RISK-SAC: Risk-Driven Soft Actor-Critic for Safety-Critical Scenario Generation in Autonomous Driving

1st Dehui Du*
dept. Software Engineering Institute
East China Normal University
Shanghai, China
dhdu@sei.ecnu.edu.cn

2nd Ermuyun Li
dept. Software Engineering Institute
East China Normal University
Shanghai, China
51275902054@stu.ecnu.edu.cn

3rd Xing Yu
dept. Software Engineering Institute
East China Normal University
Shanghai, China
51265902017@stu.ecnu.edu.cn

4th Lili Tian
dept. Software Engineering Institute
East China Normal University
Shanghai, China
52215901023@stu.ecnu.edu.cn

***Abstract*—Simulation-based testing has become an essential approach for validating autonomous driving systems. However, existing scenario generation methods often struggle to efficiently discover rare safety-critical interactions in high-dimensional traffic environments and may produce physically implausible behaviors due to unconstrained or manually designed exploration strategies. To address these limitations, we propose RISK-SAC, a risk-driven scenario generation method built upon Soft Actor-Critic (SAC) to generate diverse and physically realistic safety-critical driving scenarios. The proposed approach leverages stochastic exploration and risk-aware learning signals, together with explicit realism constraints, to guide exploration toward hazardous yet valid interaction patterns. Extensive experiments in diverse scenarios within the CARLA simulator demonstrate that RISK-SAC consistently outperforms representative baselines in terms of collision discovery efficiency, scenario diversity, and physical realism. Ablation studies further confirm the effectiveness of the proposed components in improving exploration efficiency and training stability. These results indicate that RISK-SAC provides a practical solution for efficient and realistic safety-critical scenario generation in autonomous driving.**

***Index Terms*—Safety-Critical Scenario Generation, Reinforcement Learning, Autonomous Driving Systems**

I. Introduction

Autonomous driving systems (ADSs) are widely regarded as a transformative force in future transportation, with the potential to significantly reduce accidents and improve traffic efficiency. However, despite rapid technological advances, the deployment of Level-4 ADS remains hindered by safety and reliability concerns. Real-world road testing is prohibitively expensive, time-consuming, and inherently risky, making it infeasible to rely solely on naturalistic driving to validate safety [1]. Consequently, scenario-based simulation has become the dominant paradigm for ADS testing, enabling systematic evaluation under diverse and complex traffic conditions [2], [3]. The effectiveness of this paradigm, however, critically depends on the ability to generate safety-critical scenarios. Such scenarios reside in the long tail of the scenario distribution and rarely occur under random sampling or naturalistic replay, leaving many high-risk behaviors insufficiently tested [4], [5]. This sparsity fundamentally limits coverage and slows the discovery of rare but decisive failure cases.

To address this challenge, recent work has explored adversarial and optimization-based scenario generation. Falsification and search-based approaches actively optimize scenario parameters to violate safety specifications [6]. Reinforcement learning (RL) has further gained attention due to its capability to interact with the environment and adaptively explore failure-prone regions. Compared with scripted or offline replay-based testing, closed-loop and interactive generation allows scenarios to evolve in response to the behavior of the ego vehicle, which is critical for exposing adaptive and adversarial failure modes in multi-agent traffic environments.

Nevertheless, generating realistic safety-critical scenarios remains a challenge. This difficulty arises because driving risk emerges from the joint effect of *static parameters* (e.g., initial positions, velocities, road layouts, and traffic signals) and *dynamic interactions* (e.g., acceleration, braking, and steering). Existing methods often optimize only one of these aspects or rely on loosely coupled pipelines, limiting both search efficiency and interaction realism [7]. Moreover, unconstrained exploration may exploit simulator artifacts and lead to physically implausible behaviors.

Furthermore, many existing generators rely on proxy reward formulations that only partially reflect scenario risk. However, safety-critical risk is inherently multifaceted, including proximity to collision, violation of safety specifications, and hazardous interaction patterns. Explicitly incorporating such

risk signals enables learning to focus on safety-critical regions of the scenario space rather than indirect reward objectives. From a learning perspective, targeting high-risk regions further motivates stochastic policy optimization. Deterministic methods such as Deep Deterministic Policy Gradient (DDPG) [8] often suffer from limited exploration and training instability in interactive traffic environments. In contrast, Soft Actor-Critic (SAC) [9] introduces entropy regularization to promote diversified exploration and improved stability, making it well suited for discovering rare and adversarial interactions.

Motivated by these observations, we propose **RISK-SAC**, a risk-driven reinforcement learning approach for safety-critical scenario generation in simulation. The main contributions of this work are threefold:

- We design an online generation method that jointly optimizes static scenario configurations and dynamic interactions, enabling efficient discovery of long-tail safety-critical scenarios in continuous driving environments.
- We incorporate formal safety specifications into the learning loop through robustness-aware feedback and combine them with realism regularization, allowing the generation process to explicitly target meaningful safety-critical behaviors while avoiding physically implausible solutions.
- We introduce robustness-guided reward shaping and risk-prioritized experience replay to improve exploration efficiency and training stability, resulting in faster convergence and improved scenario diversity while preserving physical realism compared with representative baselines.

