IEEE Open Journal of Intelligent Transportation Systems

Received 20 October 2025; revised 11 December 2025 and 13 April 2026; accepted 10 May 2026. Date of publication 12 May 2026; date of current version 19 May 2026.

Digital Object Identifier 10.1109/OJITS.2026.3692749

Generating Test Cases for Autonomous Vehicles With Controllable Levels of Difficulty

ZIYU WANG, JING MA (Member, IEEE), AND EDMUND M.-K. LAI (Life Senior Member, IEEE)

Department of Data Science and AI, Auckland University of Technology, Auckland 1010, New Zealand

CORRESPONDING AUTHOR: J. MA (jing.ma@aut.ac.nz)

This work was supported in part by Project 111 under Grant D23006.

ABSTRACT Autonomous vehicle (AV) safety validation increasingly relies on scenario-based testing. However, existing approaches to test scenario generation do not provide mechanisms to systematically regulate scenario difficulty. To address this critical limitation, this paper introduces a novel game-theoretic framework for adversarial safety validation. The interaction between the AV-under-test and a strategically obstructing rear vehicle is modelled as a Stackelberg game. The level of adversarial intensity, which reflects the level of difficulty, of a scenario can be controlled by a single tunable parameter called aggressiveness at both the action level and the interaction level. The efficacy of this approach is studied through the highway lane-changing operational design domain. Simulation experiments demonstrate that increasing aggressiveness reduces the success rate of lane-changing and prolongs maneuver duration for the successful attempts. These results confirm that this parameter can effectively and systematically control the difficulty level of test scenarios, providing a valuable tool for rigorous and reproducible AV safety validation.

INDEX TERMS Autonomous vehicles, test scenario generation, adversarial testing, Stackelberg game, controllable difficulty level.

I. INTRODUCTION

ENSURING the safety and reliability of autonomous vehicles (AVs) remains one of the most critical and unresolved challenges impeding their widespread deployment. Conventional validation methodologies predominantly rely on the extensive accumulation of real-world driving data, aiming to statistically demonstrate system robustness across diverse operational domains. However, this brute-force approach has become increasingly impractical and inefficient, primarily due to the prohibitive time requirements and the low probability of encountering rare but safety-critical edge cases that are essential for rigorous validation [1].

More recently, in response to these limitations, there is a shift towards scenario-based testing frameworks. This methodology prioritizes the systematic design, and evaluation of representative and high-risk traffic scenarios, which could be either derived from empirical incident data or synthesized based on known failure modes. A key advantage over conventional methods is that critical interactions could be easily reproduced under controlled and repeatable conditions, significantly improving test efficiency. Moreover, scenario-based testing enables a targeted evaluation of AV performance in complex operational design domains (ODDs), where dynamic interactions and environmental factors can profoundly influence safety outcomes [2]. The effectiveness of this methodology, however, is strongly dependent on the quality of the test scenarios.

Two main types of scenarios have been the focus of previous research in this area. The first type emphasizes realism. Data are collected through cameras and other sensors from actual road traffic, resulting in naturalistic datasets such as NGSIM [3], HighD [4], and INTERACTION [5]. Driving scenarios are then generated by sampling the distribution of such traffic interactions [6], [7]. The advantage of this method is that the AV could be tested under normal driving conditions. However, to test the AV in critical boundary scenarios is difficult since such incidents are rare in these datasets. As a result, researchers have been focusing on the second type of scenarios, which are critical scenarios [8]. Such scenarios could be defined using rule-based methods. They are exemplified by national Highway Traffic Safety Administration (NHTSA) and the European New Car Assessment Programme (EuroNCAP) protocols [9], [10] that provide standardized safety-critical situations. They could provide repeatability and regulatory alignment [11], [12]. But the lack

The review of this article was arranged by Associate Editor Jianwu Fang.

of variability of such scenarios is the main disadvantage. This shortcoming could be mitigated by using randomized and parameter-sampling approaches, often implemented through traffic simulators such as SUMO [13] or VISSIM [14]. They expand coverage by stochastically varying vehicle configurations and traffic flows.

One aspect of scenario-based testing that is missing so far is the ability to provide systematic and comprehensive testing of the level of driving capability of the AV. In this paper, we propose a game-theoretic approach that is able to create test scenarios for an ODD at different levels of difficulty. This difficulty level is determined by a single parameter α which we call "aggressiveness". Similar to human driving being described as aggressive, conservative, or normal, α directly affects the behaviour of the Background Vehicles (BVs) in the scenario, making it easier or more difficult for the AV to execute its maneuvers. Game-theoretic methods are able to capture strategic interaction among vehicles and provide good interpretability. Previous use of these methods have focused on improving AV decision-making robustness [15], [16]. They have not been used for test scenario generation. Our study will focus on the highway lane-changing ODD. In this case, α serves to adjust the aggressiveness of a single BV. In our simulation experiments, we empirically validated two complementary definitions of aggressiveness, namely acceleration-focused and distance-focused formulations. Results show that, unlike prior categorical or heuristic approaches, α is able to modulate behaviours across both the action and the interaction levels.

The main contributions of this paper can be summarized as follows:

- We propose a game-theoretic scenario generation framework for autonomous vehicle testing, which enables the systematic construction of interactive and adversarial lane-changing scenarios.
- We introduce a controllable aggressiveness parameter α that serves as a unified and interpretable variable to modulate the difficulty of generated scenarios in a continuous manner.
- We design two distinct definitions of aggressiveness, namely acceleration-focused and distance-focused formulations, which capture different behavioral mechanisms of adversarial driving and provide flexible control over scenario characteristics.
- We conduct extensive simulation experiments demonstrating that increasing α leads to systematically higher scenario difficulty, validating the effectiveness and controllability of the proposed framework.

The remainder of this paper is organized as follows. Section II reviews related research on scenario generation, traffic complexity modeling, driving behaviour and aggressiveness, and applications of game theory in autonomous driving. Section III introduces the game-theoretic formulation of the lane-changing ODD. The two aggressiveness definitions and their integration into the payoff structure is explained in Section IV. Section V describes the experimental design and evaluation metrics. This is followed by the presentation and analysis of the results in Section VI. Section VII discusses the broader implications and potential extensions of the proposed framework. Finally, Section VIII concludes the paper and outlines directions for future research.

