2025 IEEE Intelligent Vehicles Symposium (IV)
June 22-25, 2025. Cluj-Napoca, Romania

Vulnerability-aware and Curiosity-driven Adversarial Reinforcement Learning Policy for Safety-Critical Scenario Generation

Xuan Cai[1], Zhiyong Cui*[1], Xuesong Bai[1], Ruimin Ke[2], Haiyang Yu[1,3], Yilong Ren[1,3], and Zechang Ye[1]

***Abstract*— Autonomous vehicles (AVs) face significant threats to their safe operation in complex traffic environments. Adversarial policy for scenario generation has been established as a robust paradigm for enhancing AV resilience against adversarial perturbations through proactive exposure to synthetically engineered safety-critical scenarios. Training an attacker within an adversarial policy, allowing the target AV to expose vulnerabilities through interaction with this attacker. However, adversarial policies in existing methodologies often get stuck in a loop of over-exploiting established vulnerabilities, resulting in poor exploration for AVs. To overcome the limitations, we introduce a pioneering framework termed the vulnerability-aware and curiosity-driven adversarial reinforcement learning policy. Specifically, during the traffic vehicle attacker training phase, a surrogate network is employed to fit the value function of the AV victim, providing dense information about the victim's inherent vulnerabilities. Subsequently, random network distillation is used to characterize the novelty of the scenario, constructing an intrinsic reward to guide the attacker in exploring unexplored territories. Experimental results demonstrated that the adversarial policy embedded within the attacker exhibited robustness in convergence and significantly enhanced the activation of policy exposure in learning-based AVs, outperforming both other adversarial modalities and alternative reinforcement learning approaches, with a notable reduction in crash rates. The code is available at https://github.com/caixxuan/VCAT.**

I. Introduction

AVs have gradually increased their market presence but have also become one of the sources of threats to public safety [1]. However, it is extremely challenging to comprehensively enhance the robustness due to sparse corner cases. Adversarial policies for safety-critical scenario generation provide an effective method [2]. By allowing attackers, i.e., traffic vehicles, to create safety-critical scenarios, learning-based AVs are expected to exhibit how to addressing risks under safety expectations. In general, existing adversarial policies face two challenges: insufficient utilization of the victim's intrinsic information and the limited variety of the attacker's attack modes.

*This work was supported by the National Key Research and Development Project of China under Grant 2022YFB4300400 and the Beijing Natural Science Foundation (project number: L243008). (*Corresponding author: Zhiyong Cui*)

[1]{Xuan Cai, Zhiyong Cui, Xuesong Bai, Zhenshu Ma, Haiyang Yu, Yilong Ren and Zechang Ye} is with State Key Laboratory of Intelligent Transportation Systems, School of Transportation Science and Technology, Beihang University, Beijing, 100191, P.R.China (E-mail: {caixuan, zhiyongc, xs_bai, hyyu, yilongren, ye_zc}@buaa.edu.cn)

[2]Ruimin Ke is with Department of Civil and Environmental Engineering, Rensselaer Polytechnic Institute, Troy, New York, 12180, USA (E-mail:ker@rpi.edu)

[3]{Haiyang Yu and Yilong Ren} is with Zhongguancun Laboratory, Beijing, 100191, P.R.China

A. Problems and Challenges

Exploitation of the intrinsic vulnerability of victim. Prevailing studies often utilize fused environmental observation via optimization [3] or learning [4] methods to pinpoint the desired attack, while often neglecting the exploitation of the victim (i.e., target AV)'s intrinsic vulnerabilities. This oversight is consequential; reliance on mere observational data can yield substantial pitfalls, as attackers may struggle to identify unfavorable states of the black-box victim, making it difficult to launch effective attacks, particularly under conditions where safety-critical scenarios are rare. Such occurrences are quite common in AVs where the "long-tail effect" [5] exists.

Exploration of the policy space of victim. Traditional attack methods might only set binary collision or not, or a continuous probability distribution [6]. However, such tactics may falter due to inadequate exploration, leading to a phenomenon known as mode collapse, particularly under conditions of sparse rewards [7]. This research gap is often exacerbated by the propensity for local optimization intrinsic to learning-based techniques.

B. Main Contribution

To address the above issues, we propose a vulnerability-aware and curiosity-driven adversarial reinforcement learning policy, with its key contributions summarized as follows.

- **Adversarial Policy**: We have constructed a vulnerability-aware and curiosity-driven adversarial policy for scenario generation. Inspired by the victim-aware and curiosity [8] mechanism, we have developed a curiosity-driven reinforcement learning (RL) attack paradigm, that leverages vulnerabilities of the victim by focusing on areas that the attacker has not fully understood or explored.
- **Adversarial Experiment**: To rigorously evaluate the effectiveness of the proposed policy, we conducted extensive adversarial simulations. The results of these experiments reveal that our proposed method markedly bolsters the risk exposure capabilities of AVs, thereby substantially strengthening the safety robustness with adversarial training in autonomous driving.

C. Construction

The overall structure of the paper is as follows. Section II reviews related research work on scenario generation and adversarial policy. In Section III, we propose the adversarial policy method for scenario generation, complemented by adversarial training [9]. Subsequently, in Section IV, the

2025 IEEE Intelligent Vehicles Symposium (IV) | 979-8-3315-3803-3/25/$31.00 ©2025 IEEE | DOI: 10.1109/IV64158.2025.11097466

proposed method is conducted in a simulation experiment, and the results are analyzed. Finally, the conclusion and future works are summarized in Section V. Some commonly acronyms are also adopted, including *w.r.t.* (with respect to) and *w.l.o.g.* (without loss of generality).

