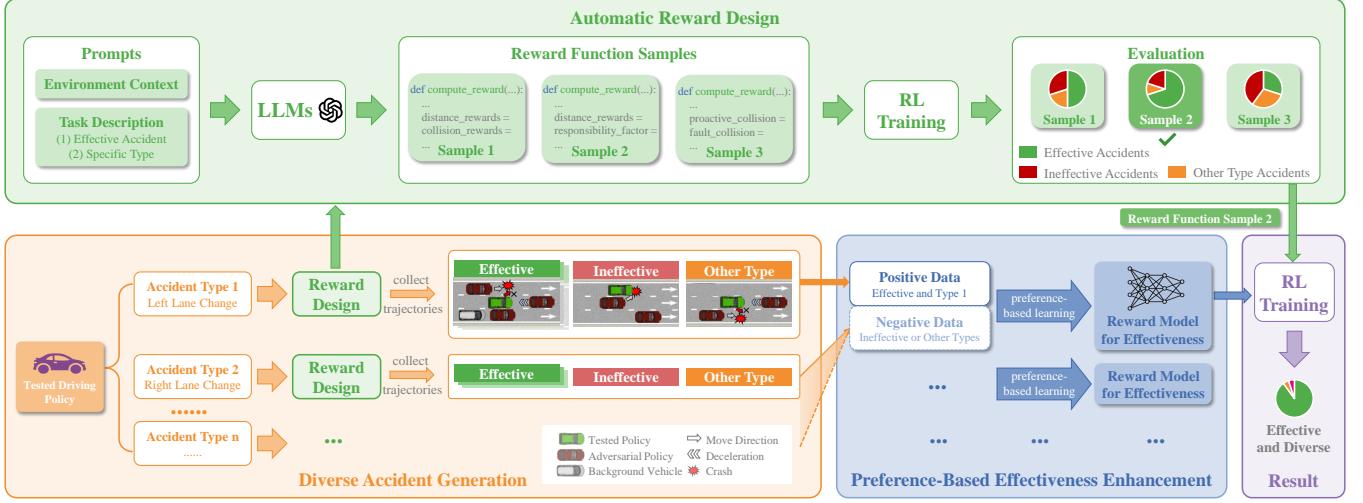


Figure 2: Overview of our proposed AED framework that uses large language models (LLMs) to automatically discover effective and diverse vulnerabilities in autonomous driving policies. Our framework first utilizes an LLM to automatically design reward for vulnerability discovery. Then we use the LLM to consider different accident types and generate a diverse set of accidents in parallel. Finally, we use preference-based learning to enhance the effectiveness of each accident type. Combining the three components leads to our framework for automatic, effective, and diverse vulnerability discovery.

feedback via trajectory preference queries or score ranking. However, querying human labels during online training is often costly and slow. To address this limitation, recent studies have proposed RL from AI feedback (Bai et al. 2022). Some leverage LLMs to provide preference based on vision-based trajectory or text descriptions of states and actions (Wang et al. 2024, 2025; Klissarov et al. 2023; Chu et al. 2023; Tu et al. 2024). Our work automatically generates preference data pairs from simulation outcomes produced by parallel adversarial policies, enabling scalable preference learning without numerous labeling while capturing diverse vulnerabilities in multi-agent interactions.

PRELIMINARY

Partially Observed Markov Decision Process

Our MARL framework is modeled as a Partially Observed Markov Decision Process (POMDP), defined by a tuple $\mathcal{M} = (\mathcal{N}, \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{P}, \mathcal{R}, \gamma)$, where $\mathcal{N} = \{1, 2, \dots, N\}$ is the set of agents, \mathcal{S} is the state space covering all possible states of the N agents, \mathcal{A} is the shared action space, \mathcal{O} denotes each agent's observation space, \mathcal{P} is the transition probability function, \mathcal{R} is the joint reward function with γ is the reward discount factor. At each step, agent i receives an observation $o_i = \mathcal{O}(s_i)$ where $s_i \in \mathcal{S}$ is the local state. The agent then takes an action a_i according to the shared policy $\pi_\theta(a_i|o_i)$, where θ represents the shared parameters. The next state progresses to s'_i according to the transition probability $p(s'_i|s_i, a_i) \in \mathcal{P}$ and the agent i receives a reward $r_i = \mathcal{R}(s_i, a_i)$. For homogeneous agents, each agent i optimizes the discounted cumulative rewards $J(\pi) = \mathbb{E}_{\tau \sim \pi} [\sum_t \gamma^t r_t]$, where a trajectory τ denotes a state-action-reward triplets sequence, i.e., $\tau = \{(s_t, a_t, r_t)\}_t$.

Diverse and Effective Vulnerability Discovery

Our objective is to discover diverse and effective vulnerabilities in a fixed, pre-trained autonomous driving policy by training adversarial policies within the previously defined POMDP framework. While the tested policy remains fixed, adversarial vehicles are optimized to induce failure scenarios that expose weaknesses in the tested policy. Specifically, the adversarial policy is optimized to maximize the rate of effective vulnerability, defined as those for which responsibility can be attributed to the tested policy. Diversity is assessed by the number of distinct vulnerability types, and effectiveness is assessed by the proportion of vulnerabilities discovered arising from wrong decisions of the tested driving policy. Formal definitions and evaluation procedures for these two metrics are provided in Sec. and Sec. .

METHODOLOGY

To automatically find effective and diverse vulnerabilities, we propose the AED framework that consists of three components including automatic reward design, diverse accident generation, and preference-based effectiveness enhancement. The overview of our framework is shown in Fig. 2. The first automatic reward design component prompts an LLM with the environment and task descriptions to automatically generate reward function samples and select the best one. Then, to enhance the diversity of vulnerabilities, we propose a diverse accident generation component that prompts the LLM to consider a broad range of accident types and trains adversarial policies for different types in parallel. Finally, to enhance the effectiveness of each vulnerability, we use preference-based learning to train a reward model that filters ineffective accidents and other accident types. Combining the three components lead to our framework that

automatically discovers effective and diverse vulnerabilities.

Automatic Reward Design

To enable automatic reward design for vulnerability discovery, we extend the In-Context Preference Learning (ICPL) framework (Yu et al. 2025) to leverage the code-synthesis and in-context learning capabilities of LLMs. In our adaptation, the LLM generates task-specific reward functions based on environment descriptions and human preferences, guiding the discovery of effective accidents in a self-improving manner.

The LLM first receives an environment description and a detailed task specification, then generates K executable reward functions. The environment text outlines the scenario and relevant code-level details, such as the agent’s observation/action space and key parameters or functions useful for reward construction. The specific details of the task description are provided in Sec. . Each reward function is used to train an agent, resulting in K agents, which are evaluated using human-written criteria for identifying effective accidents. To ensure comprehensive coverage, we also generate accident videos for human evaluation. Based on both metrics and visual inspection, evaluators select the best reward functions. The selected best function typically yields high accident effectiveness across scenarios and serves as a basis for further refinement.

Diverse Accident Generation

To enhance diversity in vulnerability discovery, the adversarial policy should capture distinct characteristics of various weaknesses in the tested policy. We introduce a Diverse Accident Generation component, which prompts LLMs to generate a set of reward functions that promote vulnerability diversity. By considering a wide range of accident types, LLMs automatically design multiple reward functions in parallel, each tailored to induce specific failure cases.

The process begins with the LLM generating a task description for a given accident type based on environment specifications. This description includes a high-level summary of the targeted accident scenario, concrete examples, justifications for attributing responsibility to the tested agent, and criteria for categorizing scenarios within this accident type. Using the component outlined in Sec. , the LLM then autonomously generates reward functions corresponding to different accident types in parallel.

Once reward functions are generated, they are integrated into the training of adversarial policies, each optimized for a specific vulnerability of the tested policy. These adversarial policies are then used to collect trajectory data, creating a structured dataset covering diverse failure cases. This dataset serves as a foundation for further analysis and improvement of the tested policy, ensuring a comprehensive evaluation of its robustness across various accident scenarios.

The prompts used to guide the LLM in generating diverse accident types, along with the corresponding environment and scenario descriptions, are provided in Appendices.

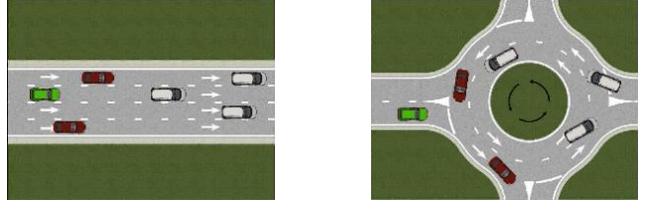


Figure 3: Demonstration of “Highway” and “Roundabout”.