II. RELATED WORKS

A. TEST SCENARIO GENERATION METHODS

The generation of representative and challenging scenarios is a cornerstone in the testing and validation of AVs. Existing approaches can broadly be classified into three categories: rule-based, randomized, and data-driven. Each of them offers distinct advantages and limitations.

1) RULE-BASED SCENARIO GENERATION

Early efforts in scenario design often relied on standardized protocols such as those developed by the National Highway Traffic Safety Administration (NHTSA) and the European New Car Assessment Programme (EuroNCAP). These frameworks prescribe canonical crash-imminent or safety-critical configurations, enabling repeatable and comparable assessments [17], [18]. As such, the parameters such as relative positions, velocities and behaviour of the BVs remain unchanged from one test to another. Consequently, an AV could be trained specifically to pass such tests. Therefore, this approach is inherently limited in diversity. They also fail to capture the full spectrum of real-world interactions, particularly adversarial and rare cases. More recently, adaptive test cases are preferred [2].

However, despite their interpretability and ease of implementation, rule-based methods remain fundamentally constrained by their reliance on predefined traffic templates and manually designed interaction rules. These handcrafted scenarios often oversimplify complex human driving behaviors and fail to reflect the adaptive and uncertain nature of real-world interactions. As a result, they tend to produce highly structured environments that lack behavioral variability, making it difficult to assess the robustness of an autonomous driving system beyond narrowly defined conditions. Furthermore, because rule-based scenarios are typically deterministic and standardized, they allow models to overfit to specific benchmark configurations rather than developing generalizable decision-making capabilities. The limited ability of such methods to represent rare, adversarial, or emergent behaviors underscores their insufficiency for testing the adaptability and resilience of advanced autonomous systems.

2) RANDOMIZED AND PARAMETER SAMPLING APPROACHES

To overcome the rigidity of fixed test cases, researchers have employed traffic simulation platforms such as SUMO and VISSIM to stochastically vary initial conditions [10]. By randomizing parameters such as relative vehicle positions, velocities, or flow densities, these approaches expand coverage of the test space and uncover corner cases [19]. However, the randomness of case generation introduces significant uncertainty, making it difficult to systematically

control scenario difficulty or guarantee the inclusion of critical interactions [20].

Although randomization increases scenario diversity and enhances the likelihood of exposing unexpected situations, it also leads to several important drawbacks. First, purely stochastic sampling often produces a large number of trivial or redundant cases that contribute little to meaningful system evaluation, resulting in high computational costs with limited testing efficiency. Second, because the scenario parameters are drawn from broad probability distributions without explicit behavioral intent, it becomes challenging to reproduce specific critical conditions or to quantify how changes in individual parameters influence scenario difficulty. This lack of controllability limits the interpretability of the testing outcomes and prevents the establishment of a consistent mapping between scenario configuration and performance metrics. Moreover, excessive randomness can obscure causal relationships between agent behaviors and outcomes, making it difficult to diagnose failure cases or identify underlying weaknesses of the autonomous driving system. In practice, these issues hinder the use of randomized approaches as a reliable method for systematic robustness validation and fine-grained difficulty adjustment in scenario-based testing.

3) DATA-DRIVEN SCENARIO GENERATION

With the availability of high-resolution naturalistic driving datasets such as HighD, NGSIM, and INTERACTION, data-driven methods have gained prominence. These approaches reconstruct realistic traffic scenes by learning distributions of vehicle trajectories and interaction patterns [21], [22]. More recent work has extended this paradigm by categorizing data-driven approaches into diversity-focused and criticality-focused generation, highlighting the trade-off between coverage and safety relevance [20]. Despite their realism, purely data-driven methods are bounded by the underlying dataset, limiting their ability to generate out-of-distribution or adversarial scenarios necessary for robustness testing [23].

Although data-driven methods provide high-fidelity reproductions of real-world traffic interactions, they also face several critical limitations. First, their performance and generalizability are inherently restricted by the coverage and balance of the source datasets. Since these datasets are collected under specific road, traffic, and cultural conditions, the resulting models often inherit dataset biases and fail to generalize to unseen driving environments or rare conflict cases. Second, data-driven approaches emphasize statistical representativeness but lack explicit control over scenario intent or difficulty. Consequently, it is difficult to systematically adjust aggressiveness, cooperation levels, or risk exposure, which limits their usefulness in controlled safety validation. Third, the heavy dependence on supervised learning and trajectory prediction models introduces uncertainty propagation from perception and labeling errors, leading to unrealistic behaviors when extrapolated beyond the training distribution. Moreover, because the data-driven paradigm focuses primarily on imitation of existing behaviors, it struggles to generate truly adversarial or safety-critical interactions that test the upper limits of autonomous vehicle decision-making. These issues collectively restrict the ability of data-driven methods to explore the full behavioral spectrum of road users and to provide controllable, stress-testing scenarios necessary for evaluating robustness and reliability.

Overall, despite substantial progress achieved through rule-based, randomized, and data-driven scenario generation methods, each paradigm still exhibits inherent limitations that constrain its effectiveness in comprehensive testing of autonomous vehicles. Rule-based frameworks ensure reproducibility and interpretability but lack behavioral richness and adaptability to unseen situations. Randomized methods enhance diversity and increase the likelihood of uncovering rare events, yet they suffer from uncontrollable variability, making it difficult to correlate parameter changes with scenario difficulty in a consistent manner. Data-driven approaches improve realism by replicating naturalistic interactions, but they remain confined by the coverage of existing datasets and cannot easily synthesize novel or adversarial situations beyond recorded data. Collectively, these limitations reveal the absence of a unified mechanism that can systematically and continuously regulate scenario difficulty while preserving behavioral plausibility and experimental reproducibility. The framework proposed in this study directly addresses this gap by introducing a controllable aggressiveness parameter that enables adaptive adjustment of scenario complexity and interaction intensity.

B. TRAFFIC SCENARIO COMPLEXITY AND DIFFICULTY MODELING

In parallel with efforts on scenario generation, a critical research direction focuses on quantifying and modelling the complexity of traffic scenarios, which is essential for benchmarking AV performances under systematically varied conditions. Two broad classes of approaches could be identified–metric-based complexity modeling and interaction-focused complexity modeling.