II. Background

A. Operational Design Domain

The Operational Design Domain (ODD) defines the operational conditions under which an autonomous driving system is expected to function [10]. According to ISO 34503, ODD can be structured into scenario elements, environmental conditions, and dynamic elements. ODD plays a central role in scenario-based safety testing by constraining the scenario space and specifying valid operating conditions. Previous studies have investigated ODD from different perspectives, including safety-oriented specification [11], task-driven boundary refinement [12], and data-driven ODD determination based on statistical risk tolerance [13]. In this work, we leverage ODD to formalize scenario constraints, which are further translated into temporal logic specifications for online monitoring and risk evaluation during scenario generation.

B. STL Robustness and Online Monitoring

Signal Temporal Logic (STL) is widely used to formally specify temporal safety requirements in cyber-physical systems (CPS), including autonomous driving systems. Beyond Boolean satisfaction, STL robustness provides a continuous measure that quantifies the degree of specification satisfaction or violation, enabling fine-grained assessment of safety margins and early detection of dangerous behaviors. In particular, lower robustness values indicate higher risk.

Compared with offline verification, online STL monitoring supports real-time evaluation of partial execution traces and provides timely feedback in dynamic environments. Previous work has demonstrated the effectiveness of robustness-based monitoring for runtime verification and control in CPS [14]–[17]. In this work, we integrate an online STL monitor [18] into the scenario generation process and use robustness values as risk-aware feedback to guide exploration toward safety-critical and near-failure interactions.

C. Reinforcement Learning for Continuous Control

Deep reinforcement learning has been widely applied to continuous control problems, with actor–critic methods such as DDPG and SAC serving as representative approaches [19]. While DDPG adopts deterministic policies, SAC introduces entropy regularization to enable stochastic policy optimization, improving exploration robustness and training stability in complex interactive environments [20]. These properties make SAC well suited for scenario generation tasks that require discovering diverse and rare interaction patterns. Standard RL, however, does not explicitly consider safety or realism constraints. To address this limitation, constrained RL based on the Constrained Markov Decision Process (CMDP) framework has been widely studied, where Lagrangian relaxation is commonly used to incorporate auxiliary constraints into policy optimization [21], [22]. In complex multi-agent settings, balancing constraint satisfaction and exploration efficiency remains a practical challenge. Motivated by these observations, we build upon SAC and integrate risk-aware guidance together with realism constraints to support efficient, diverse, and physically plausible safety-critical scenario generation.

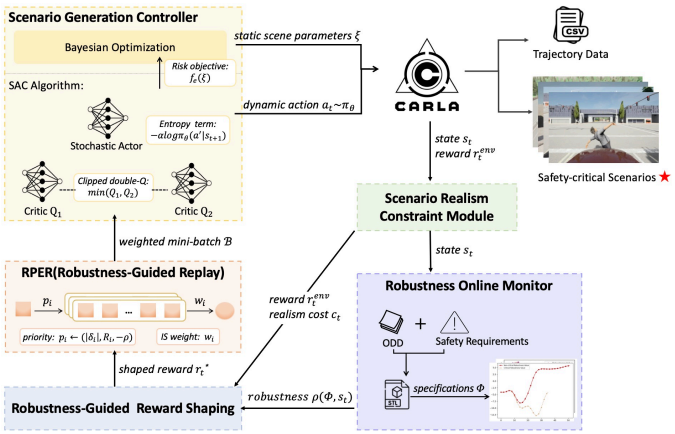


Fig. 1: Overview of RISK-SAC implementation.

III. Methodology

A. Framework Overview

As shown in Fig. 1, RISK-SAC adopts a two-stage generation pipeline. Bayesian optimization is first employed to explore the static scenario parameter space and identify high-risk initial configurations, while SAC subsequently generates continuous control actions to drive dynamic interactions. To

preserve physical feasibility and semantic validity, realism constraints are embedded into the simulation loop. During execution, trajectories are evaluated by an STL-based robustness monitor, and the resulting feedback is incorporated into policy learning via robustness-guided reward shaping and prioritized experience replay. By tightly coupling parameter search, policy optimization, and robustness feedback, RISK-SAC establishes a closed-loop optimization framework for efficient and diverse safety-critical scenario generation.

B. Scenario Realism Constraint

RISK-SAC aims to discover safety-critical driving scenarios by maximizing risk-related rewards. However, unconstrained risk optimization often leads to physically infeasible behaviors, such as overlapping initial configurations or instantaneous collisions, which produce spurious high-risk signals. To ensure physically plausible scenario generation, we explicitly incorporate realism constraints into the learning process.

We formulate scenario generation as a Constrained Markov Decision Process, where risk discovery is optimized under a realism constraint:

$$\begin{aligned} &\max_{\pi} \ \mathbb{E}_{\pi}\Big[\sum_{t}\gamma^{t}R_{risk}(s_t,a_t)\Big], \\ &\text{s.t. } \mathbb{E}_{\pi}\Big[\sum_{t}\gamma^{t}C_{real}(s_t,a_t)\Big] \le d \end{aligned} \tag{1}$$

where R_{risk} denotes the risk-driven reward, including sparse collision bonuses, dense near-miss risk signals such as time-to-collision (TTC), and safety margin violation indicators derived from RSS [23]. C_{real} represents the realism cost, γ is the discount factor, and d specifies the upper bound on the expected cumulative realism cost.

To enable efficient optimization, we apply Lagrangian relaxation and derive the following unconstrained objective:

$$\max_{\pi}\min_{\lambda\ge 0} \ \mathcal{L}(\pi,\lambda) = f(\pi) - \lambda\, g(\pi) \tag{2}$$

where $f(\pi)$ denotes the expected cumulative risk reward, $g(\pi)$ represents the realism constraint violation and $\lambda \ge 0$ is the Lagrange multiplier that adaptively controls the trade-off between risk maximization and realism constraint satisfaction.