Preference-Based Effectiveness Enhancement

Reward functions derived from Sec. for a specific accident type are often imperfect, sometimes capturing unintended accident types or ineffective scenarios. For instance, the reward function for Left Lane Change (Accident Type 1) may capture Right Lane Change (Accident Type 2) and Adversarial Vehicle cut in mistakenly (Accident Type 3). To improve the effectiveness and distinguish diverse accident types, we propose a preference-based effectiveness enhancement component to refine the LLMs-driven reward function. The key idea is to learn a new reward function \hat{r} that prefers trajectories corresponding to the intended accident types over other types. Specifically, \hat{r} is trained to assign higher cumulative rewards to Type 1 than to those from Type 2 or Type 3.

To construct training data for preference learning, we leverage trajectory data generated in parallel for different accident types using their respective LLMs-driven reward functions. We consider the Effective and Type 1 dataset as Positive Data σ^1 , and use a combination of the Ineffective or Other Types (Type 2 and Type 3) as Negative Data σ^2 .

The probability of preferring a trajectory σ^i as the intended accident type is assumed to depend exponentially on the accumulated latent reward over the trajectory’s state-action sequence (o_t^i, a_t^i) :

$$\hat{P}[\sigma^1 \succ \sigma^2] = \frac{\exp \sum_t \hat{r}(o_t^1, a_t^1)}{\exp \sum_t \hat{r}(o_t^1, a_t^1) + \exp \sum_t \hat{r}(o_t^2, a_t^2)}. \quad (1)$$

To optimize the reward function \hat{r} , we minimize the cross-entropy loss between these predictions and the intended preference. The Positive Data and Negative Data are stored in a database \mathcal{D} of triples $(\sigma^1, \sigma^2, \mu)$, where μ is a judgment label (0 or 1) to indicate whether σ^1 is the preferred accident type. The loss function is defined as:

$$\text{loss}(\hat{r}) = - \sum_{(\sigma^1, \sigma^2, \mu) \in \mathcal{D}} \mu \log \hat{P}[\sigma^1 \succ \sigma^2] + (1 - \mu) \log(1 - \hat{P}[\sigma^1 \succ \sigma^2]). \quad (2)$$

Since μ assigns full probability mass to the Positive Data, which means $\mu = 1$ across the dataset, the loss simplifies to:

$$\text{loss}(\hat{r}) = - \sum_{(\sigma^1, \sigma^2, \mu) \in \mathcal{D}} \log \hat{P}[\sigma^1 \succ \sigma^2]. \quad (3)$$

In this way, the reward function \hat{r} learns to favor Type 1 accident trajectories while effectively distinguishing them

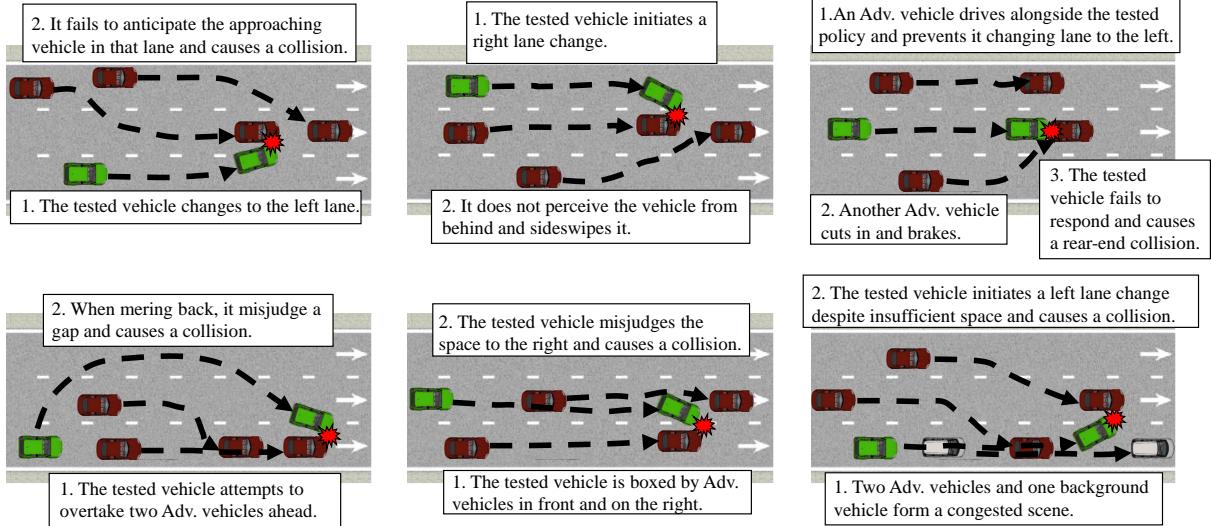


Figure 4: Examples of distinct vulnerability types discovered by AED. Adv. is abbreviation for “adversarial”.

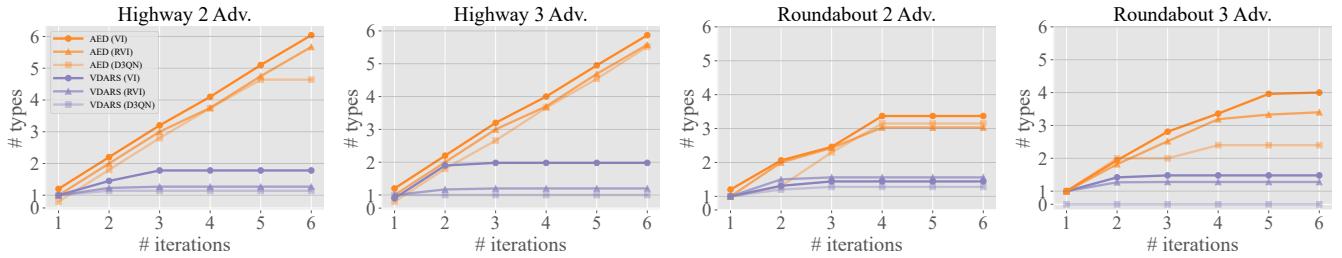


Figure 5: Number of vulnerability types discovered across different environments, tested policies, and numbers of adversarial vehicle agents. Each line represents the cumulative number over six training iterations, where the first discovered type is counted as 1, and subsequent new types are weighted by their relevant frequency among effective failures. AED consistently discovers more diverse vulnerabilities than VDARS.

from other types. We then incorporate \hat{r} and the original LLMs-driven reward function into RL training. If \hat{r} identifies trajectories as Type 2, the training will be penalized. This guides RL training to prioritize Accident Type 1 and ultimately improve the attack success rate for this category.

EXPERIMENT

We evaluate AED in two driving environments “Highway” and “Roundabout” using the Highway simulator (Leurent et al. 2018), as shown in Fig. 3. We consider three types of tested driving policy: two Planning-based methods, Value Iteration (VI) (Bellman 1966) and Robust Value Iteration (RVI) (Nilim and El Ghaoui 2004), and one RL-based method, Dueling Double Q-Network (D3QN) (Wang et al. 2016). To account for the complexity of traffic interactions, we conduct experiments with two and three adversarial vehicles in each environment.

Implementation Details

We categorize vulnerabilities into three types based on the relative positions of the tested and adversarial vehicles: **Left Lane Change (Left)**, **Right Lane Change (Right)**, and **Rear End (Rear)** collision scenarios. These categories correspond to different failure-inducing situations. For later diversity analysis, we further divide these categories into finer-grained subtypes based on common real-world traffic accident patterns. Each adversarial policy is trained using the MAPPO algorithm (Yu et al. 2022) for 10 million steps. We utilize GPT-4o and DeepSeek-R1 for diverse accident generation. Detailed environment settings and training parameters are provided in the Appendices.

We compare the performance of our method with the following baselines:

- **VDARS** (Mu et al. 2024): A multi-agent RL framework that employs handcrafted reward functions to learn the adversarial policy for vulnerability discovery.
- **Parallel Eureka**: An ablated version of our method without the preference-based effectiveness enhancement

module. This can also be regarded as a parallel version of Eureka for different vulnerability types.

Evaluation of Diversity

To evaluate AED’s ability to discover diverse vulnerability scenarios, we measure the number of distinct vulnerability types identified across multiple training iterations. We divide the three high-level collision categories into six fine-grained subtypes to enable more precise diversity assessment. Illustrative examples discovered in the Highway environment are shown in Fig. 4. A complete set of definitions and additional illustrative examples can be found in the Appendices. For each setting—defined by the environment (Highway and Roundabout), the tested policy (VI, RVI, and D3QN), and the number of adversarial vehicle agents (2 and 3)—we conduct six independent training iterations with different random seeds. In each iteration, we collect all effective failure cases. In the first iteration, we identify the most frequently occurring effective vulnerability type and assign it a count of 1. For each subsequent iteration, we exclude all previously discovered types and select the most frequent type among the remaining ones. Instead of simply incrementing the count by one, we add its proportion among all effective failures in that iteration to the cumulative count. This strategy ensures that each iteration contributes at most one new type, and the final diversity score reflects both the variety and prominence of distinct vulnerabilities uncovered. In VDARS, all iterations use the same handcrafted reward function. In contrast, AED trains the adversarial policy in each iteration with a different LLMs-generated reward function targeted at distinct failure types.