1) METRIC-BASED COMPLEXITY MODELLING

Early studies primarily adopted physical and kinematic indicators such as time-to-collision (TTC), time headway (THW), and deceleration rate to avoid crash (DRAC) to characterize traffic risk and scenario difficulty. These indicators are attractive due to their interpretability and ease of computation, and they remain widely used in safety analysis and surrogate risk assessments [24], [25]. However, such measures are often computed in a static or snapshot-based manner, reflecting instantaneous proximity or collision likelihood rather than the evolving dynamics of interaction. As a result, while they provide valuable retrospective evaluations, they are less effective as controllable levers for systematically adjusting scenario difficulty.

2) INTERACTION-FOCUSED COMPLEXITY MODELING

More recent research highlights that traffic complexity emerges not only from physical proximities but also from

the interactional structure between agents. Metrics such as "interaction intensity" or "interaction risk" explicitly consider how multi-vehicle trajectories and behavioral responses shape scenario difficulty. For instance, Cheng et al. [26] surveyed risk measures in mixed traffic and noted the limitations of single-dimensional TTC-like indicators when capturing heterogeneous interactions. Jiang et al. [27] introduced an Interactive Risk (IR) metric, combining multimodal trajectory prediction with a dynamic risk field, thereby accounting for anticipatory and multi-agent interactions. Such approaches offer a more holistic characterization of scenario difficulty, as they explicitly capture the behavioral coupling between vehicles. In addition, recent studies have further advanced scenario-based testing by addressing perception robustness and adaptability to unseen environments. Iqbal et al. [28] proposed a generative multi-modal sensor fusion framework for novelty detection in autonomous driving, enabling vehicles to identify and adapt to previously unseen or abnormal traffic scenarios. Their study underscores the importance of incorporating out-of-distribution detection and adaptability into testing frameworks, complementing scenario generation research by addressing the perceptual dimension of robustness evaluation.

Despite these advances, existing complexity models are predominantly passive in nature, focusing on post hoc evaluation of safety or interaction risk rather than enabling active modulation of scenario difficulty. Most approaches rely on multi-dimensional or context-specific indicators, which, while informative, lack a unified parameterization that can flexibly and continuously tune difficulty levels across scenarios. This limitation constrains their application in systematic AV testing, where controllable yet interpretable mechanisms for adjusting scenario complexity are of critical importance.

C. DRIVING BEHAVIOR AND AGGRESSIVENESS MODELING

Modeling human driving behavior has long been recognized as a critical component in both traffic simulation and autonomous vehicle (AV) validation. A substantial body of work has sought to categorize driving styles into canonical types such as aggressive, conservative, and normal, which are then used to parameterize simulators or to design prediction and planning modules for AVs. Such categorical distinctions are supported by empirical studies showing that driver style strongly correlates with crash likelihood and risk-taking tendencies [29].

Beyond categorical classifications, researchers have introduced risk-sensitive and psychologically grounded parameters to capture inter-driver heterogeneity. For example, utility-based and game-theoretic models often incorporate risk aversion or loss aversion coefficients to reflect how individuals trade safety against efficiency. Recent studies demonstrate that integrating risk-aware decision rules can better replicate human-like interaction patterns and improve the robustness of AV planning in mixed traffic [30].

Recent studies have further emphasized the importance of modeling human-like behavior and interaction in autonomous driving systems. For instance, Xie et al. [31] propose a data-knowledge-driven trajectory prediction framework that captures complex human driving logic, highlighting the necessity of incorporating interpretable behavioral patterns into autonomous systems. Similarly, Van Gent et al. [32] investigate lane-specific human–machine interaction and demonstrate how behavioral guidance influences driving decisions in lane-related scenarios. These works underline the significance of behavior modeling and interaction awareness in intelligent transportation systems.

Recent advances in cooperative control have further expanded the scope of behavior modeling in mixed-traffic environments. Fu et al. [33] developed a motion-sickness-oriented hierarchical model predictive control (MPC) framework that integrates safety, comfort, and efficiency through multi-objective optimization. This framework exemplifies how parameterized behavioral modeling can be incorporated into control architectures to achieve adaptive interaction regulation, resonating with the goal of systematically modulating driving behavior intensity in autonomous vehicle testing. Recent review work by Du et al. [34] provides a comprehensive synthesis of road safety evaluation methods grounded in driving behavior analysis. The review categorizes existing behavioral modeling approaches according to their descriptive, statistical, and learning-based formulations, and emphasizes that driving traits such as aggressiveness, caution, and risk sensitivity are central to understanding safety-critical interactions. However, it also points out that most behavior-based evaluation frameworks remain limited by their qualitative nature and lack explicit mechanisms for parameterized control or quantitative modulation of these behavioral traits. As a result, it remains difficult to systematically adjust behavioral intensity or aggressiveness in a way that supports reproducible and scalable testing of autonomous vehicles. This observation further motivates our approach, which formalizes aggressiveness as a tunable, continuous parameter within a game-theoretic framework to enable controlled variation of scenario difficulty and interaction intensity.

Aggressiveness, in particular, has emerged as a salient construct for modeling adversarial or high-risk maneuvers in testing environments. Prior work has employed aggressiveness parameters to simulate challenging behaviors such as sudden cut-ins, tailgating, or emergency braking, thereby exposing AVs to more demanding situations [35]. However, most of these implementations rely on heuristic or scenario-specific definitions, which lack generality and systematic control. Typically, aggressiveness is either encoded as a fixed acceleration bias or as a categorical style label, limiting its applicability for fine-grained difficulty modulation in scenario generation.

In summary, while the literature highlights the importance of aggressiveness as both a psychological and operational dimension of driving behavior, existing approaches remain fragmented. Current models tend to either emphasize discrete style categories or rely on loosely defined heuristics, leaving open the challenge of designing a unified and interpretable

parameter that can continuously modulate both action-level tendencies and interaction-level dynamics.