Accordingly, the penalized reward is defined as:

$$R(s_t,a_t) = R_{risk}(s_t,a_t) - \lambda C_{real}(s_t,a_t) \tag{3}$$

Based on this formulation, the final optimization objective of RISK-SAC is defined as:

$$\max_{\pi} \ \mathbb{E}_{\pi}\Big[\sum_{t}\gamma^{t}\big(R(s_t,a_t) - \alpha\log\pi(a_t|s_t)\big)\Big] \tag{4}$$

where α is the entropy temperature coefficient controlling the exploration–exploitation trade-off.

The realism cost is composed of hard feasibility constraints and soft continuous regularization:

$$C_{real}(s_t,a_t) = C_{hard}(s_t,a_t) + C_{soft}(s_t,a_t) \tag{5}$$

The hard constraint term C_{hard} eliminates invalid scenario configurations at the episode level, such as overlapping agent initialization and physically implausible initial TTC conditions that cause instantaneous collisions. The soft constraint term C_{soft} provides realism regularization by penalizing abrupt control variations, excessive acceleration changes, and sudden extreme proximity events between agents. Together, these realism constraints enable RISK-SAC to discover diverse safety-critical scenarios while avoiding degenerate solutions caused by physically infeasible behaviors.

C. Robustness-Guided Reward Shaping

In safety-critical scenario generation, rare events such as collisions or specification violations often lead to sparse and delayed reward signals, which significantly degrade exploration efficiency. To alleviate this issue, we adopt robustness-guided reward shaping to introduce structured domain knowledge into the learning process and provide denser feedback for policy optimization. To preserve the optimal policy of the original CMDP, we follow the potential-based shaping formulation proposed by Ng et al. [24], which guarantees policy invariance. The shaping term is defined as:

$$F(s,a,s') = \gamma\phi(s') - \phi(s) \tag{6}$$

where $\phi : S \to \mathbb{R}$ denotes a state potential function and γ is the discount factor of the underlying CMDP.

STL enables principled reward shaping by providing formally grounded robustness semantics [25]–[27]. Let $\Phi = \{\varphi_i\}_{i=1}^{M}$ denote the set of STL specifications and $\rho(\varphi_i, s)$ denote the robustness of state s with respect to specification φ_i. To capture the overall safety risk while preserving worst-case sensitivity, we aggregate robustness values using a soft-min operator:

$$\rho(\Phi,s) = -\frac{1}{\beta}\log\sum_{i=1}^{M}\exp\big(-\beta\rho(\varphi_i,s)\big), \tag{7}$$

where β controls the sharpness of the approximation and larger values yield behavior closer to the minimum operator. We then define the potential function as $\phi(s) = -\rho(\Phi, s)$, where lower aggregated robustness corresponds to higher risk. Accordingly, the final shaped reward $R'(s_t, a_t)$, which is used for policy training, is constructed as:

$$\begin{aligned} R'(s_t,a_t) &= R(s_t,a_t) + F(s_t,a_t,s_{t+1}) \\ &= R(s_t,a_t) - \gamma\rho(\Phi,s_{t+1}) + \rho(\Phi,s_t), \end{aligned} \tag{8}$$

where $R(s_t, a_t)$ denotes the previously defined penalized reward. The robustness-guided shaping term encourages exploration toward states with decreasing aggregated robustness, guiding the agent to safety-critical boundary regions.

D. Robustness-Guided Prioritized Experience Replay

Experience replay improves sample efficiency and stabilizes off-policy learning by reusing past transitions. However, uniform sampling under-utilizes rare but safety-critical experiences. Prioritized Experience Replay (PER) alleviates this issue by sampling transitions based on temporal-difference (TD) errors, focusing training on samples that are more informative for value learning [28].

Although PER improves learning efficiency, it does not explicitly prioritize experiences that are most relevant to safety-critical scenario generation. To address this limitation, we propose Robustness-Guided Prioritized Experience Replay (RPER), which integrates risk awareness and formal safety semantics into the prioritization mechanism. Specifically, the priority of a replay buffer transition e_i is defined as:

$$p_i = w_1|\delta_t| + w_2\hat{R}_t + w_3\big(-\rho(\Phi, s_t)\big), \tag{9}$$

where $|\delta_t|$ denotes the TD error magnitude, $\hat{R}_t$ represents the accumulated reward and $\rho(\Phi, s_t)$ denotes the aggregated STL robustness defined above. Since lower robustness indicates proximity to specification violations, its negative value assigns higher priority to safety-critical transitions. The weights w_1, w_2, and w_3 control the relative contributions of learning progress, risk sensitivity, and safety relevance.

To correct the sampling bias introduced by prioritization, we apply standard importance sampling (IS) weights:

$$w_i = \left(\frac{1}{N \cdot P(i)}\right)^{\eta}, \tag{10}$$

where $P(i) = \frac{p_i}{\sum_j p_j}$ is the normalized sampling probability, N denotes the replay buffer size, and $\eta \in [0, 1]$ controls the strength of bias correction. This formulation increases the sampling frequency of safety-critical experiences while preserving sufficient coverage of low-priority samples, thereby ensuring stable training and convergence.