II. Related Works

A. Scenario Generation

Scenario generation employing RL has garnered significant academic attention by training adversarial agents to effectively execute attacks. Especially in the field of AVs, artificial intelligence (AI) testing AI is a common way. Through adversarial training, one can enhance the robustness of the target AI agent, a concept commonly seen in Generative Adversarial Network (GAN) [10], Generative Adversarial Imitation Learning [11], and Game Theory [12].

In response to the limitations of traditional adversarial RL, some literature aims to improve the performance in specific autonomous driving adversarial scenario generation tasks. For instance, the series RL method proposed by Cai et al. [13] considerably diversified the range of adversarial scenarios. Huang et al. [14] leveraged Stackelberg game dynamics by factoring in the adaptivity of the agent, generating challenging yet solvable scenarios, thus enhancing the stability and robustness of RL training.

Despite extensive research suggesting that constructing safety-critical scenarios with RL aids in the training and validation, the potential benefits of integrating vulnerability-evaluation and curiosity-exploration within adversarial policies for learning tasks remain an unresolved issue.

B. Adversarial Policy

Adversarial policy is a crucial method for evaluating the robustness of AI agents [15], [16], [17], and it has accumulated substantial empirical research. The perspectives on adversarial phenomena can be dichotomized into adversarial policies and training strategies. In terms of adversarial policies, Ding et al. [18] devised a generative adversarial network aimed at stabilizing adversarial training to enhance contextual prediction in AVs through the restoration of visually degraded images. Kloukiniotis et al. [19] reviewed denoising techniques as a countermeasure to adversarial attacks on AVs, emphasizing the role of adversarial training in improving adversarial robustness. Adversarial policy and adversarial training are frequently interdependent and mutually reinforcing. Zhang et al. [20] introduced a closed-loop adversarial training framework aimed at improving the robustness and safety of AV control.

However, existing adversarial policy methods have not exploited the intrinsic vulnerability nor explored the policy space of the victim, which hinders the advancements in the robustness of AI-driven AVs.

III. Proposed Method

This section introduces a novel adversarial policy for scenario generation, thereby strengthen the robustness of AVs via adversarial training. It first provides an overview of the proposed adversarial framework and then elaborates on the proposed adversarial policy and adversarial training protocols.

A. Overview of the Adversarial Framework

The overview of the proposed adversarial framework is illustrated in Fig.1. It divides into dual stages of adversarial policy and training, based on the victim's state, which alternates between being fixed (frozen) or variable (thawed) during the training and evaluation phases. In essence, the adversarial policy studied in this paper models the game between the attacker and the victim as a two-player minimax Markov Game (MG), which models the strategies of agents as part of the Markov Decision Process. In MGs, multiple agents perform a series of actions to maximize their collective or individual benefits. Specifically, two-player zero-sum MGs [21] involve a pair of agents with completely opposite interests. This study relaxes the zero-sum game problem due to the complicated traffic interactions.

B. Adversarial Policy

Before conducting the adversarial policy, it is imperative that the victim (target AV) be subject to extensive training using standard datasets (e.g., road-collected or random-generated data) to ascertain it possesses fundamental navigational proficiencies, albeit with a deficiency in managing anomalous or corner cases. Once the AV agent has been thoroughly trained, its parameters should be frozen to play the role of the victim, υ, thus being attacked by the training attacker, α.

Victim and attacker constitute a two-player MG. When both are RL-driven, their value functions are $V^{\upsilon}_{\pi_{\theta_\alpha}}(s)$ and $V^{\alpha}_{\pi_{\theta_\alpha}}(s)$, also known as expected rewards [22]. Therefore, the goal of adversarial policy is for the attacker to learn to adeptly discern and exploit the victim's vulnerabilities, specifically by minimizing $V^{\upsilon}_{\pi_{\theta_\alpha}}(s)$. Given the network parameters θ_υ remain frozen, policy π_{θ_υ} is thereby fixed, which effectively incorporates the victim as an integral component of the environmental construct. Thus, the objective of the adversarial policy is formulated as:

$$J = arg\max_{\theta_\alpha}\left(V^{\alpha}_{\pi_{\theta_\alpha}}(s) - V^{\upsilon}_{\pi_{\theta_\alpha}}(s)\right) \tag{1}$$

Therefore, an important insight is that if we can estimate the victim's $V^{\upsilon}_{\pi_{\theta_\alpha}}(s)$, it would help find its weaknesses more accurately. The Proximal Policy Optimization (PPO) paradigm [22] is used to train the π_{θ_α}.

1) Victim Value Approximation Network: We use an approximation network (parameterized by θ_υ) to fit the state-value function of υ, which aids in the explicit formulation of Eq.1. Adopting the Temporal Difference (TD) learning paradigm, we define the loss function of the approximation network equivalent to the TD-error:

$$arg\min_{\theta_\upsilon}\left\|V^{\upsilon}_{\pi_{\theta_\alpha}}(s_t) - \left(\widehat{r^{\upsilon}}(s_t,a_t) + \gamma\mathbb{E}_{s_{t+1}\sim P}[V^{\upsilon}_{\pi_{\theta_\alpha}}(s_{t+1})]\right)\right\|^2 \tag{2}$$

where γ is the discount factor, P is the state transition probability, $a_t = (a^{\alpha}_t, a^{\upsilon}_t)$ is the sampled joint action, and