D. APPLICATIONS OF GAME THEORY FOR DRIVING INTERACTION MODELING

Game-theoretic frameworks have been widely adopted in the study of autonomous driving interactions, as they provide a principled way to capture the strategic coupling between autonomous vehicles (AVs) and human-driven vehicles (HVs). Several classical formulations have emerged in the literature. Stackelberg games, for example, are commonly employed to represent hierarchical interactions such as merging or lane-changing, where the leader-follower dynamics reflect asymmetric information and decision timing [36], [37]. Zero-sum and differential games have been applied in adversarial planning, modeling competitive situations such as highway racing, pursuit-evasion, or defensive driving, where one agent's gain corresponds directly to the other's loss [38], [39].

Beyond decision planning, game-theoretic methods have also gained traction in the design of adversarial testing environments. By explicitly modeling an adversarial agent's objective, such as obstructing or deceiving the AV, game theory provides an interpretable and controllable framework for generating safety-critical scenarios. For instance, Ramos et al. [40] demonstrated how Stackelberg and zero-sum games can be extended to capture adversarial maneuvers that challenge safety controllers. Similarly, recent studies highlight the use of multi-agent potential games to emulate complex crowd–vehicle interactions, thereby creating diverse and safety-relevant test cases [41].

Despite these advances, most existing game-theoretic research remains centered on improving the decision-making robustness of AVs, with relatively limited attention to their potential as systematic scenario generation mechanisms. Current applications typically operationalize game equilibria to derive optimal driving strategies, but they seldom exploit the parametric flexibility of game models for generating controllable gradients of scenario difficulty. This limits the integration of game theory into large-scale, automated testing pipelines where both interpretability and adjustable complexity are required.

In summary, prior research has advanced the field of autonomous vehicle testing through diverse approaches to scenario generation, complexity modeling, behavioral representation, and game-theoretic interaction analysis. Rule-based, randomized, and data-driven generation methods respectively emphasize reproducibility, diversity, and realism, yet they lack a unified and controllable mechanism for adjusting scenario difficulty. Complexity modeling efforts, while providing valuable insights through physical indicators and interaction-based metrics, remain largely passive measures rather than active tools for regulating test conditions. Behavioral studies have highlighted aggressiveness as a key dimension of human driving style, but most formulations adopt heuristic or categorical definitions that do not support continuous and parameterized control. Game-theoretic methods, although effective in modeling interactive decision-making, have predominantly been applied to enhance planning robustness rather than to design systematic adversarial test scenarios.

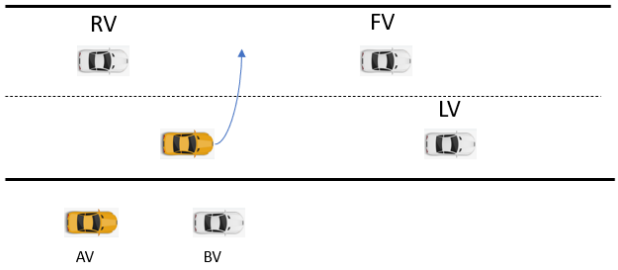

FIGURE 1. Illustration of the lane-changing scenario. The AV attempts to change lanes into the target lane occupied by the RV, which may adopt adversarial strategies to obstruct the maneuver. The LV is positioned ahead of the AV in its original lane, while the FV is located ahead of the RV in the target lane.

These limitations converge on a critical need: an interpretable and unified framework that can actively and continuously regulate scenario difficulty in a principled manner. Addressing this gap, this paper proposes a game-theoretic lane-changing model in which aggressiveness is formalized as a single tunable parameter, α, enabling systematic modulation of adversarial intensity and supporting comprehensive testing and validation of autonomous vehicles.

III. PROBLEM FORMULATION

In the testing and validation of AVs, relying solely on static or low-complexity traffic scenarios is insufficient to reveal their robustness under high-difficulty conditions. To construct more diverse and challenging environments, it is essential to introduce a background vehicle as a critical interactive agent with potential adversarial behaviors. In lane-changing scenarios, the background vehicle (BV) may adopt varying levels of aggressiveness, such as accelerating to block, thereby increasing the difficulty for the AV to successfully complete the maneuver and elevating the risk of collisions or failure. Hence, to systematically characterize AV performance under complex interactive conditions, this study proposes a formalized adversarial lane-changing model, which enables the generation of test cases with controllable difficulty and supports comprehensive performance evaluation.

A. LANE-CHANGING ODD

We consider a lane-changing ODD, illustrated by Figure 1. This ODD consists of a two-lane section of road where the AV is initially driving in one lane. Its goal is change into the second lane which will be referred to as the target lane. There are two BVs in the target lane, referred to in this paper as the rear vehicle (RV), which may exhibit adversarial behaviors to obstruct the lane-changing maneuver. To better approximate diverse and challenging traffic conditions for testing purposes, two additional vehicles are included in the scene: a leading vehicle (LV), positioned ahead of the AV, and a front vehicle (FV), positioned ahead of the RV. Their dynamic

characteristics introduce stochasticity and variability, thereby enhancing the overall complexity of the scenario. The specific scenario representation is illustrated in Figure 1.

The state vector of each vehicle is defined as:

$$s_i = (y_i, x_i, v_i), \quad i \in \{\text{AV}, \text{RV}, \text{LV}, \text{FV}\}, \tag{1}$$

where y_i denotes the longitudinal position, $x_i \in \{0, 1\}$ represents the lane index, and v_i denotes the longitudinal velocity.

To ensure the model is realistic, we limit the initial vehicle placement. Vehicles must maintain a longitudinal gap that is neither too small (avoiding overlap) nor too large (avoiding overly sparse distribution). Vehicle speeds are also restricted to the range $[v_{\min}, v_{\max}]$, consistent with typical traffic behavior.

B. STRATEGY SPACE

In the proposed adversarial lane-changing scenario, both the AV and the RV are modeled as rational decision-making agents. Their interaction is naturally structured as a Stackelberg game. The AV acts first as the leader, committing to a lane-change and acceleration plan. The RV, upon observing this commitment, functions as the follower, strategically choosing its acceleration to maximize its own payoff.

The strategy space of the AV can be expressed as

$$\Pi_{AV} = \{(u^{lat}, u^{lon}) \mid u^{lat} \in \{0, 1\},\ u^{lon} \in [a_{\min}, a_{\max}]\}, \tag{2}$$

where u^{lat} denotes the lateral decision, with $u^{lat} = 0$ representing lane keeping and $u^{lat} = 1$ representing a lane change, while u^{lon} represents the longitudinal control through acceleration selected from the continuous interval $[a_{\min}, a_{\max}]$.