E. Safety-Critical Scenario Generation via RISK-SAC

Unlike data-driven scenario generation approaches that rely on limited real-world datasets and manually designed test cases that often fail to expose rare but hazardous situations, our framework formulates safety-critical scenario discovery as a risk-driven reinforcement learning problem. By leveraging stochastic policy optimization and maximum-entropy exploration, RISK-SAC actively samples diverse behaviors in continuous scenario and interaction spaces. Integrated with global parameter search, this design enables systematic discovery of diverse high-risk scenario modes that are difficult to anticipate through heuristic design or random sampling.

As illustrated in Algorithm 1, scenario generation is decomposed into two tightly coupled stages: static initial scene optimization and dynamic interaction generation.

i) **Static initial scene optimization.** Bayesian Optimization (BO) is employed to efficiently explore the high-dimensional continuous parameter space and identify initial scene configurations that induce high-risk interactions. To stabilize optimization across episodes with varying lengths and reward scales, we define the BO objective based on the episode-level average shaped return:

$$f_e(\xi) = \frac{\max\big(0, \widetilde{M}_e - \bar{R}_e(\xi)\big)}{|\widetilde{M}_e| + \varepsilon}, \tag{11}$$

where $\bar{R}_e = \frac{1}{\tau_e}\sum_{t=1}^{\tau_e} \tilde{r}_t$ denotes the average normalized shaped reward obtained in episode e, τ_e is the episode length, and $\widetilde{M}_e$ is a smoothed running maximum of historical returns. Minimizing $f_e(\xi)$ drives the optimizer toward parameter configurations associated with higher-risk scenarios, while the adaptive normalization term improves stability and comparability across episodes.

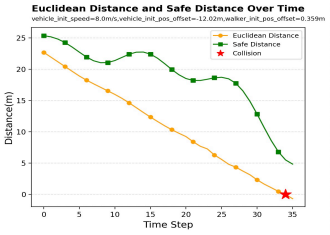
ii) **Dynamic interaction generation.** Starting from the selected initial scene, RISK-SAC samples continuous actions from a stochastic policy to promote entropy-regularized exploration. Agent behaviors are optimized under the proposed realism-constrained, risk-driven reward formulation, while robustness-guided shaping and RPER further steer exploration toward safety-critical interactions. Each episode terminates upon collision or when the maximum simulation horizon is reached.

Algorithm 1 RISK-SAC with Realism Constraint

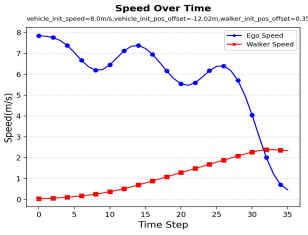
Initialize Bayesian optimizer opt, replay buffer D
Initialize RISK-SAC networks $\pi_\theta, Q_{\psi_1}, Q_{\psi_2}$, entropy temperature α and dual variable λ
for episode $e = 1$ to E **do**
  Sample initial scene $\xi_e \sim opt.\textbf{ask}()$
  Reset environment with ξ_e, observe s_0
  for $t = 1$ to T **do**
    Sample action $a_t \sim \pi_\theta(\cdot|s_t)$
    Execute a_t, observe $(s_{t+1}, r_t^{env}, done)$
    Compute realism cost c_t and penalized reward $R_t = r_t^{env} - \lambda c_t$
    Compute robustness shaping reward $r_t \leftarrow R_t - \gamma\rho(\Phi, s_{t+1}) + \rho(\Phi, s_t)$
    Compute TD error δ_t and RPER priority p_t
    Store $(s_t, a_t, r_t, c_t, s_{t+1}, done, p_t)$ into D
    Sample mini-batch $\mathcal{B}$ from D with RPER and IS weights
    Update $\pi_\theta, Q_{\psi_1}, Q_{\psi_2}, \alpha$ using SAC losses
    Compute $\bar{C}_{real}$ as the mini-batch mean realism cost
    Update dual variable: $\lambda \leftarrow \big[\lambda + \eta_\lambda(\bar{C}_{real} - d)\big]_+$
    if $done$ **then**
      break
    end if
  end for
  Compute episode score $\bar{R}_e$ and BO objective $f_e(\xi_e)$
  $opt.\textbf{tell}(\xi_e, f_e)$
end for

The two stages are connected through a closed-loop feedback mechanism. After each episode, the episode-level average shaped return is fed back to the Bayesian optimizer as the evaluation signal for updating the parameter search model. To avoid instability caused by mismatched update frequencies, BO is updated once per episode, while RISK-SAC performs step-wise policy updates within each episode. This hierarchical optimization scheme enables coordinated exploration of both static scene parameters and dynamic interaction behaviors.

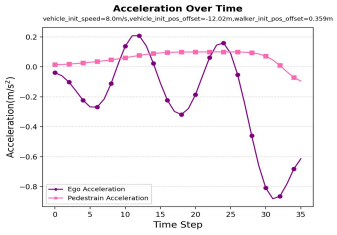
(a) Safety-critical scenario



(b) Distance over time (collision case)

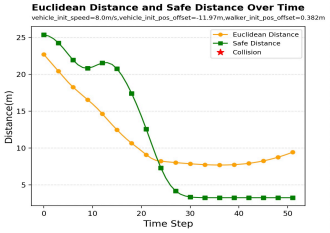


(c) Speed over time

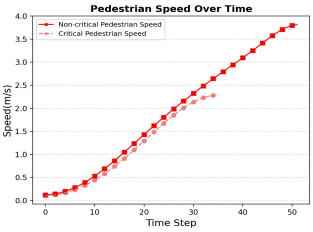


(d) Acceleration Over Time

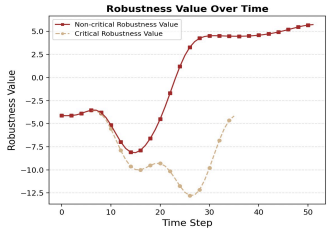
(e) Non-critical scenario



(f) Distance over time (safe case)