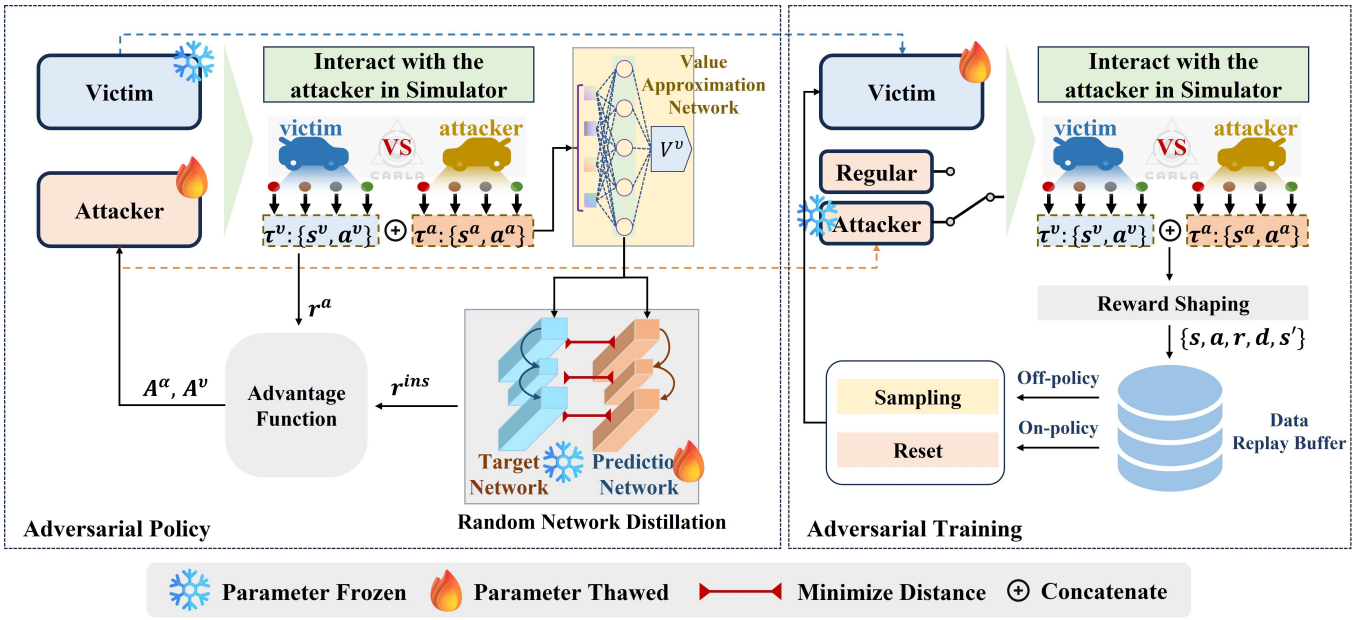

Fig. 1. Overview of the proposed adversarial framework. It is divided into two stages: adversarial policy, enclosed by the left dashed box, and adversarial training, enclosed by the right one. The snowflake pattern indicates that the neural network parameters are frozen, while the flame pattern indicates that the parameters can be adjusted for learning. The horizontal line with the reverse triangle arrow represents the minimum Euclidean distance between the two. The circled cross signifies data concatenation.

$\widehat{r^v}$ is the estimated reward function for the victim, *w.l.o.g.*, under the black-box assumption. This is done to extend the victim's generality, which is beneficial for comprehensive training and validation endeavors:

$$\widehat{r^v} = \lambda_1 \cdot r_{target} - \lambda_2 \cdot r_{acc} - \lambda_3 \cdot r_{collision} \tag{3}$$

where λ is the weight, r_{target} is the reward for the victim reaching the goal, r_{acc} is the acceleration reward, and $r_{collision}$ is the collision reward.

2) Curiosity-Driven Exploration: The value approximation network (VAN) is capable of approximating the significance of the current state s_t *w.r.t.* the victim; hence, θ_v encapsulates dense information about the state value of v at the said s_t. If a certain s_t represents an unfamiliar state to v, it becomes imperative to induce the adversarial agent to probe and exploit this state. Inspired by the random network distillation mechanism (RND) [8], two networks are constructed: a stationary target network, ϱ and a dynamic predictor network, $\hat{\varrho}$. The parameter of ϱ is randomized and then fixed, while $\hat{\varrho}$ is continuously optimized after randomization, with the aim of continuously approximating ϱ. This iterative optimization process is driven by the intent to minimize prediction disparities. When $\hat{\varrho}$ encounters a fresh state, the prediction error will be high, resulting in a high intrinsic reward output. Considering that the last hidden layer of θ_v (denoted as φ^v) due to its potent representation of the characteristics of the dense state *w.r.t.* $V^v_{\pi_{\theta_\alpha}}$, it is used as input for RND. Therefore, the mean square error of RND is:

$$r^{ins} = \|\hat{\varrho}(\varphi^v(s_t)) - \varrho(\varphi^v(s_t))\|^2 \tag{4}$$

where r^{ins} is the intrinsic reward, which can adaptively adjust the exploration value of s_t to steer the exploration of the attacker.