The strategy space of the RV, in contrast, is defined by a discrete set of longitudinal accelerations:

$$\Pi_{RV} = \{a_j \mid a_j \in [a_{\min}, a_{\max}],\ j = 1, 2, \ldots, M\}, \tag{3}$$

where M denotes the number of discretized acceleration options. The choice of a_j not only determines the RV's subsequent longitudinal position and velocity but also critically influences its ability to obstruct the AV's lane-changing maneuver.

C. PAYOFF FUNCTIONS

Within the game-theoretic formulation, the payoff functions quantify the objectives of both agents' success by combining factors like safety, efficiency, and adversarial intent. The AV's payoff measures the trade-off between successfully complete its maneuver quickly while avoiding safety-critical outcomes. Meanwhile the RV's payoff reflects its goal to obstruct the AV, subject to its own safety and dynamic interaction.

The payoff of the AV is defined as

$$U_{AV} = Q_{\text{success}} - P_{\text{collision}} - C_{\text{time}}, \tag{4}$$

where Q_{success} denotes the reward obtained when the lane change is successfully completed, $P_{\text{collision}}$ represents a large negative penalty associated with collisions, and C_{time} is the cost reflecting the time consumed to complete the maneuver. Specifically, it is

$$\begin{aligned} U_{AV} = {} & \mathbb{I}_{\text{success}} \cdot \frac{R_{\text{base}}}{T_{\text{success}} + 1} + \mathbb{I}_{\text{failure}} \cdot (-2R_{\text{base}}) \\ & + \mathbb{I}_{\text{collision}} \cdot P_{\text{collision}} + T \cdot P_{\text{time}}, \end{aligned} \tag{5}$$

where $\mathbb{I}_{\text{success}}$, $\mathbb{I}_{\text{failure}}$, and $\mathbb{I}_{\text{collision}}$ are indicator variables for the event of a successful lane change, a failed attempt, or a collision, respectively. The constant R_{base} denotes the nominal success reward, T_{success} is the number of time steps required to complete the maneuver, and P_{time} is the time-dependent penalty applied at each step T. This formulation penalizes both prolonged maneuvers and unsafe outcomes, while rewarding rapid and safe completion.

The RV's payoff incorporates blocking effectiveness, collision avoidance, and temporal incentives, and is explicitly modulated by the aggressiveness parameter α. It is given by

$$U_{RV} = Q_{\text{block}}(\alpha) - P_{\text{collision}} - C_{\text{time}}, \tag{6}$$

where $Q_{\text{block}}(\alpha)$ represents the obstruction reward, which is explicitly parameterized by the aggressiveness level α. A larger value of α corresponds to a more aggressive strategy by the RV, thereby amplifying the incentive to block the AV's maneuver. Nevertheless, similar to the AV, the RV is subject to a severe penalty $P_{\text{collision}}$ in the case of a crash, ensuring that safety considerations remain an integral part of the decision process. Specifically, it is

$$\begin{aligned} U_{RV} = {} & \mathbb{I}_{\neg\text{success}} \cdot \Bigg(R_{\text{block}} \cdot T \cdot \frac{\alpha}{50} + \lambda_1 \cdot f_d(d_{AV,RV}, v_{AV}) \\ & + \lambda_2 \cdot f_v(\Delta v_{AV,RV}) \Bigg) \\ & - \mathbb{I}_{\text{success}} \cdot \left(2R_{\text{block}} \cdot \left(1 - \tfrac{\alpha}{50}\right)\right) \\ & + \mathbb{I}_{\text{collision}} \cdot P_{\text{collision}} + T \cdot R_{\text{delay}}, \end{aligned} \tag{7}$$

where R_{block} denotes the nominal blocking reward, scaled by both time T and aggressiveness α. Here, the constant value 50 represents the predefined upper bound of the aggressiveness parameter α. The term $\frac{\alpha}{50}$ normalizes α to the range $[0, 1]$, ensuring that its contribution remains comparable to other components in the payoff function and avoiding disproportionate scaling effects. The choice of this upper bound serves as a normalization factor for numerical stability and interpretability, and does not affect the qualitative behavior of the model, as α functions as a monotonic control variable regulating the intensity of adversarial interaction. The term $f_d(d_{AV,RV}, v_{AV})$ measures the proximity of the AV–RV distance $d_{AV,RV}$ to a dynamically defined target obstruction distance $d^*(v_{AV}) = v_{AV} \cdot \text{THW}_{\min} + d_{\text{buffer}}$, encouraging the RV to position itself effectively relative to the AV. Here, $\text{THW}_{\min}$ represents the minimum time headway, which specifies a safe temporal gap between two vehicles as a function of the AV's speed, while d_{buffer} is a fixed safety margin ensuring nonzero separation even at very low velocities. The function $f_v(\Delta v_{AV,RV})$ penalizes large velocity differences $\Delta v_{AV,RV}$, incentivizing the RV to maintain a speed profile compatible

with that of the AV. The weighting coefficients λ_1 and λ_2 control the relative importance of distance regulation and velocity matching in shaping the RV's payoff. When the AV ultimately succeeds in lane changing, the RV incurs a penalty whose magnitude decreases linearly with the aggressiveness parameter α, reflecting that more aggressive blocking strategies reduce the cost of failure. Collisions, however, always trigger a severe fixed penalty $P_{\text{collision}}$. Finally, a positive term $T \cdot R_{\text{delay}}$ is added to reward the RV for delaying the AV's maneuver, independent of success or failure.

This detailed payoff structure highlights the asymmetry of objectives: the AV prioritizes for safety and efficiency, while the RV strategically balances aggressiveness, proximity, and velocity matching in order to obstruct the AV without causing a catastrophic collision. The explicit role of α ensures that the RV's behavioral tendency can be systematically tuned, thus providing a principled mechanism to control the difficulty of the generated scenarios.