(g) Pedestrian speed comparison



(h) Robustness comparison

Fig. 2: Visualization of safety-critical and non-critical pedestrian crossing scenarios generated by RISK-SAC in CARLA

IV. Experiments and Analysis

A. Setup

1) Environment: All experiments are conducted in the CARLA simulator, a high-fidelity autonomous driving platform widely used for safety evaluation and reinforcement learning research. CARLA provides realistic urban traffic environments with configurable road layouts and dynamic agents, enabling reproducible evaluation of safety-critical interactions. The built-in CARLA autonomous driving agent is used as the ego controller, while surrounding traffic participants are driven by RISK-SAC policies. This setup allows us to assess the ability of RISK-SAC to expose failure modes of a realistic ADS under diverse interaction conditions.

2) Evaluation Scenarios: To evaluate the performance of RISK-SAC under different traffic interaction complexities, we design three categories of driving scenarios according to the number of interacting agents and road topology. Low-interaction scenarios correspond to urban car-following on straight roads with a limited number of vehicles, focusing on longitudinal interactions such as braking and speed adaptation. Mixed-interaction scenarios involve pedestrian crossing events at intersections or crosswalks, introducing heterogeneous interactions between vehicles and vulnerable road users. Dense multi-agent scenarios represent intersection environments with crossing traffic flows, traffic lights, and turning maneuvers, forming highly coupled interaction patterns. By organizing scenarios with progressively increasing interaction complexity, we systematically evaluate RISK-SAC under diverse traffic conditions and compare its scenario generation performance across different interaction settings.

3) Baselines and Metrics: To enable controlled comparison and isolate the contribution of different algorithmic components, we construct a set of representative baselines. DDPG [8] and SAC [9] are included as deterministic and stochastic continuous-control baselines, together with their constrained variants, DDPG-Lag and SAC-Lag, which incorporate realism costs via Lagrangian relaxation. In addition, two ablation variants are considered: RISK-SAC (only-RPER) without reward shaping and RISK-SAC (only-RS) without prioritized experience replay. The complete RISK-SAC framework is used as the final comparison target.

Correspondingly, we adopt evaluation metrics to assess performance from multiple complementary aspects. Safety criticality is measured by Effective Collision Rate (ECR), defined as the proportion of generated scenarios that result in valid collision events after filtering physically infeasible or degenerate cases. During training, ECR is computed using a sliding window to reflect online convergence behavior. Generation efficiency is evaluated using an Efficiency Score (ES), derived from the average number of steps required to trigger a critical scenario and normalized so that higher values indicate better efficiency. Scenario realism is evaluated using $\bar{\rho}_c$ [22], defined as the average realism cost rate computed from the overall realism regularization term C_{real}. Finally, following prior studies [29], we measure scenario diversity using Outcome Feature Dispersion (OFD), which quantifies the dispersion of risk-related outcome features (e.g., minimum TTC, relative speed, lateral deviation, and STL robustness) across discovered safety-critical scenarios.

B. Experiment Results

A representative safety-critical pedestrian crossing scenario generated by RISK-SAC is shown in Fig. 2a. This case is obtained after policy convergence and illustrates a realistic interaction pattern leading to collision. Fig. 2b–2d show the corresponding distance, speed, and acceleration time series.

Trajectory analysis indicates that the ego vehicle initially decelerates after detecting braking behavior from a truck in

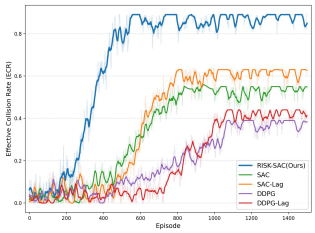
(a) Near rear-end collision

(b) Pedestrian collision

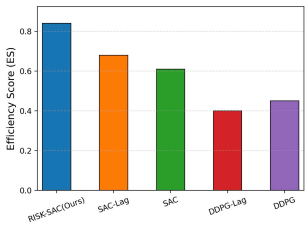
(c) Intersection collision (case I)

(d) Intersection collision (case II)

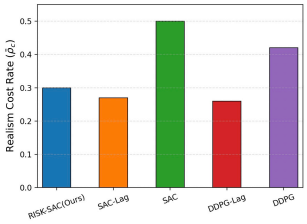
Fig. 3: Discovered safety-critical scenarios in CARLA Town10



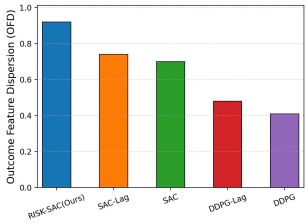
(a) Training convergence of ECR



(b) Normalized comparison of ES



(c) Normalized comparison of $\bar{\rho}_c$



(d) Normalized comparison of OFD

Fig. 4: Performance comparison of different methods under dense intersection scenarios

the adjacent lane. During this phase, the pedestrian moves slowly and remains within the ego vehicle's sensing blind region for an extended period. Consequently, the pedestrian is not immediately perceived, leading the planning module to assume a clear crossing path and resume acceleration to pass the green-light intersection. When the pedestrian finally enters the perception range around $t = 25$, emergency braking is triggered. However, the remaining stopping distance is insufficient, resulting in a collision at approximately $t = 34$. This failure case reflects a realistic perception–decision–reaction coupling, rather than being caused by artificially aggressive or physically implausible control actions.