As shown in Fig. 5, AED consistently discovers more diverse vulnerability types than VDARS. Across all settings, AED reaches or approaches the maximum of six types within six iterations, while VDARS typically uncovers only two or three. This discrepancy highlights the limited exploration capability of a single handcrafted reward and AED’s strength in promoting broader vulnerability discovery through multiple LLMs-generated reward functions.

Evaluation of Effectiveness

We evaluate the effectiveness of the proposed preference-based enhancement module through both qualitative and quantitative analyses. These evaluations demonstrate its utility by: (1) assessing the learned reward model’s ability to distinguish goal-aligned vulnerabilities from unrelated failures, and (2) measuring its impact on the overall effective vulnerability rate when integrated into RL training.

Qualitative Evaluation We first illustrate the reward model’s discriminative capability using a representative case from the Highway environment with three adversarial vehicle agents and a D3QN tested policy. When trained with a LLMs-generated reward function targeting Left Lane Change (Left) collisions, the resulting accidents include both targeted Left collisions, unintended Rear-End collisions, and other ineffective failures. After applying preference-based reward learning, the learned model is evaluated on different accident trajectories. As shown in

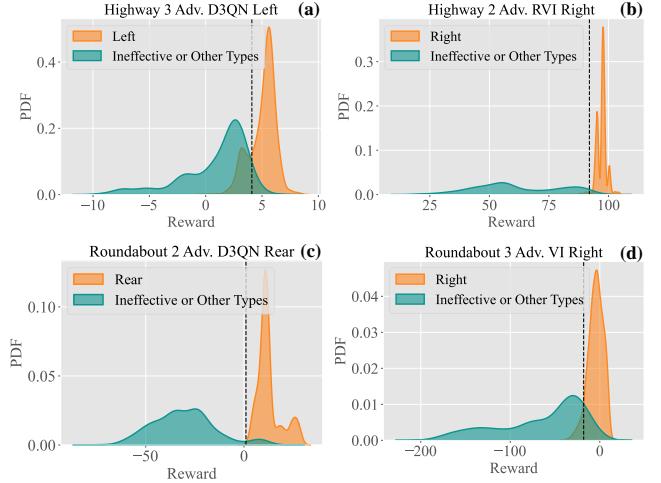


Figure 6: Probability density distribution of reward for accident trajectories across different environments and policy configurations. In all cases, the learned reward model assigns higher scores to goal-aligned failures (orange), effectively separating them from unrelated accidents (green).

Fig. 6(a), it assigns higher reward scores to Left collisions (orange) than to unrelated accidents (green), demonstrating clear separation. Similar reward distributions are observed across other environments and policies (Fig. 6(b)(c)(d)), further validating the model’s ability to distinguish goal-aligned vulnerabilities from irrelevant ones.

Quantitative Evaluation To quantify the benefit of preference-based enhancement, we measure the **effective vulnerability rate**—defined as the percentage of simulation runs where the adversarial policy successfully triggers a specific failure type attributable to the tested policy. We compare the full AED framework against a baseline ablation, Parallel Eureka, which uses LLMs-generated reward functions without the final enhancement module. Parallel Eureka can be seen as a variant of VDARS, replacing a fixed hand-crafted reward function with static LLMs-generated ones. In AED, if the learned reward model assigns a score below a predefined threshold (black dashed lines in Fig. 6), the corresponding LLMs-generated reward is suppressed to zero, thus filtering out off-target behaviors.

The effective vulnerability rate reported corresponds to the maximum among all fine-grained subtypes in each high-level collision category. Results for all subtypes are provided in the Appendices. As shown in Fig. 7, AED consistently outperforms or matches Parallel Eureka in nearly all settings. The improvement is especially notable where the initial LLMs-generated reward is suboptimal. For example, in the Highway environment with a D3QN-driven tested policy and three adversarial vehicles, AED increases the effective Left collision rate from 32.22% to 88.89%. This substantial improvement demonstrates the impact of preference-based enhancement in guiding adversarial vehicles to consistently induce failure scenarios that the tested policy struggles with. Even in cases where Parallel Eureka already performs well,

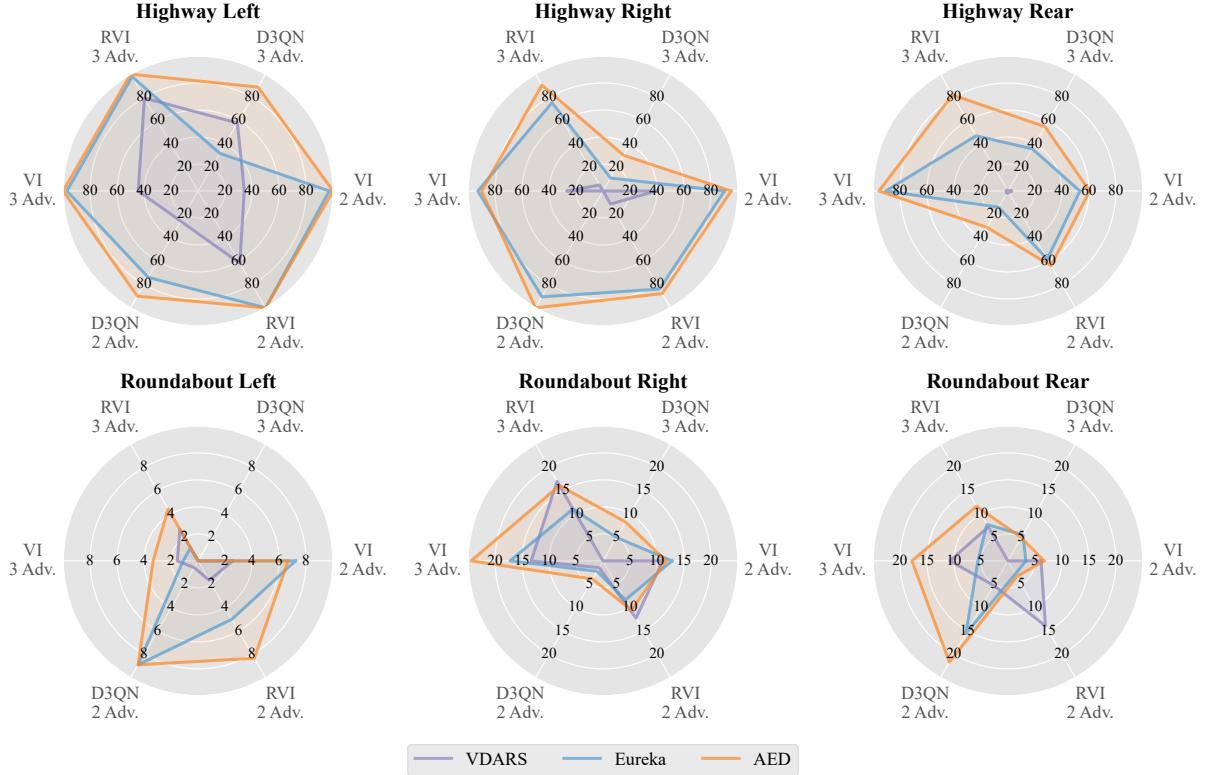


Figure 7: Effective Vulnerability Rates (%) across different environments, tested policies, and numbers of adversarial vehicle agents. Each radar chart corresponds to one environment with one high-level accident category. Adv. is abbreviation for “adversarial”. AED consistently achieves the highest effective vulnerability rates.

AED maintains or improves performance, indicating that the refinement module enhances reward fidelity without degrading strong initial signals.

Overall Performance

AED consistently outperforms VDARS in both diversity and effectiveness. It approaches a maximum of six distinct vulnerability types in most settings. It also consistently achieves significantly higher effective vulnerability rates (Fig. 7). These gains stem from AED’s design: its parallel accident generation enables broader exploration of failure modes, while the preference-based enhancement module refines noisy LLMs-generated rewards into precision training objects. Together, they make AED a more comprehensive framework for adversarial vulnerability discovery.

Interpretability and Discussion

To illustrate how AED transforms high-level vulnerability descriptions into effective adversarial policies, we present a case study in Fig. 8, targeting Left Lane Change collisions in the Highway environment with a D3QN-driven tested policy and three adversarial vehicles.

Task Specification and Reward Generation We begin by prompting the LLMs with relevant environment context (e.g., road layout, lane width, vehicle length, and action

set) and task description, including the targeted vulnerability type (“Left Lane Change Collisions”), expected behaviors (e.g., “the tested AV unexpectedly changes lanes to the left...”), and criteria for assigning responsibility. Based on this, the LLM generates K ($K = 6$ in our experiments) executable reward functions. Each candidate function is used to train an adversarial policy, and the best-performing one—measured by the proportion of Left Lane Change collisions in simulation runs—is selected.