D. GAME-THEORETIC SOLUTION

The interaction between the AV and the RV can be abstracted as a two-player dynamic game, where both agents seek to maximize their respective payoffs as defined in Equations (5) and (7). Given a specified aggressiveness parameter α, the equilibrium strategies of the two players can be expressed as

$$\begin{aligned} \pi^*_{AV} &= \arg\max_{\pi_{AV}\in\Pi_{AV}} U_{AV}(\pi_{AV}, \pi_{RV}; \alpha), \\ \pi^*_{RV} &= \arg\max_{\pi_{RV}\in\Pi_{RV}} U_{RV}(\pi_{AV}, \pi_{RV}; \alpha). \end{aligned} \tag{8}$$

where π_{AV} and π_{RV} denote the strategy profiles of the AV and RV, respectively. The existence of such equilibria ensures that the adversarial behaviors of both players under rational decision-making can be consistently captured within the parameterized model.

Formally, under the Stackelberg game formulation, the AV acts as the leader and optimizes its strategy anticipating the best response of the RV [42]:

$$\pi^*_{AV} = \arg\max_{\pi_{AV}\in\Pi_{AV}} U_{AV}\big(\pi_{AV}, \pi^*_{RV}(\pi_{AV})\big), \tag{9}$$

where the RV's best response is defined as [43]

$$\pi^*_{RV}(\pi_{AV}) = \arg\max_{\pi_{RV}\in\Pi_{RV}} U_{RV}(\pi_{AV}, \pi_{RV}). \tag{10}$$

Although the Stackelberg framework effectively models the hierarchical and asymmetric structure of the interaction, where the AV acts first and anticipates the RV's response, this study primarily focuses on analyzing the RV's optimal response, as defined by Equation (10). This emphasis is due to the RV's strategy, which, through the aggressiveness parameter α, directly indicates the dynamic difficulty of the lane-change and provides a systematic way to control the complexity of the AV testing environment.

E. FORWARD-LOOKING EVALUATION IN RV STRATEGY

To operationalize the game-theoretic formulation, the rear vehicle (RV) adopts a forward-looking evaluation mechanism when selecting its action. Rather than greedily reacting to the current state, the RV enumerates a set of candidate accelerations $a_j \in \Pi_{RV}$ and simulates their effect on its next-step state:

$$s_{RV}(t+1; a_j) = \begin{pmatrix} y_{RV}(t) + v_{RV}(t)\Delta t + \frac{1}{2}a_j\Delta t^2, \\ v_{RV}(t) + a_j\Delta t \end{pmatrix}, \tag{11}$$

where y_{RV} and v_{RV} denote the longitudinal position and velocity, respectively.

Given the predicted next-step state, the RV evaluates the corresponding payoff under aggressiveness parameter α:

$$U_{RV}(a_j; \alpha) = f\Big(s_{RV}(t+1; a_j),\, s_{AV}(t),\, \alpha\Big), \tag{12}$$

where $f(\cdot)$ encapsulates obstruction rewards, collision penalties, and temporal incentives as defined in Equation. (7).

The optimal action is then determined by

$$a^*_{RV}(t) = \arg\max_{a_j\in\Pi_{RV}} U_{RV}(a_j; \alpha). \tag{13}$$

This one-step predictive look-ahead ensures that the impact of aggressiveness α is propagated consistently across time, shaping not only instantaneous control decisions but also the temporal evolution of the AV–RV interaction.

IV. DEFINITIONS OF AGGRESSIVENESS

The core innovation of this work is the introduction of the aggressiveness parameter α, which systematically controls the intensity of the rear vehicle (RV)'s behavior. As formulated in Section C, the payoff structure of the RV in Equation (7), which includes a blocking reward term $Q_{\text{block}}(\alpha)$ that explicitly relies on α. By adjusting this term, we can continuously and predictably tune the difficulty of the lane-changing scenario. In this section, we will analyze two distinct but practical definitions for $Q_{\text{block}}(\alpha)$, representing different types of adversarial driving.

A. DEFINITION I: ACCELERATION-FOCUSED AGGRESSIVENESS

The first definition treats α as a direct control over the RV's acceleration tendencies when reacting to the AV. A larger α encourages the RV to choose more extreme accelerations designed to obstruct the AV, while a smaller α promotes smoother, more comfortable driving. This effect is mathematically represented by decomposing the blocking reward $Q_{\text{block}}(\alpha)$ into a convex combination of two distinct components:

$$Q^{(1)}_{\text{block}}(\alpha) = \left(1 - \tfrac{\alpha}{50}\right) \cdot \Phi_{\text{comfort}} + \tfrac{\alpha}{50} \cdot \Phi_{\text{obstruction}}, \tag{14}$$

where Φ_{comfort} applies a penalty for unnecessary accelerations to promote driving smoothness, and $\Phi_{\text{obstruction}}$, which provides a reward for actions that reduce the distance or velocity gap between the AV and RV, effectively hindering the AV's maneuver.

This definition closely matches the concept of aggressiveness as a *driving style parameter*, common in behavioral modeling and risk-sensitive decision-making. It highlights how the intensity of action chosen by the RV directly affects the dynamic interaction between the AV and the RV.

B. DEFINITION II: DISTANCE-FOCUSED AGGRESSIVENESS

The second definition treats α as a spatial threshold that determines when the RV starts its adversarial response, based on the AV's longitudinal position. Instead of changing the magnitude of acceleration, α extends the effective range within which the RV begins to actively obstruct the AV. Specifically, we define this boundary as the *target speed-matching distance* as

$$d_{\text{match}}(\alpha) = d_{\text{base}} + \alpha, \tag{15}$$

where d_{base} is a nominal minimum distance (e.g., 20 m). When the AV–RV distance $d_{AV,RV}$ falls below $d_{\text{match}}(\alpha)$, the RV is incentivized to align its velocity with that of the AV and begin obstruction. The blocking reward is thus defined as

$$\begin{aligned} Q^{(2)}_{\text{block}}(\alpha) = \Phi_{\text{distance}}(d_{AV,RV}, d_{\text{match}}(\alpha)) \\ + \Phi_{\text{velocity}}(\Delta v_{AV,RV}, \alpha), \end{aligned} \tag{16}$$

where Φ_{distance} rewards RV positioning close to a dynamically defined target obstruction distance $d^*(v_{AV}) = v_{AV} \cdot \text{THW}_{\min} + d_{\text{buffer}}$, while Φ_{velocity} penalizes large velocity mismatches, with the penalty weight increasing with α.