For comparison, Fig. 2e–2f present a paired non-collision scenario generated under similar initial conditions. As shown in Fig. 2g, the pedestrian in this case enters the perception range earlier due to higher walking speed, enabling the ego vehicle to initiate braking in time and come to a complete stop. The robustness comparison in Fig. 2h further shows that the critical case consistently exhibits lower robustness values, indicating closer proximity to specification violation boundaries. Overall, these results demonstrate that RISK-SAC is able to generate safety-critical scenarios arising from realistic multi-agent interaction dynamics.

We further evaluate RISK-SAC across different scenario complexity levels. Fig. 3 presents representative safety-critical cases in low-interaction, mixed-interaction, and dense traffic settings, covering typical hazardous patterns such as rear-end collisions, vehicle–pedestrian conflicts, and intersection conflicts. Since dense intersections involve tightly coupled interactions and complex right-of-way constraints, we focus on this setting for quantitative evaluation.

Fig. 4a shows the training convergence of ECR under dense intersection scenarios. RISK-SAC consistently achieves the highest ECR and converges substantially faster than all baselines, indicating more efficient and stable discovery of safety-critical interactions. Compared with Lagrangian-based methods, SAC-Lag and DDPG-Lag exhibit slower early-stage convergence, which can be attributed to the additional non-stationarity introduced by dual variable updates before the Lagrange multiplier stabilizes. After convergence of λ, their ECR improves, but remains notably lower than that of RISK-SAC. This result highlights the advantage of robustness-guided reward shaping and prioritized experience replay in accelerating the accumulation of informative high-risk trajectories. The comparison between stochastic and deterministic policies further reveals clear performance differences. As shown in Fig. 4a, SAC-based methods (RISK-SAC and SAC) consistently outperform DDPG variants in both convergence speed and final ECR. This is mainly due to entropy-regularized exploration, which maintains action diversity and prevents premature convergence to narrow failure modes. Such exploration behavior is particularly critical in dense multi-agent environments, where rare safety-critical interactions often arise from complex and tightly coupled agent behaviors.

To further analyze the trade-offs among efficiency, realism, and diversity, Fig. 4b–4d present a multi-objective comparison of ES, $\bar{\rho}_c$, and OFD. Overall, RISK-SAC achieves the most balanced performance across these metrics. As shown in Fig. 4b, RISK-SAC attains the highest efficiency score, indicating faster discovery of safety-critical events with fewer interaction steps. Meanwhile, Fig. 4c shows that its realism cost remains comparable to constrained baselines such as SAC-Lag, suggesting that improved exploration efficiency does not come at the expense of severe physical infeasibility. Although RISK-SAC occasionally permits mild violations of soft realism constraints C_{soft} (e.g., emergency braking or close-proximity interactions) to expose critical behaviors, these deviations are

TABLE I: Ablation study of RISK-SAC across different scenario complexity levels

Method	Scenario Type											
	Low Interaction				Mixed Interaction				Dense Interaction			
	ECR ↑	ES ↑	$\bar{\rho}_c$ ↓	OFD ↑	ECR ↑	ES ↑	$\bar{\rho}_c$ ↓	OFD ↑	ECR ↑	ES ↑	$\bar{\rho}_c$ ↓	OFD ↑
SAC [9]	0.74	0.76	0.25	0.62	0.65	0.66	0.37	0.66	0.55	0.61	0.50	0.70
SAC-Lag	0.81	0.78	**0.12**	0.66	0.70	0.74	0.22	0.71	0.63	0.68	**0.27**	0.74
RISK-SAC (only-RPER)	0.92	0.81	0.14	0.69	0.88	0.76	**0.21**	0.72	0.85	0.73	0.28	0.78
RISK-SAC (only-RS)	0.89	0.86	0.14	0.77	0.84	0.82	0.23	0.80	0.81	0.79	0.29	0.89
RISK-SAC (Ours)	**0.97**	**0.94**	0.15	**0.82**	**0.92**	**0.88**	0.25	**0.85**	**0.89**	**0.84**	0.30	**0.92**

explicitly regularized and bounded, preserving overall physical plausibility. In contrast, the unconstrained SAC baseline, while achieving relatively high efficiency, incurs a substantially higher realism cost, indicating that its generated scenarios often rely on aggressive or physically less plausible behaviors. Lagrangian-based methods achieve lower realism cost, but at the expense of reduced exploration efficiency and lower ECR. Regarding diversity, RISK-SAC also achieves the highest OFD score, demonstrating its ability to uncover diverse failure patterns rather than repeatedly exploiting a limited subset of critical behaviors. This advantage becomes particularly evident in dense interaction settings, where naive exploration strategies tend to collapse into repetitive adversarial patterns.

Finally, we conduct ablation experiments to analyze the contribution of individual components, as summarized in Table I. Across all scenario complexity levels, the full RISK-SAC consistently achieves the highest ECR and efficiency score, confirming the complementary effects of robustness-guided reward shaping and risk-prioritized experience replay. Removing either component leads to noticeable performance degradation. Comparing the two partial variants, RPER primarily improves training stability and effective reuse of informative high-risk transitions, resulting in higher sustained ECR after convergence, while robustness-guided reward shaping directly steers exploration toward safety-critical regions, contributing more to early risk targeting and diversity expansion. When combined, these two mechanisms jointly enable efficient and stable discovery of diverse hazardous interactions. Moreover, both RISK-SAC variants significantly outperform SAC and SAC-Lag, especially in dense interaction settings, highlighting the advantage of risk-driven optimization over generic exploration and purely constraint-driven baselines. Finally, the full RISK-SAC achieves the highest OFD, with the performance gap becoming more pronounced as scenario complexity increases, demonstrating its stronger capability in covering diverse safety-critical interaction patterns.