Reward Function Analysis The selected reward function is shown in Fig. 8(b). It consists of modular checks for: (1) whether the tested vehicle performed a left lane change before collision, (2) whether adversarial vehicles behaved non-aggressively (e.g., within steering angle and acceleration constraints), and (3) whether the collision can be attributed to the tested vehicle. The differences between the LLMs-generated reward and the expert-crafted version are visualized using a diff-style format. While both functions share a similar core structure—detecting collisions and evaluating adversarial passiveness—the expert reward relies on a simplified branching logic that only distinguishes between lane change and rear-end collisions. In contrast, the LLMs-generated reward introduces two key advantages. First, it explicitly encodes the targeted collision type (Left Lane Change) as part of the reward condition. Second, it exam-

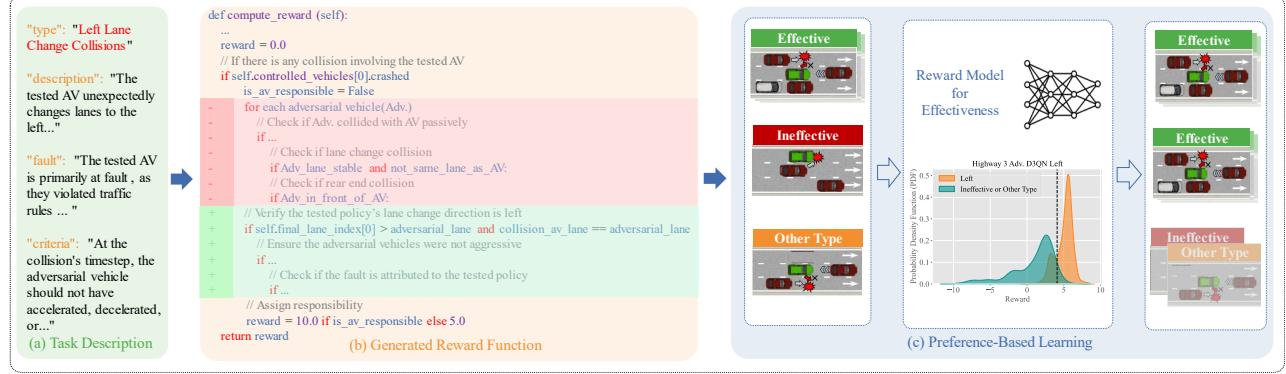


Figure 8: Case study demonstrating the AED framework. (a) Task description provided to LLMs specifies the targeted Left Lane Change vulnerability, scenario details, and responsibility criteria. (b) The LLMs-generated reward function verifies the tested policy’s maneuver, ensures non-aggressive behavior from adversarial vehicles, and assigns responsibility accordingly. Differences from the expert-crafted reward are highlighted using a diff-style format. (c) The preference-based reward model differentiates targeted failure scenarios from unintended ones.

ines both the current and previous actions of the AV to more accurately determine responsibility. These improvements ensure that rewarded collisions are not only aligned with the intended failure type, but also effectively attributed to decision-making flaws in the tested policy.

Preference-based Enhancement Initial adversarial training with the LLMs-generated reward function yields a sub-optimal distribution of accident types: Left Lane Change collisions (36.67%), Rear-End collisions (1.00%), and ineffective accidents (62.33%). As shown in Fig. 8(c), the preference-based model corrects this by assigning higher rewards to targeted Left Lane Change collisions and lower rewards to unintended or ineffective accidents. During RL training, the preference model is integrated with the original LLMs-generated reward function. If the preference score for a trajectory during training falls below a predefined threshold, the LLMs-generated reward is suppressed to zero. This filtering mechanism shifts the training focus towards high-quality, goal-aligned failures, resulting in a substantial increase in effective Left Lane Change collisions to 88.89%.

CONCLUSION

We propose AED, an LLMs-based framework that uses reinforcement learning to automatically discover effective and diverse vulnerabilities of autonomous driving policies. We use an LLM to automatically design reward functions without the need of human experts. To enhance the diversity of vulnerabilities, we use the LLM to consider a broad range of potential accidents and train adversarial policies using RL in parallel. We further improve the effectiveness of each vulnerability by training a reward model with preference-based learning to filter ineffective or other types of accidents. Experiments on various driving scenarios with both planning-based policies and learning-based policies show that AED discovers a diverse range of vulnerabilities and achieves higher attack success rates compared to existing methods, showing the potential of using LLMs to reduce the need for

manual reward engineering and improve the comprehensiveness of autonomous driving safety evaluation.

References

- Akrour, R.; Schoenauer, M.; and Sebag, M. 2012. April: Active preference learning-based reinforcement learning. In *ECML PKDD*, 116–131.
- Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.
- Bellman, R. 1966. Dynamic programming. *science*, 153(3731): 34–37.
- Carta, T.; Romac, C.; Wolf, T.; Lamprier, S.; Sigaud, O.; and Oudeyer, P.-Y. 2024. Grounding Large Language Models in Interactive Environments with Online Reinforcement Learning. *arXiv:2302.02662*.
- Chen, B.; Chen, X.; Wu, Q.; and Li, L. 2021. Adversarial evaluation of autonomous vehicles in lane-change scenarios. *IEEE T-ITS*, 23(8): 10333–10342.
- Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *NeurIPS*, 30.
- Chu, K.; Zhao, X.; Weber, C.; Li, M.; and Wermter, S. 2023. Accelerating reinforcement learning of robotic manipulations via feedback from large language models. *arXiv preprint arXiv:2311.02379*.
- Corso, A.; Du, P.; Driggs-Campbell, K.; and Kochenderfer, M. J. 2019. Adaptive stress testing with reward augmentation for autonomous vehicle validation. In *IEEE ITSC*.
- Du, Y.; Watkins, O.; Wang, Z.; Colas, C.; Darrell, T.; Abbeel, P.; Gupta, A.; and Andreas, J. 2023. Guiding Pre-training in Reinforcement Learning with Large Language Models. *arXiv:2302.06692*.
- Eysenbach, B.; Gupta, A.; Ibarz, J.; and Levine, S. 2018. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*.