3) Adversarial Policy Training: The crux of the PPO lies in the calculation of the advantage function. **Algorithm 1** incorporates r^{ins} into the advantage function, simultaneously coordinating r^v and r^α. This training requires the initialization of six networks. The calculations of the advantage functions A^α_t, A^v_t, and $A^{\alpha,ins}_t$ are shown from lines 7 to 9. Subsequently, the training objective for PPO as in Eq.2 can be computed:

$$\begin{cases} arg\max_{\theta_\alpha} \mathbb{E}_{(a^\alpha_t, s_t)\sim\pi_{\theta_{\alpha,k}}}[min\big(\rho_t A^\alpha_t, clip\big(\rho_t, 1-\epsilon, 1+\epsilon\big) \\ \qquad A^\alpha_t\big) - min\big(\rho_t A^v_t, clip\big(\rho_t, 1-\epsilon, 1+\epsilon\big) A^v_t\big)] \\ \rho_t = \frac{\pi_\alpha(a^\alpha_t|s_t)}{\pi_{\alpha,k}(a^\alpha_t|s_t)} \\ \begin{cases} A^v_t = A^v_{\pi_{\alpha,k}}(a^\alpha_t, s_t) \\ A^\alpha_t = A^\alpha_{\pi_{\alpha,k}}(a^\alpha_t, s_t) + \lambda A^{\alpha,ins}_{\pi_{\alpha,k}}(a^\alpha_t, s_t) \end{cases} \end{cases} \tag{5}$$

where λ denotes a hyperparameter that signifies the degree of exploration. This objective function is designed to leverage both the victim's value function and the intrinsic value of exploration, aiming to expeditiously navigate towards a state that maximizes expected rewards. Consequently, r^{ins} can be internalized as the advantage function of PPO, with λ balancing exploitation and exploration, and its value setting is referenced to [23].

C. Adversarial Training

Upon successful execution of the adversarial policy training, that is, once the policy governing the attacker has satisfied the pre-established criteria for convergence, the network parameters attributed to the attacker are henceforth frozen. Concurrently, the victim's parameters are thawed to learn robustness strategy against the onslaught of the well-

Algorithm 1 Adversarial Policy

Require: $V^{\upsilon}_{\pi_{\theta_\alpha}}$: state value of the victim; ϱ: target network of the RND; $\hat{\varrho}$: predictor network of the RND; $V^{\alpha,ins}_{\pi_{\theta_\alpha}}$: state value of the intrinsic reward; $V^{\alpha}_{\pi_{\theta_\alpha}}$: state value of the attacker; π_{θ_α}: attacker policy;

for $n = 1, 2, ..., N$ **do**
  while not done **do**
    $s_t = env.step(\upsilon, \alpha)$
    Collect trajectory: $\mathcal{T}.\boldsymbol{append}(s_t)$
  end while
  Compute r^{ins}_t in each step of $\mathcal{T}$ ▷ Based on Eq.4
  for $i = 1, 2, ..., T$ in $\mathcal{T}$ **do**
    $A^{\alpha}_t = r^{\alpha}_t + \gamma V^{\alpha(t)}_{\pi_{\theta_\alpha}}\left(s^{\alpha}_{t+1}\right) - V^{\alpha(t)}_{\pi_{\theta_\alpha}}\left(s^{\alpha}_t\right)$
    $A^{\upsilon}_t = r^{\upsilon}_t + \gamma V^{\upsilon(t)}_{\pi_{\theta_\alpha}}\left(s^{\alpha}_{t+1}\right) - V^{\upsilon(t)}_{\pi_{\theta_\alpha}}\left(s^{\alpha}_t\right)$
    $A^{\alpha,ins}_t = r^{\upsilon}_t + \gamma V^{\alpha(t),ins}_{\pi_{\theta_\alpha}}\left(s^{\alpha}_{t+1}\right) - V^{\alpha(t),ins}_{\pi_{\theta_\alpha}}\left(s^{\alpha}_t\right)$
  end for
  Update π_{θ_α} by minimizing the loss ▷ Based on Eq.5
  Update $V^{\upsilon}_{\pi_{\theta_\alpha}}$, $V^{\alpha,ins}_{\pi_{\theta_\alpha}}$, $V^{\alpha}_{\pi_{\theta_\alpha}}$ by minimizing the TD error
  Update $\hat{\varrho}$ by minimizing the loss ▷ Based on Eq.4
end for
return $\mathcal{T}$

trained attacker. Similarly, inverting Eq.1 as follows:

$$J = arg\min_{\theta_\upsilon}\left(V^{\alpha}_{\pi_{\theta_\upsilon}}(s) - V^{\upsilon}_{\pi_{\theta_\upsilon}}(s)\right) \tag{6}$$

where the parameters θ_α are frozen, meaning π_{θ_α} is fixed, while θ_υ is thawed to learn to minimize Eq.6.

Note that the victim can be any construct, but the PPO is adopted as the model to assess the potency of the adversarial training. Adversarial participants are incorporated into regular scenarios referring to the proportion of safety-critical scenarios derived from natural driving environment (NDE) [16], [24].

IV. Experiment

This study selects three scenarios for experiments. The simulation is conducted on a desktop PC equipped with a CPU Core i7 and a GPU NVIDIA 4070 Ti, using the highway-env [25]. This section details the experiment setup, research questions, results, and analysis.

A. Experiment Setup

1) Scenario Setup: The experiments set up three typical interactive scenarios, as illustrated in Fig.2, all of which are interactive dual-vehicle intersections that are recognized as safety-critical hotspots. The black attacker (referred to as the traffic vehicle) is equipped with an adversarial policy protocol, π_{θ_α}, enabling it to methodically engineer safety-critical scenarios that challenge the response robustness of the victim (referred to as the target AV dominated by π_{θ_υ}).

2) Hyper-parameter Setup: The generalized training regimen of the victim before the adversarial policy training is outside the scope of this study, and the key detail of the hyperparameters in this study is shown in Tab.I referred to [23], [24].