V. Related Work

Safety-critical scenario generation has been widely studied to expose failures of ADS. Existing approaches are mainly categorized into data-driven, rule-based, and search-based methods. Data-driven methods leverage naturalistic driving data [30]–[32] to improve realism but are limited by data coverage, making rare long-tail hazards difficult to discover. Rule-based approaches such as LawBreaker [33] and TARGET [34] offer strong interpretability, but suffer from limited scalability and expressiveness in complex scenarios.

Search-based and reinforcement learning-based methods have therefore attracted increasing attention due to their ability to actively explore high-risk regions of the scenario space [7], [35]–[39]. Representative works formulate scenario generation as an optimization problem and employ RL to discover adversarial interactions, such as actor–critic-based lane-change risk exploration [35] and multi-objective adversarial scenario generation in Baidu Apollo [36]. However, existing methods often rely on heuristic reward design and lack explicit realism modeling, which can lead to inefficient exploration, limited behavioral diversity, and physically implausible scenarios. In contrast, RISK-SAC systematically addresses these challenges through a unified risk-driven learning framework.

VI. Conclusion

In this paper, we proposed RISK-SAC, a risk-driven reinforcement learning approach for efficient and realistic safety-critical scenario generation. By combining entropy-regularized exploration with robustness-guided reward shaping, risk-prioritized experience replay, and realism-aware constraints, RISK-SAC enables targeted exploration of high-risk regions while preserving physical feasibility and scenario diversity. Extensive experiments and ablation studies demonstrate its superiority over representative baselines in discovery efficiency, realism, and diversity. Future work will explore integrating the proposed method with real-world autonomous driving algorithms for closed-loop evaluation under safety-critical scenarios, as a step toward addressing sim-to-real generalization.

References

[1] D. Zhao, H. Lam, H. Peng, S. Bao, D. J. LeBlanc, K. Nobukawa, and C. S. Pan, "Accelerated evaluation of automated vehicles safety in lane-change scenarios based on importance sampling techniques," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 3, pp. 595–607, 2017.

[2] C. Birchler, S. Khatiri, P. Rani, T. Kehrer, and S. Panichella, "A roadmap for simulation-based testing of autonomous cyber-physical systems: Challenges and future direction," *ACM Trans. Softw. Eng. Methodol.*, vol. 34, no. 5, pp. 1–9, May 2025.

[3] D. Nalic, T. Mihalj, M. Baeumler, M. Lehmann, A. Eichberger, and S. Bernsteiner, "Scenario based testing of automated driving systems: A literature survey," in *FISITA web Congress*, 10 2020.

[4] N. Kalra and S. M. Paddock, "Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?" *Transportation Research Part A: Policy and Practice*, vol. 94, pp. 182–193, 2016.

[5] Y. Mei, T. Nie, J. Sun, and Y. Tian, "Seeking to collide: Online safety-critical scenario generation for autonomous driving with retrieval augmented large language models," in *Proceedings of the 2025 IEEE 28th International Conference on Intelligent Transportation Systems (ITSC)*, Gold Coast, Australia, Nov. 2025, accepted, to appear.

[6] D. Karunakaran, S. Worrall, and E. Nebot, "Efficient statistical validation with edge cases to evaluate highly automated vehicles," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, 2020, pp. 1–8.

[7] Z. Wang, X. Li, D. Wei, L. Wang, and Y. Huang, "Efficient generation of safety-critical scenarios combining dynamic and static scenario parameters," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 12, pp. 7812–7827, 2024.

[8] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[9] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*. Pmlr, 2018, pp. 1861–1870.

[10] Society of Automotive Engineers, "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles," SAE International Standard J3016_202104, 2021, warrendale, Pennsylvania. [Online]. Available: https://www.sae.org/standards/content/j3016_202104/

[11] M. Ito, "Odd description methods for automated driving vehicle and verifiability for safety," *JUCS - Journal of Universal Computer Science*, vol. 27, pp. 796–810, 08 2021.

[12] C. Sun, Z. Deng, W. Chu, S. Li, and D. Cao, "Acclimatizing the operational design domain for autonomous driving systems," *IEEE Intelligent Transportation Systems Magazine*, vol. 14, no. 2, pp. 10–24, 2022.

[13] C. W. Lee, N. Nayeer, D. E. Garcia, A. Agrawal, and B. Liu, "Identifying the operational design domain for an automated driving system through assessed risk," in *2020 IEEE Intelligent Vehicles Symposium (IV)*, 2020, pp. 1317–1322.

[14] C. Agbo and H. Mehrpouyan, "Resilience of industrial control systems using signal temporal logic and autotuning mechanism," in *2023 IEEE Intl Conf on Dependable, Autonomic and Secure Computing (DASC)*, 2023, pp. 284–293.

[15] Y. Yao, J. Sun, and Y. Zhang, "Multitask synthesis of hybrid systems via temporal logic," *IEEE Transactions on Automatic Control*, vol. 68, no. 11, pp. 6883–6890, 2023.