- Feng, S.; Yan, X.; Sun, H.; Feng, Y.; and Liu, H. X. 2021. Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment. *Nature communications*, 12(1): 748.
- Guo, J.; Kurup, U.; and Shah, M. 2019. Is it safe to drive? An overview of factors, metrics, and datasets for driveability assessment in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 21(8): 3135–3151.
- Karunakaran, D.; Worrall, S.; and Nebot, E. 2020. Efficient statistical validation with edge cases to evaluate highly automated vehicles. In *IEEE ITSC*.
- Klissarov, M.; D’Oro, P.; Sodhani, S.; Raileanu, R.; Bacon, P.-L.; Vincent, P.; Zhang, A.; and Henaff, M. 2023. Motif: Intrinsic motivation from artificial intelligence feedback. *arXiv preprint arXiv:2310.00166*.
- Koren, M.; and Kochenderfer, M. J. 2020. Adaptive stress testing without domain heuristics using go-explore. In *IEEE ITSC*.
- Kuutti, S.; Fallah, S.; and Bowden, R. 2020. Training adversarial agents to exploit weaknesses in deep control policies. In *IEEE ICRA*.
- Kwon, M.; Xie, S. M.; Bullard, K.; and Sadigh, D. 2023. Reward design with language models. *arXiv preprint arXiv:2303.00001*.
- Lee, K.; Smith, L.; and Abbeel, P. 2021. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. *arXiv preprint arXiv:2106.05091*.
- Leurent, E.; et al. 2018. An environment for autonomous driving decision-making.
- Li, H.; Yang, X.; Wang, Z.; Zhu, X.; Zhou, J.; Qiao, Y.; Wang, X.; Li, H.; Lu, L.; and Dai, J. 2024. Auto MC-Reward: Automated Dense Reward Design with Large Language Models for Minecraft. *arXiv:2312.09238*.
- Lillicrap, T. P.; Hunt, J. J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; and Wierstra, D. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.
- Liu, H.; and Schneider, M. 2012. Similarity measurement of moving object trajectories. In *ACM SIGSPATIAL GeoStreaming*, 19–22.
- Lowe, R.; Wu, Y. I.; Tamar, A.; Harb, J.; Pieter Abbeel, O.; and Mordatch, I. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. *NeurIPS*, 30.
- Ma, Y. J.; Liang, W.; Wang, G.; Huang, D.-A.; Bastani, O.; Jayaraman, D.; Zhu, Y.; Fan, L.; and Anandkumar, A. 2023. Eureka: Human-level reward design via coding large language models. *arXiv preprint arXiv:2310.12931*.
- Mathewson, K. W.; and Pilarski, P. M. 2017. Actor-critic reinforcement learning with simultaneous human control and feedback. *arXiv preprint arXiv:1703.01274*.
- Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533.
- Mu, Y.; Liu, W.; Yu, C.; Ning, X.; Cao, Z.; Xu, Z.; Liang, S.; Yang, H.; and Wang, Y. 2024. Multi-Agent Vulnerability Discovery for Autonomous Driving Policy by Finding AV-Responsible Scenarios. In *IEEE CASE*.
- Myers, V.; Biyik, E.; Anari, N.; and Sadigh, D. 2022. Learning multimodal rewards from rankings. In *CoRL*, 342–352.
- Nilim, A.; and El Ghaoui, L. 2004. *Robust markov decision processes with uncertain transition matrices*. Ph.D. thesis, University of California, Berkeley.
- Rastogi, A.; and Nygard, K. E. 2020. Threats and Alert Analytics in Autonomous Vehicles. In *CATA*, 48–59.
- Schulman, J.; Levine, S.; Abbeel, P.; Jordan, M.; and Moritz, P. 2015. Trust region policy optimization. In *ICML*, 1889–1897.
- Shalev-Shwartz, S. 2017. On a formal model of safe and scalable self-driving cars. *arXiv preprint arXiv:1708.06374*.
- Shinn, N.; Cassano, F.; Berman, E.; Gopinath, A.; Narasimhan, K.; and Yao, S. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. *arXiv:2303.11366*.
- Singh, S.; Lewis, R. L.; and Barto, A. G. 2009. Where do rewards come from. In *CogSci*, 2601–2606.
- Sun, H.; Feng, S.; Yan, X.; and Liu, H. X. 2021. Corner case generation and analysis for safety assessment of autonomous vehicles. *Transportation research record*, 2675(11): 587–600.
- Sun, H.; Poskitt, C. M.; Sun, Y.; Sun, J.; and Chen, Y. 2024. ACAV: a framework for automatic causality analysis in autonomous vehicle accident recordings. In *ICSE*, 1–13.
- Sutton, R. S. 2018. Reinforcement learning: An introduction. *A Bradford Book*.
- Tu, S.; Sun, J.; Zhang, Q.; Lan, X.; and Zhao, D. 2024. Online Preference-based Reinforcement Learning with Self-augmented Feedback from Large Language Model. *arXiv preprint arXiv:2412.16878*.
- Wachi, A. 2019. Failure-scenario maker for rule-based agent using multi-agent adversarial reinforcement learning and its application to autonomous driving. *arXiv preprint arXiv:1903.10654*.
- Wan, Z.; Shen, J.; Chuang, J.; Xia, X.; Garcia, J.; Ma, J.; and Chen, Q. A. 2022. Too afraid to drive: systematic discovery of semantic dos vulnerability in autonomous driving planning under physical-world attacks. *arXiv preprint arXiv:2201.04610*.
- Wang, R.; Zhao, D.; Yuan, Z.; Obi, I.; and Min, B.-C. 2025. Prefclm: Enhancing preference-based reinforcement learning with crowdsourced large language models. *IEEE RAL*.
- Wang, Y.; Sun, Z.; Zhang, J.; Xian, Z.; Biyik, E.; Held, D.; and Erickson, Z. 2024. Rl-vlm-f: Reinforcement learning from vision language foundation model feedback. *arXiv preprint arXiv:2402.03681*.
- Wang, Z. 2024. Towards Socially and Morally Aware RL agent: Reward Design With LLM. *arXiv:2401.12459*.
- Wang, Z.; Schaul, T.; Hessel, M.; Hasselt, H.; Lanctot, M.; and Freitas, N. 2016. Dueling network architectures for deep reinforcement learning. In *ICML*, 1995–2003.

Wilson, A.; Fern, A.; and Tadepalli, P. 2012. A bayesian approach for policy learning from trajectory preference queries. *NeurIPS*, 25.

Wirth, C.; Fürnkranz, J.; and Neumann, G. 2016. Model-free preference-based reinforcement learning. In *AAAI*, volume 30.

Wu, Y.; Fan, Y.; Liang, P. P.; Azaria, A.; Li, Y.; and Mitchell, T. M. 2024. Read and Reap the Rewards: Learning to Play Atari with the Help of Instruction Manuals. arXiv:2302.04449.

Yu, C.; Tan, Q.; Lu, H.; Gao, J.; Yang, X.; Wang, Y.; Wu, Y.; and Vinitsky, E. 2025. Large Language Models are In-context Preference Learners. arXiv:2410.17233.

Yu, C.; Velu, A.; Vinitsky, E.; Gao, J.; Wang, Y.; Bayen, A.; and Wu, Y. 2022. The surprising effectiveness of ppo in cooperative multi-agent games. *NeurIPS*, 35: 24611–24624.

Yu, W.; Gileadi, N.; Fu, C.; Kirmani, S.; Lee, K.-H.; Arenas, M. G.; Chiang, H.-T. L.; Erez, T.; Hasenclever, L.; Humplik, J.; Ichter, B.; Xiao, T.; Xu, P.; Zeng, A.; Zhang, T.; Heess, N.; Sadigh, D.; Tan, J.; Tassa, Y.; and Xia, F. 2023. Language to Rewards for Robotic Skill Synthesis. arXiv:2306.08647.

Appendices

MAPPO Training

Observation and Action Space In our experiments, all agents share the same observation and action space. Each agent receives an observation vector $o_i = [k_0^i, k_1^i, k_2^i, k_3^i, k_4^i]$, where each k_j^i represents the state of one of the five nearest neighboring vehicles. Each k consists of five features $k = [p, x, y, v_x, v_y]$ as detailed in Table. 1. Each agent selects from a discrete action space consisting of five maneuver options, as shown in Table. 2.

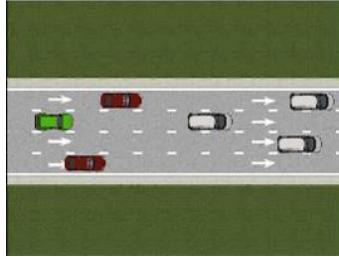
Feature	Description	Range
p	Binary indicator of vehicle presence	{0, 1}
x	Relative longitudinal position	[-150, 150]m
y	Relative lateral position	[-16, 16]m
v_x	Relative velocity along x-axis	[-60, 60]m/s
v_y	Relative velocity along y-axis	[-60, 60]m/s

Table 1: Observation space in MAPPO training.

Index	Description
0	Change to the left lane
1	Maintain current lane and speed
2	Change to the right lane
3	Accelerate
4	Decelerate

Table 2: Discrete action space available to agents.

Environment Setting To account for diverse and realistic traffic scenarios, we conduct experiments in two representative simulation environments: ‘Highway’ and ‘Roundabout’ as demonstrated in Fig. 9. *Highway* scenarios involve high-speed interactions, frequent lane changes, rear-end collisions and overtake collisions, which are typical in structured multi-lane traffic. *Roundabout* scenarios involve frequent turning, right-of-way decisions and merging behaviors. The key configuration parameters are summarized in Table. 3, which define the road structure, background vehicle behavior, and task-specific constraints.



(a) Highway



(b) Roundabout

Figure 9: Demonstration of ‘Highway’ and ‘Roundabout’.

Parameter	Value
Number of Lanes	4
Number of Background Vehicles	5
Background Vehicle Control Policy	Value Iteration
Simulation Frequency	5Hz
Max Task Length	20

Table 3: Environment Configurations used in MAPPO training.

Training Hyperparameters Table 4 summarizes the hyperparameters used in our MAPPO training for adversarial vehicle agents. These configurations are kept consistent across all environments.

Illustrations of Discovered Vulnerability Scenarios

To better reflect real-world traffic incident patterns, we further decompose the three high-level categories—Left Lane Change, Right Lane Change, and Rear End collisions—into six fine-grained subtypes. These subtypes enable a more detailed assessment of vulnerability diversity:

- Left-Sparse: The tested vehicle performs a left lane change without properly estimating the distance of an approaching adversarial vehicle in the adjacent lane, resulting in a sideswipe collision in sparse traffic.

Hyperparameter	Value
Network Architecture	MLP+RNN
Hidden Layer Size	64
Initial Learning Rate	5e-4
Discount Factor	0.99
GAE Parameter	0.95
Gradient clipping	0.5
Value Loss Coefficient	0.5
Entropy Coefficient	0.01
PPO Clipping Parameter	0.2
PPO Epochs	15
Training Steps	10M
Rollout Threads	100
Max Episode Length	40

Table 4: Hyperparameters used in MAPPO training.