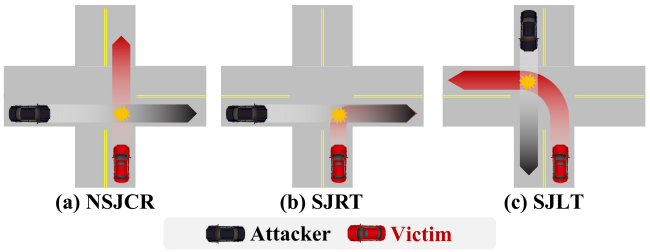

Fig. 2. Illustration for the setup of the three scenarios. The trajectory of the AV (victim) is represented by the red line, while the trajectory of the traffic vehicle (attacker) is represented by the black line. The scenarios are (a) # NoSignalJunctionCrossingRoute (# NSJCR), (b) SignalizedJunctionRight-Turn (# SJRT), and (c) SignalizedJunctionLeftTurn (# SJLT), respectively. The abbreviations are used hereafter.

TABLE I
HYPER-PARAMETER SETUP. BOLD NUMBERS ARE THE FINAL CHOICE.

PPO Adversarial Policy	
buffer capacity	5000
batch size	[64, **128**, 256]
learning rate of policy	[**5.0e-4**, 5.0e-3, 5.0e-2]
learning rate of value	[5.0e-4, **5.0e-3**, 5.0e-2
ϵ	[0.8, **0.9**, 0.95]
train iteration	[5, **10**, 20]
network dimension of policy	{state dim, 128, 64, action dim}
network dimension of value	{state dim, 128, 64, 1}
λ (curiosity exploration)	[0.05, **0.2**, 0.4, 0.6, 0.8]
γ	[0.8, 0.9, **0.95**, 0.98]
Value Approximation Network	
learning rate	[1.0e-4, **1.0e-3**, 1.0e-2]
network dimension	{input dim, 64, output dim}
Random Network Distillation	
learning rate	[1.0e-4, **1.0e-3**, 1.0e-2]
network dimension	{input dim, 128×3, output dim}

3) Baseline Setup of Adversarial Policy: This paper selects several state-of-the-art scenario generation methods as baselines, particularly focusing on the RL-based family that shares the same origin as the proposed method.

- **Monte Carlo Sampling/Random (MC)** [26]: The control of the attacker in the overall horizon is set randomly.
- **REINFORCE/Learning-to-Collide (LC)** [16]: The concept of GAN is utilized to generate safety-critical scenarios.
- **NormalizingFlow (NF)** [17]: The normalizing flow generator is leveraged to create natural and adversarial safety-critical scenarios.
- **RL-PPO** [27] / **RL-DDPG** [28] / **RL-TD3** [13] / **RL-SAC** [27]: RL-based agents are employed to play the role of attacker.

To be fair, the rewards or losses of attackers (r^{α}) in all methods are set as indicator functions.

B. Research Questions

Prior to the initiation of experimental procedures, we have articulated three research inquiries to steer the experimental design and execution:

- **RQ.1.** What is the efficacy of the proposed adversarial policy in supporting adversarial attacks?

- **RQ.2.** Does the proposed adversarial policy provide a superior level of resilience against adversarial maneuvers compared to others?
- **RQ.3.** How does each component of the proposed adversarial policy contribute to the attack capability (i.e., ablation studies)?

C. Experiment Result

1) RQ1. Efficacy of Adversarial Policy: **Metrics**. The crash rate characterizes the efficiency of generating safety-critical scenarios *w.r.t.* adversarial policies [26]. A more rapid increase in the crash rate signifies greater efficiency. To measure the coverage of adversarial policies, t-SNE [23] is used to visualize all action vectors from the slice trajectories of the victim interacting with different adversarial attackers in 2-D space. The wider the coverage of t-SNE, the richer the behaviors activated by π_{θ_v}, and the more vulnerabilities exposed. The number of different crash types is an important metric specifically used to measure the realism compared with NDE [29]; better match, better outcome, the better. For the features of all the crashes, we distinguish four categories to examine the distribution realism of the scenarios generated. The monte carlo method serves as the NDE.

Results. Fig.3 shows the crash rates under the three scenarios. The following characteristics can be identified: 1) Many baseline methods struggle to form effective attacks with the sparse incentives, prone to convergence failure in the limited iteration, such as LC and PPO in the #SJRT. The proposed method, however, can avoid this issue, with the crash rate rising to a high level. 2) The proposed adversarial policy experiences a distinct "V"-shaped phase of decline followed by an increase during early stages, as emphasized by the orange V-shaped arrows. Fig.4 presents the 2-D t-SNE visualization of the victim's action vector. It can be observed that the data distribution of the proposed is more widespread, suggesting that, compared to other counterparts, it can activate a richer policy within the victim, helping to uncover more vulnerabilities. Fig.5 illustrates the number of different crash types. The crash type distribution generated by the proposed adversarial policy exhibits the highest degree of consistency with NDEs.

Analysis. The method introduced herein adeptly circumvents convergence failure and assimilates potent adversarial patterns, achieving a higher crash rate. The V-shaped feature in Fig.3 and the extensive data distribution in Fig.4 further demonstrate the enhanced exploration capability of our approach without the exploitation of the internal workings within the victim. Although the proposed does not consistently achieve the highest crash rate, as seen in the #SJRT where it performs slightly worse than TD3 and DDPG, it improves the learning efficiency of RL under the sparse incentive condition, maintaining a balanced exploration and exploitation, especially suitable for such rare safety-critical conditions. For instance, DDPG exhibits convergence failure in the other two scenarios. Simultaneously, with the improvement in exploration capability, the level of randomness is increased, bringing it closer to the nature of NDEs.