[16] Z. Huang, W. Lan, and X. Yu, "A formal control framework of autonomous vehicle for signal temporal logic tasks and obstacle avoidance," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 1930–1940, 2024.

[17] A. Donzé, T. Ferrère, and O. Maler, "Efficient robust monitoring for stl," vol. 8044, 07 2013.

[18] T. G. Dehui Du and M. Zhang, "A statistical model detection method based on signal temporal logic online monitor," China Patent CN111523225A, 2020, retrieved from http://www.sipo.gov.cn/.

[19] X. Wang, S. Wang, X. Liang, D. Zhao, J. Huang, X. Xu, B. Dai, and Q. Miao, "Deep reinforcement learning: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 4, pp. 5064–5078, 2024.

[20] Y. Park, W. Jun, and S. Lee, "A comparative study of deep reinforcement learning algorithms for urban autonomous driving: Addressing the geographic and regulatory challenges in carla," *Applied Sciences*, vol. 15, no. 12, 2025.

[21] A. Wachi, X. Shen, and Y. Sui, "A survey of constraint formulations in safe reinforcement learning," ser. IJCAI '24, 2024. [Online]. Available: https://doi.org/10.24963/ijcai.2024/913

[22] J. Ji, B. Zhang, J. Zhou, X. Pan, W. Huang, R. Sun, Y. Geng, Y. Zhong, J. Dai, and Y. Yang, "Safety-gymnasium: a unified safe reinforcement learning benchmark," ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023.

[23] B. Gassmann, F. Oboril, C. Buerkle, S. Liu, S. Yan, M. S. Elli *et al.*, "Towards standardization of av safety: C++ library for responsibility sensitive safety," in *2019 IEEE Intelligent Vehicles Symposium (IV)*, 2019, pp. 2265–2271.

[24] A. Y. Ng, D. Harada, and S. J. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in *Proceedings of the Sixteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999, pp. 278–287.

[25] L. Lindemann, C. K. Verginis, and D. V. Dimarogonas, "Prescribed performance control for signal temporal logic specifications," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, 2017, pp. 2997–3002.

[26] A. Donzé and O. Maler, "Robust satisfaction of temporal logic over real-valued signals," in *Formal Modeling and Analysis of Timed Systems*, K. Chatterjee and T. A. Henzinger, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 92–106.

[27] J. Chen, Y. Zou, and S. Li, "Safe reinforcement learning for signal temporal logic tasks using robust control barrier functions," in *2023 42nd Chinese Control Conference (CCC)*, 2023, pp. 8627–8632.

[28] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," *CoRR*, vol. abs/1511.05952, 2015. [Online]. Available: https://arxiv.org/abs/1511.05952

[29] W. Ding, C. Xu, M. Arief, H. ming Lin, B. Li, and D. Zhao, "A survey on safety-critical driving scenario generation—a methodological perspective," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, pp. 6971–6988, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247159072

[30] F. Montanari, C. Stadler, J. Sichermann, R. German, and A. Djanatliev, "Maneuver-based resimulation of driving scenarios based on real driving data," in *2021 IEEE Intelligent Vehicles Symposium (IV)*, 2021, pp. 1124–1131.

[31] D. Guo, M. M. Sánchez, E. de Gelder, and T. P. van der Sande, "Scenario extraction from a large real-world dataset for the assessment of automated vehicles," in *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, Sep. 2023, pp. 1570–1575.

[32] A. Gambi, T. Huynh, and G. Fraser, "Generating effective test cases for self-driving cars from police reports," in *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ser. ESEC/FSE 2019. New York, NY, USA: Association for Computing Machinery, 2019, pp. 257–267.

[33] Y. Sun, C. M. Poskitt, J. Sun, Y. Chen, and Z. Yang, "Lawbreaker: An approach for specifying traffic laws and fuzzing autonomous vehicles," in *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, ser. ASE '22. ACM, Oct. 2022, pp. 1–12.

[34] Y. Deng, Z. Tu, J. Yao, M. Zhang, T. Zhang, and X. Zheng, "Target: Traffic rule-based test generation for autonomous driving via validated llm-guided knowledge extraction," *IEEE Transactions on Software Engineering*, vol. 51, no. 7, pp. 1950–1968, 2025.

[35] B. Chen, X. Chen, Q. Wu, and L. Li, "Adversarial evaluation of autonomous vehicles in lane-change scenarios," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 10 333–10 342, 2021.

[36] H. Tian, Y. Jiang, G. Wu, J. Yan, J. Wei, and et al., "Mosat: finding safety violations of autonomous driving systems using multi-objective genetic algorithm," in *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ser. ESEC/FSE 2022. New York, NY, USA: Association for Computing Machinery, 2022, pp. 94–106.

[37] Y. Luo, X.-Y. Zhang, P. Arcaini, Z. Jin, H. Zhao, F. Ishikawa, R. Wu, and T. Xie, "Targeting requirements violations of autonomous driving systems by dynamic evolutionary search," in *2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 2021, pp. 279–291.

[38] Z. Zhong, G. Kaiser, and B. Ray, "Neural network guided evolutionary fuzzing for finding traffic violations of autonomous vehicles," *IEEE Transactions on Software Engineering*, vol. 49, no. 4, pp. 1860–1875, 2023.

[39] J. Zhou, S. Tang, Y. Guo, Y.-F. Li, and Y. Xue, "From collision to verdict: Responsibility attribution for autonomous driving systems testing," in *2023 IEEE 34th International Symposium on Software Reliability Engineering (ISSRE)*, 2023, pp. 321–332.