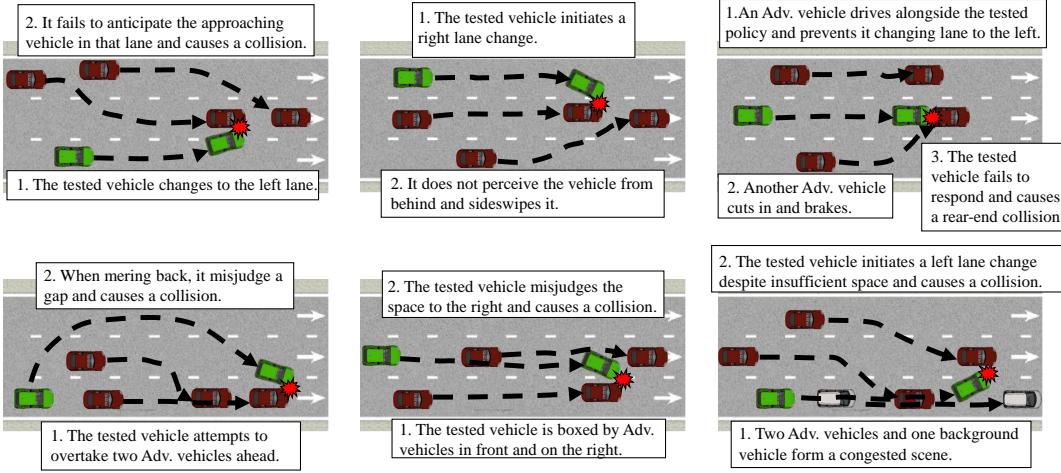


Figure 10: Examples of distinct vulnerability types discovered by AED. Adv. is abbreviation for “adversarial”.

- Right-Sparse: The tested policy executes a right lane change while misjudging the lateral distance to a nearby adjacent vehicle, causing under sparse traffic conditions.
- Rear End: The tested vehicle fails to maintain a safe following distance or reacts too slowly, resulting in a rear-end collision with an adversarial vehicle in the same lane.
- Overtake: The tested policy attempts to overtake a slower adversarial vehicle ahead, but misjudges the timing and crashes during the maneuver, typically due to its failure to account for adversarial vehicles blocking the return lane.
- Left-Dense: When facing congestion caused by an adversarial vehicle ahead, the tested policy attempts a left lane change but collides with another adversarial vehicle in the adjacent lane.
- Right-Dense: Similarly, due to congestion caused by a leading adversarial vehicle, the tested policy attempts a right lane change and causes a sideswipe collision.

Fig. 10 shows example trajectories for each of the above vulnerability subtypes in the Highway Environment. In the Roundabout environment, overtaking accidents are really rare in real-world scenarios and are therefore excluded from our analysis. Instead, we identify a distinct type of vulnerability unique to this setting as shown in Fig. 11:

- T-Bone: The tested vehicle misjudges the right-of-way when entering the roundabout, and fails to yield to an adversarial vehicle already circulating within it. This results in a perpendicular collision where the front of the tested vehicle crashes into the side of another.

Full Breakdown of Effective Vulnerability Rates

To complement the main text, we provide detailed results for all fine-grained vulnerability subtypes across environments, tested policies, and adversarial vehicles. The full breakdown, as presented in Table 5, highlights that AED outperforms or matches

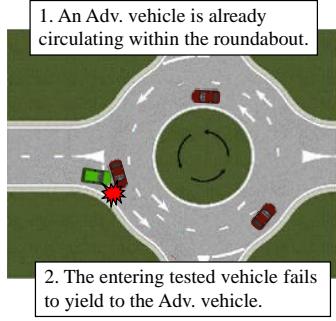


Figure 11: Example of T-Bone collision.

		2 Adversaries			3 Adversaries		
		VI	RFI	D3QN	VI	RFI	D3QN
Highway	Left-Sparse	VDARS	34.24% _(0.09)	61.88% _(0.52)	25.81% _(0.07)	44.38% _(0.14)	79.51% _(0.10)
		Eureka	97.22% _(0.08)	100% _(0.00)	73.89% _(0.55)	97.22% _(0.08)	97.78% _(0.08)
		AED	100% _(0.00)	100% _(0.00)	90.00% _(0.04)	100% _(0.00)	100% _(0.00)
	Right-Sparse	VDARS	39.56% _(0.11)	11.18% _(0.09)	0.00% _(0.00)	26.92% _(0.08)	5.13% _(0.01)
		Eureka	90.00% _(0.00)	83.89% _(0.28)	90.56% _(0.39)	92.78% _(0.48)	75.56% _(0.31)
		AED	95.00% _(0.04)	87.77% _(0.03)	100% _(0.00)	90.00% _(0.01)	90.56% _(0.10)
	Rear	VDARS	2.28% _(0.03)	1.49% _(0.01)	1.94% _(0.01)	1.45% _(0.01)	0.85% _(0.01)
		Eureka	53.33% _(0.72)	58.33% _(0.36)	13.89% _(0.15)	90.56% _(0.44)	47.22% _(0.39)
		AED	60.56% _(0.05)	63.89% _(0.06)	31.11% _(0.06)	95.60% _(0.01)	82.22% _(0.10)
Roundabout	Overtake	VDARS	0.00% _(0.00)				
		Eureka	46.67% _(0.28)	48.33% _(0.57)	1.67% _(0.01)	51.67% _(0.15)	51.67% _(0.07)
		AED	56.11% _(0.01)	51.11% _(0.05)	1.00% _(0.01)	56.00% _(0.07)	58.89% _(0.08)
	Left-Dense	VDARS	0.00% _(0.00)				
		Eureka	35.00% _(0.28)	40.00% _(0.36)	8.33% _(0.15)	58.33% _(0.23)	45.00% _(0.86)
		AED	40.56% _(0.07)	36.11% _(0.02)	8.00% _(0.02)	88.33% _(0.03)	78.89% _(0.05)
	Right-Dense	VDARS	0.00% _(0.00)				
		Eureka	70.00% _(0.23)	75.00% _(0.05)	20.00% _(0.07)	28.33% _(0.23)	75.00% _(0.34)
		AED	76.11% _(0.03)	75.00% _(0.00)	35.56% _(0.08)	61.11% _(0.43)	82.22% _(0.05)
T-Bone	Left-Sparse	VDARS	2.57% _(0.02)	1.67% _(0.01)	0.58% _(0.01)	1.54% _(0.01)	2.70% _(0.03)
		Eureka	7.22% _(0.08)	5.00% _(0.08)	8.89% _(0.20)	1.11% _(0.15)	1.11% _(0.07)
		AED	6.67% _(0.01)	8.33% _(0.03)	8.89% _(0.01)	3.33% _(0.01)	4.44% _(0.01)
	Right-Sparse	VDARS	11.07% _(0.07)	12.23% _(0.07)	1.46% _(0.03)	13.31% _(0.02)	17.01% _(0.19)
		Eureka	12.78% _(0.28)	8.33% _(0.13)	2.22% _(0.08)	17.22% _(0.61)	11.11% _(0.15)
		AED	11.67% _(0.05)	10.56% _(0.01)	3.89% _(0.01)	24.44% _(0.04)	16.11% _(0.03)
	Rear	VDARS	6.13% _(0.04)	13.90% _(0.08)	5.26% _(0.09)	10.74% _(0.01)	7.29% _(0.08)
		Eureka	3.33% _(0.36)	2.78% _(0.08)	15.56% _(0.43)	5.56% _(0.20)	7.78% _(0.20)
		AED	6.67% _(0.01)	3.33% _(0.03)	21.67% _(0.02)	17.78% _(0.05)	11.67% _(0.04)
Left-Dense	T-Bone	VDARS	0.00% _(0.00)				
		Eureka	8.00% _(0.04)	3.33% _(0.02)	2.00% _(0.08)	3.33% _(0.15)	6.67% _(0.04)
		AED	11.67% _(0.05)	8.33% _(0.063)	3.33% _(0.01)	2.78% _(0.01)	10.00% _(0.05)
	Right-Dense	VDARS	0.00% _(0.00)				
		Eureka	0.00% _(0.00)	0.00% _(0.00)	0.00% _(0.00)	1.00% _(0.02)	1.67% _(0.05)
		AED	0.00% _(0.00)	0.00% _(0.00)	0.00% _(0.00)	3.33% _(0.01)	1.67% _(0.01)

Table 5: Full breakdown of effective Vulnerability Rates (standard deviation) for all fine-grained subtypes. AED consistently achieves the highest effective vulnerability rates.

other baselines across nearly all individual subtypes.

Prompt for LLMs

The prompts used to guide the LLM in generating diverse collision scenarios are shown in Prompt. 1 and Prompt. 2.

After multiple queries to the LLM and incorporating traffic rules, we selected seven representative collision types that are both realistic and commonly encountered. These seven types constitute the core collision scenarios studied in this paper. The LLM was then prompted again to generate the final task descriptions for each scenario.

The prompts used to generating reward function for a specific environment and collision are presented in Prompt.3, 4 and 5. The specific environment descriptions and task descriptions for all collision types are presented in Sec..