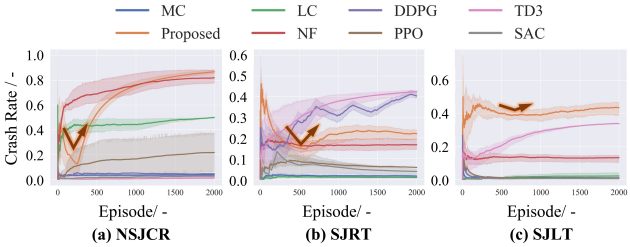

Fig. 3. Crash rate in the adversarial policy training with different methods. The orange "V"-shaped arrows highlight the decline-rise process experienced by the proposed method.

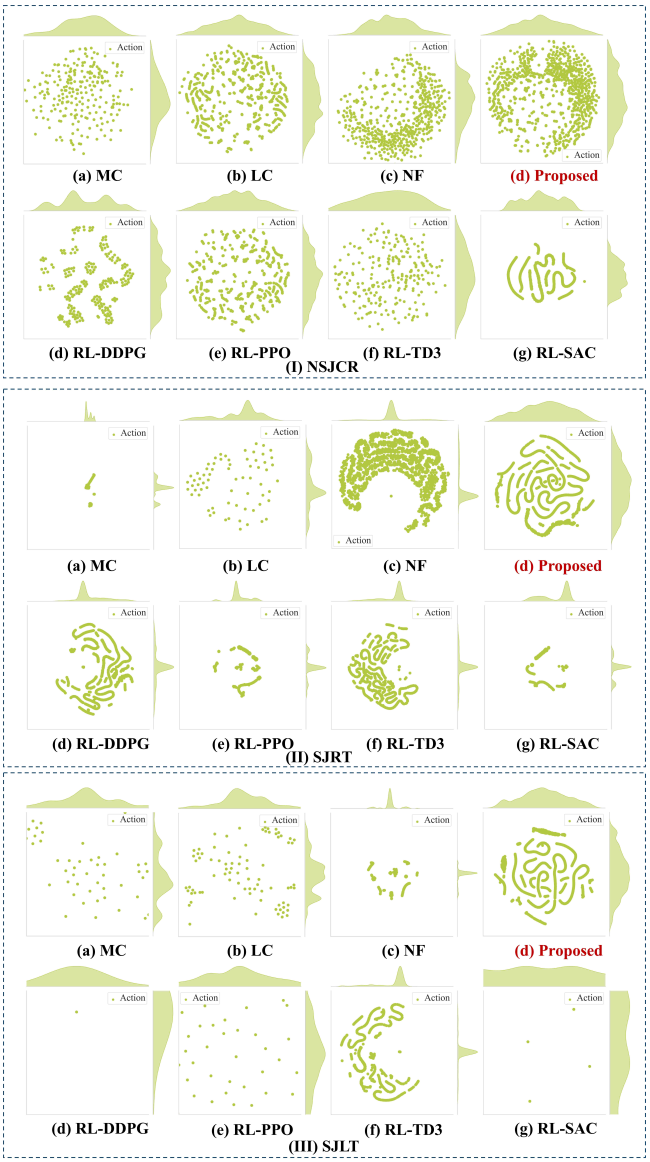

Fig. 4. t-SNE visualization of the victim (target AV) in the adversarial policy training under the three scenarios. The size of the coordinate axis is consistent for each scenario.

2) RQ2. Comparison of Adversarial Training: **Metrics**. Non-Crash Rate (as shown in Tab.II). Comparing the non-crash rate validated by different adversarial policies under various adversarial training methods, a higher non-crash

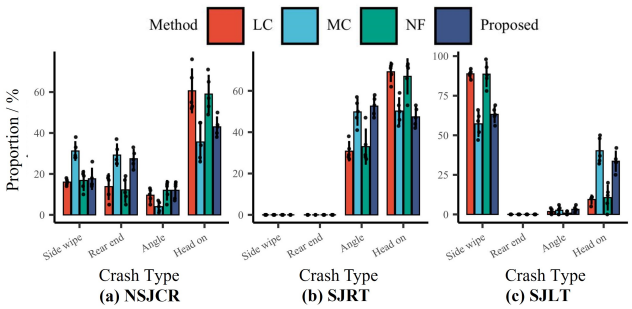

Fig. 5. Number of different crash types under the three scenarios. The monte carlo method serves as the NDE.

TABLE II

NON-CRASH RATE UNDER DIFFERENT VALIDATION METHODS AFTER ADVERSARIAL TRAINING. (AT: ADVERSARIAL TRAINING; VAL.: ATTACK METHODS USED TO VALIDATE ADVERSARIAL TRAINING).

Non-Crash Rate (↑) /%	(I) AT:MC+ Val.:Prop.	(II) AT:NF+ Val.:Prop.	(III) AT:DDPG+ Val.:NF	(IV) AT:Prop.+ Val.:NF
#NSJCR	16.4±2.6%	88.2±1.6%	18.0±2.9%	96.0±1.5%
#SJRT	76.8±3.9%	92.0±4.7%	83.6±1.3%	94.7±3.7%
#SJLT	57.1±5.4%	72.9±3.3%	86.3±2.7%	93.6±1.0%

rate is preferable. To test the effectiveness of adversarial training, cross-training and validation are employed. Taking the second column in Tab.II as an example, the adversarial training method uses MC, followed by the validation method using the proposed, to test whether the victim can withstand the attack from the proposed after being trained in MC.