This formulation offers an anticipatory view of aggressiveness: a larger α causes the RV to initiate blocking maneuvers from a greater distance. This action, in turn, prolongs the adversarial interaction and systematically increases the scenario difficulty.

C. COMPARISON OF THE TWO DEFINITIONS

The two definitions reveal distinct ways that aggressiveness molds adversarial driving behavior:

- **Acceleration-Focused Aggressiveness (Definition I):** This view treats α as a driver style parameter that directly influences the RV's chosen accelerations. Higher α immediately pushes the RV toward aggressive, decisive actions aimed at blocking the AV.
- **Distance-Focused Aggressiveness (Definition II):** Here, α acts as an anticipatory trigger defined by the spatial distance. It doesn't change how hard the RV accelerates, but how early it starts blocking, making the RV engage from longer distances and thus extending the scenario's difficulty.

Beyond their qualitative descriptions, the two definitions differ fundamentally in how the aggressiveness parameter α modulates the rear vehicle's (RV's) decision-making process and temporal behavior. In the acceleration-focused definition, α directly scales the magnitude of the RV's longitudinal acceleration response within each decision cycle. A higher α value therefore leads to stronger and faster changes in acceleration, allowing the RV to perform more abrupt blocking maneuvers and to exert immediate pressure on the autonomous vehicle (AV). This formulation emphasizes instantaneous intensity, where aggressiveness manifests through the strength and decisiveness of short-term control actions. As a result, scenario difficulty increases mainly due to rapid conflict escalation and reduced reaction margins for the AV.

By contrast, in the distance-focused definition, α no longer alters the instantaneous acceleration but instead modifies the spatial threshold that determines when the RV starts to engage in adversarial behavior. A larger α expands the RV's effective perception or anticipation range, prompting it to begin blocking from a greater distance and maintain the engagement over a longer time horizon. This leads to a more gradual yet persistent interaction process, where the difficulty arises from the extended temporal duration of conflict and the increased need for the AV to plan its response strategically.

Overall, the two definitions highlight different but complementary mechanisms by which aggressiveness influences the scenario. The acceleration-focused formulation captures reactive aggressiveness, representing how forcefully a vehicle responds once an interaction has begun, whereas the distance-focused formulation reflects proactive aggressiveness, describing how early a vehicle chooses to act upon detecting potential competition. Together, these definitions provide a more comprehensive understanding of adversarial behavior modulation and demonstrate the flexibility of the proposed framework in generating test scenarios with controllable difficulty levels.

D. INTEGRATION WITH THE GENERAL PAYOFF FUNCTION

It's important to stress that the two definitions, $Q^{(1)}_{\text{block}}(\alpha)$ and $Q^{(2)}_{\text{block}}(\alpha)$, are not substitutes for the overall RV payoff function in Equation (7). Instead, they are alternative ways to define the internal blocking reward term, $Q_{\text{block}}(\alpha)$, which is part of that function. The total RV payoff always follows Equation (7), which accounts for blocking effectiveness, penalties, collision costs, and time incentives. The difference lies in how $Q_{\text{block}}(\alpha)$ is implemented: Definition I models α as acceleration-focused decisional intensity, while Definition II models it as distance-focused anticipatory engagement.

These definitions are put into practice using a forward-looking evaluation mechanism as described in Section E. The RV uses this mechanism to predict its next state for various acceleration choices before selecting the payoff-maximizing strategy. This predictive look-ahead ensures that the influence of α is dynamically consistent, shaping not just the immediate action but also the temporal evolution of the AV–RV interaction. By systematically adjusting α under $Q^{(1)}_{\text{block}}(\alpha)$ and $Q^{(2)}_{\text{block}}(\alpha)$, this framework can generate a wide spectrum of scenario difficulties, from minor, late interference to aggressive, early obstruction.

V. EXPERIMENTS

To empirically validate the effectiveness of the proposed aggressiveness formulations for shaping the difficulty of lane-changing scenarios, extensive simulation experiments were conducted. The experiments were designed with two primary objectives: first, to demonstrate that aggressiveness α can systematically adjust the adversarial behavior of the rear

vehicle (RV) and consequently, the scenario difficulty; and second, to ensure that the observed difficulty changes are a direct result of the α parameter and not just random variations from the simulation's initialization. In the simulation, the background traffic consists of three types of vehicles: the rear vehicle (RV), the front vehicle (FV), and the lead vehicle (LV). Among them, only the RV is modeled as an interactive agent using the proposed game-theoretic framework. The RV determines its action by maximizing its payoff function under safety constraints, making it responsive to the AV's state and lane-changing intention. In contrast, the FV and LV follow simplified kinematic rules with bounded accelerations and do not explicitly optimize any objective function. They serve as passive traffic participants to provide realistic driving context rather than adversarial interaction. In the original scenario generator, randomness is introduced to enhance scenario diversity and to better capture the variability of real-world traffic conditions. In contrast, for the two fixed experimental settings designed for the two aggressiveness definitions, the initial conditions are kept constant, thereby removing the effect of stochastic initialization from the reported evaluation. Since all experimental results presented in this paper are obtained from these fixed settings, the resulting variations in performance metrics can be directly interpreted as the effect of the aggressiveness parameter α, which supports a controlled and meaningful assessment of its influence on scenario difficulty.

A. EXPERIMENTAL DESIGN

The original scenario generator relied on stochastic sampling for initial vehicle states (like position and velocity). While this created diverse scenarios, the resulting run-to-run variability made it hard to prove that differences in outcomes were solely due to the aggressiveness parameter α. To solve this problem, we implemented two fixed-initialization experimental settings corresponding to the acceleration-focused and distance-focused definitions of aggressiveness.