Prompt 1: System Prompts of Generating Diverse Collision Patterns

```
You are a helpful assistant designed to generate diverse collision scenarios for an autonomous driving simulation involving a defender, attackers, and dummy vehicles.  
Your task is to generate different types of collisions based on traffic regulations where the defender is at fault. These collision scenarios will be used to write a reward function for reinforcement learning to improve the defender's strategy.  
You should consider various collision types, ensuring that the defender is primarily responsible according to traffic rules.
```

Prompt 2: Prompts of Generating Diverse Collision Patterns

```
I am working on a simulation environment that models autonomous driving strategies. The description of the environment is shown below.  
  
<Environment Description>  
  
I hope to write a reward function for the attackers to facilitate reinforcement learning for the following goal. My goal is to generate a variety of collision scenarios as many diverse collision scenarios between the defender and attackers during the simulation. However, the key requirement is that, according to traffic regulations, the defender should bear the primary responsibility for these collisions. In other words, the defender must be at fault under traffic rules, which will help identify vulnerabilities in the defender's strategy for further improvement. Therefore, the task here is to propose different types of collisions that could occur in this environment. The collisions should be varied in nature, easy to determine based on vehicle states (e.g., position, speed, lane), and focused on the defender being primarily at fault. Examples of collisions could involve blocking an attacker's lane, sudden speed changes that force collisions, or any other traffic violation where the defender is responsible. I need these types for use in writing a reward function for reinforcement learning and for simulating realistic traffic behavior.  
Please sort all collision types by their likelihood of occurrence, from highest to lowest, and output the result. Additionally, please format your response in a structured way, such that it can be easily extracted and used programmatically. I would like the response to follow this structure:  
  
```json  
{
 "collision_types": [
 {
 "type": "Collision Type Name",
 "description": "A detailed description of the collision type.",
 "fault": "Who is at fault for this collision (typically the defender).",
 "scenario": "A sample scenario where this collision would occur. It should contain specific actions of the attackers and the defender before and during the collision."
 "criteria": "Explain how to determine whether a collision belongs to the current type."
 },
 {
 "type": "Another Collision Type Name",
 "description": "Another description.",
 "fault": "Who is at fault.",
 "scenario": "Sample scenario."
 "criteria": "Explain how to determine whether a collision belongs to the current type."
 }
]
}
```

```

```
Here is a possible example for lane change collisions:
{
  "collision_types": [
    "type": "Lane Change Collisions",
    "description": "The defender changes lanes unexpectedly, causing a collision with an attacker or another vehicle already occupying that lane."
    "fault": "The defender is primarily at fault, as it violated traffic rules by changing lanes unsafely."
    "scenario": "The defender makes a lane change to the left or right sufficient clearance, directly causing the attacker or a dummy vehicle to crash into the defender."
    "criteria": "Before the collision's timestep, the defender and attacker should not be in the same lane; at the collision's timestep, the defender and attacker should be in the same lane; at the collision's timestep, the attacker should not have accelerated, decelerated, or changed lanes."
  ]
}

```

Prompt 3: System Prompts of Generating Reward Function

You are a reward engineer trying to write reward functions to solve reinforcement learning tasks as effective as possible.

Your goal is to write a reward function for the environment that will help the agent learn the task described in text.

When provided with class methods or incomplete code, you are required to complete the code to the best of your understanding based on common patterns and conventions.

Your reward function should use useful variables and class functions from the environment by access the local variables of the class through 'controlled_vehicles' for example.

The reward function signature can be:

```

def compute_reward(self) -> Tuple[torch.Tensor, Dict[str, torch.Tensor]]:
    ...
    return reward, {}

```

Make sure any new tensor or variable you introduce is on the same device as "self.speed".

Prompt 4: Prompts of Generating Reward Function

The environment description is as follows:

<Environment Description>

and the specific python environment code is

<Environment Python Code>

Write a reward function for the following task:

<Task Description>

Prompt 5: Prompts of Tips for Writing Reward Function

The output of the reward function should consist of two items:

- (1) the total reward (should be a tensor of shape (num_attackers, 1)),
- (2) a dictionary of each individual reward component.

The code output should be formatted as a python code string: ````python ... ````.

Some helpful tips for writing the reward function code:

- (1) You may find it helpful to normalize the reward to a fixed range by applying transformations like `torch.exp` to the overall reward or its components
- (2) `torch.norm` can only be used on a tensor. Therefore, if the input is of type `numpy.ndarray`, you need to first convert it to a tensor, pass it to `torch.norm`, and then convert it back to `ndarray` (if needed).
- (3) If you choose to transform a reward component, then you must also introduce a temperature parameter inside the transformation function; this parameter must be a named variable in the reward function and it must not be an input variable. Each transformed reward component should have its own temperature variable
- (4) Make sure the type of each input variable is correctly specified; a float input variable should not be specified as `torch.Tensor`
- (5) Most importantly, the reward code's input variables must contain only attributes of the provided environment class definition (namely, variables that have prefix `self.`). Under no circumstance can you introduce new input

variables.

(6) Your reward should contain a component which rewards the defender-responsible collision and you should carefully consider how to determine whether a collision is defender-responsible. If a defender-responsible collision is created, you must grant a reward of 100 to all attackers.

(7) Aside from this component for defender-responsible collision, you should normalize all other reward components to the range [0,10] using transformations such as torch.exp.

Detailed Description

The environment descriptions for the Highway and Roundabout settings are provided in Prompt. 6 and 7, respectively. As an illustrative example, the detailed task description for a rear-end collision scenario is presented in Prompt. 8. The descriptions of other scenarios differ only in the text specifying the collision configurations, which are provided in Prompt. 9.

Prompt 6: Environment Description of Highway

This environment is a simulation environment designed to model autonomous driving strategies.

This is a one-way, four-lane highway, with travel occurring in the positive x-axis direction. The right side of the travel direction corresponds to the positive y-axis direction. The width of each lane is 4 and the length of each vehicle is 5.

Multiple vehicles operate within the lanes based on their respective autonomous driving strategies. Among these vehicles, there is 1 defender, {n_attackers} attackers, and the remaining are dummy vehicles.

At any given moment, a vehicle can perform one of the following actions:

'LANE_LEFT': Change lanes to the left,
'IDLE': Remain idle,
'LANE_RIGHT': Change lanes to the right,
'FASTER': Accelerate, OR
'SLOWER': Decelerate.

Prompt 7: Environment Description of Roundabout

This environment is a simulation environment designed to model autonomous driving strategies.

In this task, the ego-vehicle is approaching a roundabout with flowing traffic. It will follow its planned route automatically, but has to handle lane changes and longitudinal control to pass the roundabout as fast as possible while avoiding collisions. The width of each lane is 4 and the length of each vehicle is 5.

Multiple vehicles operate within the lanes based on their respective autonomous driving strategies. Among these vehicles, there is 1 defender, {n_attackers} attackers, and the remaining are dummy vehicles.

At any given moment, a vehicle can perform one of the following actions:

'LANE_LEFT': Change lanes to the left,
'IDLE': Remain idle,
'LANE_RIGHT': Change lanes to the right,
'FASTER': Accelerate, OR
'SLOWER': Decelerate.

Prompt 8: Task Description of Rear End Collision

Your task is to design a reward function for the attackers to facilitate reinforcement learning.

The goal is to generate as many diverse collision scenarios between the defender and attackers (collisions between the defender and other dummy vehicles are also ok) as possible during the simulation.

However, the key requirement is that, according to traffic regulations, the defender should bear the primary responsibility for these collisions. In other words, the defender must be at fault under traffic rules, which will help identify vulnerabilities in the defender's strategy for further improvement.

Specifically, we aim to maximize the following type of collision, its key information is as follows:

```
{
  "type": "Rear End Collision",
  "description": "The defender collides with the attacker from behind, typically caused by the defender not maintaining a safe following distance or being unable to stop in time.",
  "fault": "The defender is primarily at fault, as it failed to maintain an adequate following distance or was unable to react in time to avoid the collision.",
  "scenario": "The defender is traveling on the same straight line as the attacker but from behind (at a lower x-coordinate). The attacker remains in the same lane throughout the event.",
  "criteria": "Before the collision's timestep, the defender and attacker should be on the same straight path (same lane), with the defender positioned behind the attacker (lower x-coordinate). At the collision's timestep, the defender should be in the same lane as the attacker, but with the defender's position behind. The attacker should not have performed any significant acceleration or lane change prior to the collision. The defender should not change lanes before the collision, and at the time of the collision, the defender's and attacker's y-coordinates should be close to each other."
}
```

When you try to determine responsibility for an accident, you can simply check whether it is collision at the above type, and then give the attacker a corresponding reward based on whether the defender is primarily responsible.

Here are some suggestions for designing the reward function:

1. Access the state of each vehicle.

You can access the state of each vehicle at the current step by using `self.controlled_vehicles`.

This is a list of `MDPVehicle` objects, where `self.controlled_vehicles[0]` is the defender, `self.controlled_vehicles[self.n_defenders:self.n_defenders + self.n_attackers]` are attackers.

You can see the definition of the `Vehicle`, `ControlledVehicle`(inherited from `Vehicle`) and `MDPVehicle`(inherited from `ControlledVehicle`) in the environment code to get more information about the attributes of the vehicle and some useful class function you can use.