Results. We selected MC, DDPG, and NF as baselines and compared four cross-adversarial training and validation categories. 1) In comparing groups (I) and (II), the adversarial training methods differ, while the validation methods remain consistent. The overall non-crash rate is relatively low, suggesting that the robust training effect of the adversarial policy is suboptimal. A comparison between the two groups revealed that the adversarial policy of NF outperformed that of MC. 2) In comparison with groups (III) and (IV), the non-crash rate of group (IV) is higher, demonstrating that, under the proposed adversarial policy, the system can maintain robust safety even when validated against another adversarial policy. This indicates that the proposed one is more effective. 3) In comparison with groups (II) and (IV), the non-crash rate of group (IV) is higher, which intuitively suggests that the proposed adversarial policy exhibits greater attack capability.

Analysis. The proposed method effectively uncovers a comprehensive attack space, encompassing a broader range of safety-critical scenarios. Adversarial training with this approach significantly enhances the victim's robustness, enabling it to effectively handle NDE and resist malicious attacks from other counterparts to a large extent. However, despite successful adversarial training, other methods exhibit limited policy action activation exploration, thereby constraining their generalization performance.

TABLE III

ABLATION STUDIES FOR THE PROPOSED ADVERSARIAL POLICY.

Crash Rate (↑) /%	MC	PPO	PPO-VA	Proposed
#NSJCR	1.2±0.2%	21.6±2.9%	22.2±2.4%	83.4±6.4%
#SJRT	1.0±0.1%	7.3±1.4%	7.2±1.7%	23.8±3.2%
#SJLT	1.9±0.4%	2.2±0.3%	3.5±0.3%	46.0±4.7%

3) RQ3. Ablation Studies: We focus on the ablation studies of attacking efficacy. **Ablation baseline:**

- PPO: The raw PPO adversarial policy method;
- PPO-VA: Vulnerability-aware PPO, in which the curiosity exploration hyperparameter is set to zero, i.e., $\lambda = 0$;
- Proposed: The full method introduced in this paper, $\lambda = 0.2$.

Results. The ablation experiment results are shown in Tab.III. MC is clearly inferior to the PPO method. However, the PPO still exhibits low attack efficiency, with a maximum of only about 21.6%. When the vulnerability-aware module is incorporated, the improvement in the crash rate is minimal and even decreases, with a maximum increase of only about 1.3%. When the proposed method is fully implemented, the crash rate significantly increases, particularly achieving a high crash rate of 83.4% in the first scenario.

Analysis. The utility of using the vulnerability-aware module alone is limited. This is because without the introduction of the exploration mechanism, it merely weights the states where the attacked victim may have vulnerabilities. However, some error exists in the estimated reward (see Eq.3), making it difficult to achieve improvements using only the VAN. The curiosity mechanism must be combined to explore a larger space; otherwise it will result in excessive exploitation.

V. CONCLUSIONS AND FUTURE WORKS

This study introduces the vulnerability-aware and curiosity-driven adversarial policy to address the challenge of generating diverse safety-critical scenarios for AVs under sparse reward conditions.Traditional adversarial policies often suffer from over-exploitation of known vulnerabilities and insufficient exploration of the victim's policy space. To overcome these limitations, the proposed adversarial policy integrates two key innovations: (1) a value approximation network that explicitly identifies the AV victim's decision-making vulnerabilities by fitting its state-value function, and (2) a random network distillation mechanism that generates intrinsic rewards to guide the attacker toward novel states, ensuring balanced exploration-exploitation trade-offs. Experimental results demonstrated that the proposed adversarial policy method exhibits enhanced convergence robustness, a wider coverage of activated target victims, more realistic crash types, and superior adversarial training capabilities compared to counterparts.

Future work will focus on incorporating real-world data into the training process, expanding the range of adversarial scenarios, and strengthening the system's resilience against adaptive adversaries.

DECLARATIONS

- **Acknowledgements** The authors would like to appreciate the financial support of the National Key R&D Program of China, No. 2023YFB4301802-02, the Beijing Natural Science Foundation (project number: L243008) and the National Natural Science Foundation of China (project number: 52441202).
- **Conflict of interest** On behalf of all the authors, the corresponding author states that there is no conflict of interest.

REFERENCES

[1] Xiaoqiang Sun, F Richard Yu, and Peng Zhang. A survey on cyber-security of connected and autonomous vehicles (cavs). *IEEE Transactions on Intelligent Transportation Systems*, 23(7):6240–6259, 2021.

[2] Jung Im Choi and Qing Tian. Adversarial attack and defense of yolo detectors in autonomous driving scenarios. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 1011–1017. IEEE, 2022.

[3] Jia Cheng Han and Zhi Quan Zhou. Metamorphic fuzz testing of autonomous vehicles. In *Proceedings of the IEEE/ACM 42nd International Conference on Software Engineering Workshops*, pages 380–385, 2020.

[4] Cumhur Erkan Tuncali, Georgios Fainekos, Danil Prokhorov, Hisahiro Ito, and James Kapinski. Requirements-driven test generation for autonomous vehicles with machine learning components. *IEEE Transactions on Intelligent Vehicles*, 5(2):265–280, 2019.

[5] Henry X Liu and Shuo Feng. Curse of rarity for autonomous vehicles. *nature communications*, 15(1):4808, 2024.

[6] Peng Chen, Haoyuan Ni, Liang Wang, Guizhen Yu, and Jian Sun. Safety performance evaluation of freeway merging areas under autonomous vehicles environment using a co-simulation platform. *Accident Analysis & Prevention*, 199:107530, 2024.

[7] Szilárd Aradi. Survey of deep reinforcement learning for motion planning of autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 23(2):740–759, 2020.

[8] Pierre-Yves Oudeyer. Computational theories of curiosity-driven learning. *arXiv preprint arXiv:1802.10546*, 2018.

[9] Han Xu, Yaxin Li, Wei Jin, and Jiliang Tang. Adversarial attacks and defenses: Frontiers, advances and practice. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3541–3542, 2020.

[10] Alankrita Aggarwal, Mamta Mittal, and Gopi Battineni. Generative adversarial network: An overview of theory and applications. *International Journal of Information Management Data Insights*, 1(1):100004, 2021.

[11] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

[12] Guillermo Owen. *Game theory*. Emerald Group Publishing, 2013.

[13] Xuan Cai, Xuesong Bai, Zhiyong Cui, Peng Hang, Haiyang Yu, and Yilong Ren. Adversarial stress test for autonomous vehicle via series reinforcement learning tasks with reward shaping. *IEEE Transactions on Intelligent Vehicles*, 2024.

[14] Peide Huang, Mengdi Xu, Fei Fang, and Ding Zhao. Robust reinforcement learning as a stackelberg game via adaptively-regularized adversarial training. *arXiv preprint arXiv:2202.09514*, 2022.

[15] Adnan Qayyum, Muhammad Usama, Junaid Qadir, and Ala Al-Fuqaha. Securing connected & autonomous vehicles: Challenges posed by adversarial machine learning and the way forward. *IEEE Communications Surveys & Tutorials*, 22(2):998–1026, 2020.

[16] Wenhao Ding, Baiming Chen, Minjun Xu, and Ding Zhao. Learning to collide: An adaptive safety-critical scenarios generating method. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2243–2250. IEEE, 2020.

[17] Wenhao Ding, Baiming Chen, Bo Li, Kim Ji Eun, and Ding Zhao. Multimodal safety-critical scenarios generation for decision-making algorithms evaluation. *IEEE Robotics and Automation Letters*, 6(2):1551–1558, 2021.

[18] Feng Ding, Keping Yu, Zonghua Gu, Xiangjun Li, and Yunqing Shi. Perceptual enhancement for autonomous vehicles: Restoring visually degraded images for context prediction via adversarial training. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):9430–9441, 2021.

[19] A Kloukiniotis, A Papandreou, A Lalos, P Kapsalas, D-V Nguyen, and K Moustakas. Countering adversarial attacks on autonomous vehicles using denoising techniques: A review. *IEEE Open Journal of Intelligent Transportation Systems*, 3:61–80, 2022.

[20] Linrui Zhang, Zhenghao Peng, Quanyi Li, and Bolei Zhou. Cat: Closed-loop adversarial training for safe end-to-end driving. In *Conference on Robot Learning*, pages 2357–2372. PMLR, 2023.

[21] Trapit Bansal, Jakub Pachocki, Szymon Sidor, Ilya Sutskever, and Igor Mordatch. Emergent complexity via multi-agent competition. *arXiv preprint arXiv:1710.03748*, 2017.

[22] Kunkun Hao, Wen Cui, Yonggang Luo, Lecheng Xie, Yuqiao Bai, Jucheng Yang, Songyang Yan, Yuxi Pan, and Zijiang Yang. Adversarial safety-critical scenario generation using naturalistic human driving priors. *IEEE Transactions on Intelligent Vehicles*, 2023.

[23] Chen Gong, Zhou Yang, Yunpeng Bai, Jieke Shi, Arunesh Sinha, Bowen Xu, David Lo, Xinwen Hou, and Guoliang Fan. Curiosity-driven and victim-aware adversarial policies. In *Proceedings of the 38th Annual Computer Security Applications Conference*, pages 186–200, 2022.

[24] Hanlin Tian, Kethan Reddy, Yuxiang Feng, Mohammed Quddus, Yiannis Demiris, and Panagiotis Angeloudis. Enhancing autonomous vehicle training with language model integration and critical scenario generation. *arXiv preprint arXiv:2404.08570*, 2024.

[25] Edouard Leurent. An environment for autonomous driving decision-making. `https://github.com/eleurent/highway-env`, 2018.

[26] Matthew O'Kelly, Aman Sinha, Hongseok Namkoong, Russ Tedrake, and John C Duchi. Scalable end-to-end autonomous vehicle testing via rare-event simulation. *Advances in neural information processing systems*, 31, 2018.

[27] Chejian Xu, Wenhao Ding, Weijie Lyu, Zuxin Liu, Shuai Wang, Yihan He, Hanjiang Hu, Ding Zhao, and Bo Li. Safebench: A benchmarking platform for safety evaluation of autonomous vehicles. *Advances in Neural Information Processing Systems*, 35:25667–25682, 2022.

[28] Baiming Chen, Xiang Chen, Qiong Wu, and Liang Li. Adversarial evaluation of autonomous vehicles in lane-change scenarios. *IEEE transactions on intelligent transportation systems*, 23(8):10333–10342, 2021.

[29] Shuo Feng, Haowei Sun, Xintao Yan, Haojie Zhu, Zhengxia Zou, Shengyin Shen, and Henry X Liu. Dense reinforcement learning for safety validation of autonomous vehicles. *Nature*, 615(7953):620–627, 2023.