Besides, you can access `self.controlled_vehicles_trajectory[-1]` to get the trajectory of all vehicles in the last step, which is also a list of `MDPVehicle` objects.

Then you can compare the state of the vehicles in the last step and the current step to determine the action of the vehicle in the last step.

2. Rewarding Attackers for Generating Defender-Responsible Collisions

You should design a component of the reward function that rewards attackers for successfully creating collision scenarios where the defender is at fault.

To achieve this, you can access the state of each vehicles, analyze whether there is a collision between the defender and attackers and then determine whether the defender is responsible for the collision.

This will be a challenging task, so you are encouraged to create a relatively complex constraint to assess whether the defender is responsible for the collision. If the collision scenario meets the criteria, reward the attackers generously. If not, reward them less.

To determine if the defender is at fault, you should base the judgment on traffic rules. If this proves difficult, consider whether the defender could have avoided the collision by adjusting their strategy beforehand. For example, if the defender makes an incorrect decision, such as an unnecessary lane change or failure to decelerate in time, resulting in a collision, then the defender should be deemed responsible.

3. Avoiding Unnecessary Collisions

The reward function should avoid rewarding collisions that occurring between attackers or between attackers and dummy vehicles.

4. Encouraging Attackers to Stay Close to the Defender

You should also design a component of the reward function that encourages attackers to remain close to the defender throughout the simulation, such as a distance-based reward. Without this, attackers may fail to generate any collisions and, as a result, may not receive any reward.

5. You are also encouraged to design additional components for the reward function to further enhance the learning process.

6. Since the accident scenarios are generated by the collaboration of the attackers, you should reward all the attackers for an defender-responsible accident scenario.

7. If the attacker performs some erroneous actions, such as changing lanes or slowing down, at the moment of

collision with the defender, it is highly likely that the attacker should also bear some responsibility for the incident. You need to carefully assess this situation. If the attacker is determined to be at fault in the current accident, then the attacker should receive a smaller reward or no reward at all.

Prompt 9: Configurations of All Collisions

```

Left-Sparse:
{
  "type": "Left Lane Change Collisions",
  "description": "The defender unexpectedly changes lanes to the left(decreasing the lane index), causing a collision with an attacker or another vehicle already occupying that lane.",
  "fault": "The defender is primarily at fault, as they violated traffic rules by changing lanes unsafely to the left without proper clearance.",
  "scenario": "The defender moves from their original lane to the left (decreasing the lane index), directly causing the attacker or another vehicle to crash into the defender's vehicle due to insufficient clearance or the attacker's inability to avoid the collision.",
  "criteria": "Before the collision's timestep, the defender and attacker should not be in the same lane; at the collision's timestep, the defender and attacker should be in the same lane; at the collision's timestep, the attacker should not have accelerated, decelerated, or changed lanes."
}

Right-Sparse:
{
  "type": "Right Lane Change Collisions",
  "description": "The defender unexpectedly changes lanes to the right(increasing the lane index), causing a collision with an attacker or another vehicle already occupying that lane.",
  "fault": "The defender is primarily at fault, as they violated traffic rules by changing lanes unsafely to the right without proper clearance.",
  "scenario": "The defender moves from their original lane to the right (increasing the lane index), directly causing the attacker or another vehicle to crash into the defender's vehicle due to insufficient clearance or the attacker's inability to avoid the collision.",
  "criteria": "Before the collision's timestep, the defender and attacker should not be in the same lane; at the collision's timestep, the defender and attacker should be in the same lane; at the collision's timestep, the attacker should not have accelerated, decelerated, or changed lanes."
}

Rear End:
{
  "type": "Rear End Collision",
  "description": "The defender collides with the attacker from behind, typically caused by the defender not maintaining a safe following distance or being unable to stop in time.",
  "fault": "The defender is primarily at fault, as it failed to maintain an adequate following distance or was unable to react in time to avoid the collision.",
  "scenario": "The defender is traveling on the same straight line as the attacker but from behind (at a lower x-coordinate). The attacker remains in the same lane throughout the event.",
  "criteria": "Before the collision's timestep, the defender and attacker should be on the same straight path (same lane), with the defender positioned behind the attacker (lower x-coordinate). At the collision's timestep, the defender should be in the same lane as the attacker, but with the defender's position behind. The attacker should not have performed any significant acceleration or lane change prior to the collision. The defender should not change lanes before the collision, and at the time of the collision, the defender's and attacker's y-coordinates should be close to each other."
}

Overtake:
{
  "type": "Failed Overtake Return to Original Lane",
  "description": "The defender moves to an adjacent lane to overtake a vehicle, but fails to accelerate enough to pass safely and attempts to return to the original lane too soon, resulting in a side or rear-end collision with the original vehicle.",
  "fault": "The defender is at fault for not completing the overtaking maneuver properly and merging back into the original lane unsafely.",
  "scenario": "The defender is behind a slower attacker in Lane 2. It changes to Lane 3 (right) using 'LANE_RIGHT', intending to pass. However, the defender does not gain sufficient speed to get ahead. Despite that, it executes a 'LANE_LEFT' to return to Lane 2, where the attacker is still present. This causes a collision as the attacker impacts the side or rear of the defender."
}

```

```

    "criteria": "Defender initially changes lanes to pass, then attempts to return to the original lane. The attacker or dummy vehicle remains in the original lane and does not change speed or lanes. The collision occurs during or immediately after the defender tries to return to the original lane. The x-position of the defender is still close to or behind the attacker. You can check self.controlled_vehicles_trajectory[-5:-1] to get the previous actions of the defender before the collision."
}

Left-Dense:
{
    "type": "Left Lane Change During Congestion",
    "description": "The defender attempts to switch to the left lane in dense traffic due to a slower vehicle ahead but fails to properly assess the space in the target lane. The defender merges into a gap that does not provide a safe buffer with the vehicle in the left lane, causing a side or rear-side collision.",
    "fault": "The defender is at fault for merging left without maintaining a safe gap in heavy traffic, violating lane-change safety protocols.",
    "scenario": "In a congested traffic setting, the defender's current lane has a slower vehicle directly ahead at a close distance. Seeking to maintain momentum, the defender initiates a 'LANE_LEFT' maneuver toward a seemingly faster adjacent lane. However, the left lane already contains a vehicle (attacker or dummy) very close to the intended merge location. The defender does not sufficiently reduce speed and clips the front or side of the vehicle in the left lane.",
    "criteria": "There must be high traffic density in both lanes (i.e., low spacing between vehicles). A vehicle must be present in close proximity ahead of the defender in its original lane. The defender must initiate a left lane change when there is insufficient gap to the vehicle ahead in the target lane. A collision must occur during or immediately after the lane change, with the point of impact on the defender's left rear-side or side, and the other vehicle's front or front-side."
}

Right-Dense:
{
    "type": "Right Lane Change During Congestion",
    "description": "The defender attempts to switch to the right lane in dense traffic due to a slower vehicle ahead but fails to properly assess the space in the target lane. The defender merges into a gap that does not provide a safe buffer with the vehicle in the right lane, causing a side or rear-side collision.",
    "fault": "The defender is at fault for merging right without maintaining a safe gap in heavy traffic, violating lane-change safety protocols.",
    "scenario": "In a congested traffic setting, the defender's current lane has a slower vehicle directly ahead at a close distance. Seeking to improve progress, the defender initiates a 'LANE_RIGHT' maneuver toward a seemingly faster adjacent lane. However, the right lane already contains a vehicle (attacker or dummy) very close to the intended merge location. The defender does not sufficiently reduce speed and clips the front or side of the vehicle in the right lane.",
    "criteria": "There must be high traffic density in both lanes (i.e., low spacing between vehicles). A vehicle must be present in close proximity ahead of the defender in its original lane. The defender must initiate a right lane change when there is insufficient gap to the vehicle ahead in the target lane. A collision must occur during or immediately after the lane change, with the point of impact on the defender's right rear-side or side, and the other vehicle's front or front-side."
}

T-Bone:
{
    "type": "T-Bone Collision at Intersection (Roundabout Only)",
    "description": "A side-impact (T-bone) collision occurs when the front of the defender hits the side of another vehicle (attacker or dummy) at an intersection or roundabout due to the defender ignoring right-of-way or entering improperly.",
    "fault": "The defender is at fault for not yielding at the intersection or roundabout, violating the right-of-way of the attacker or another vehicle already in the roundabout.",
    "scenario": "In a roundabout environment, the attacker is traveling along the roundabout lane, already engaged in the circular path. The defender enters the roundabout from an entry lane without yielding, performing a 'FASTER' or even an 'IDLE' to merge. The defender collides with the side of the attacker's vehicle at the intersection of their paths.",
    "criteria": "Defender should be at the roundabout entry point. Attacker is already in the roundabout, moving along its circular or curved path. Collision must involve the front of the defender and the side of the attacker. Attacker's speed and path should remain stable. Defender must enter without stopping or yielding."
}

```