In both of these deterministic settings, the initial states were fixed as follows: the autonomous vehicle (AV) was initialized at a longitudinal position of 50 m with a velocity of 30 km/h, while the RV was placed 10 m behind in the target lane, also traveling at the same 30 km/h. A leading vehicle (LV) and a front vehicle (FV) were positioned 20 m ahead of the AV and RV, respectively, with velocities set to 25 km/h and 35 km/h. In the fixed-initialization setting, the chosen initial states represent typical highway lane-changing conditions and ensure controlled yet realistic interactions. The AV's initial position of 50 m and speed of 30 km/h emulate a normal cruising state before initiating a lane change, while placing the RV 10 m behind the AV in the target lane keeps it within an effective range to exhibit obstructive behavior. Likewise, the LV and FV were configured within suitable distance and speed ranges to maintain plausible car-following dynamics and ensure safe, reproducible experimental conditions. This deterministic configuration ensured that every simulation started identically, effectively isolating the role of the aggressiveness parameter α for clear validation.

B. EVALUATION METRICS

To quantify the influence of aggressiveness, we employed three key metrics that together provide a holistic assessment of agent performance and scenario difficulty: (i) the expected payoff of the RV, reflecting the effectiveness of obstruction strategies; (ii) the success rate of the AV's lane change, measuring the difficulty of the scenario; and (iii) the average time required for the AV to complete a successful lane change, which assess the temporal efficiency of the successful maneuvers.

C. SIMULATION PROCEDURE

For each aggressiveness definition, we systematically tested 11 difficulty levels, corresponding to aggressiveness α values ranging from 0 to 50 in increments of 5. At every level, 100 independent simulations were conducted to estimate the expected values of the evaluation metrics. In all simulations, the RV used the payoff functions and decision strategies described in Section III, specifically selecting acceleration to maximize its expected return while adhering to collision avoidance constraints.

The experimental pipeline's key strength is that it systematically varies only the aggressiveness parameter α while keeping all other variables constant. This design significantly enhances reproducibility and guarantees that any trends observed in the metrics (whether increasing or decreasing) are directly attributable to changes in α.

D. EXPECTED OUTCOMES

The two aggressiveness definitions predict distinct adversarial mechanisms: Under the acceleration-focused aggressiveness, higher α values are expected to bias the RV toward stronger, more forceful accelerations to block the AV. In turn, reduce the AV's success rate, increase the RV's obstruction payoff, and lengthen the time needed for a successful lane change. Under the distance-focused aggressiveness, higher α values here expand the spatial zone where the RV initiates obstruction, leading to earlier and more sustained adversarial engagement. This anticipatory behavior is anticipated to lower AV success rates and increase delays, even without the RV using extreme acceleration.

By contrasting the results from these two definitions, the experiments will reveal not just the quantitative effect of α on scenario difficulty, but also the qualitative mechanisms by which aggressiveness works, whether through decisional intensity (strong acceleration) or spatial anticipation (early engagement).

VI. RESULTS AND ANALYSIS

We present the experimental findings obtained using the fixed-initialization settings for both the acceleration-focused and distance-focused aggressiveness definitions. The core of the analysis is to demonstrate the systematic influence of the aggressiveness parameter α on the expected payoff of the rear vehicle (RV), the autonomous vehicle (AV)'s lane-change success rate, and the time required for successful maneuvers.

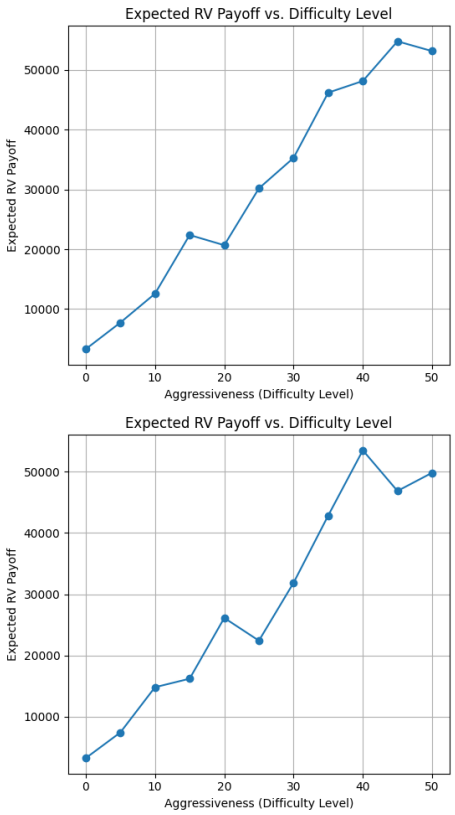

FIGURE 2. Expected RV payoff versus aggressiveness under acceleration-focused (top) and distance-focused (bottom) definitions.

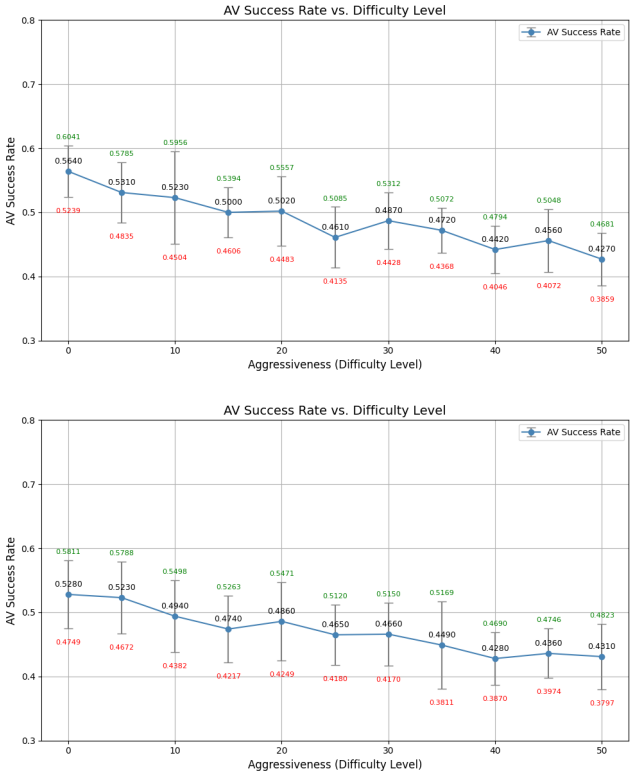

FIGURE 3. AV lane-change success rate versus aggressiveness under the acceleration-focused definition (top) and the distance-focused definition (bottom).

A. RV EXPECTED PAYOFF

As shown in Figure 2, the expected payoff of the RV consistently increases as the aggressiveness parameter α rises, a trend observed for both the acceleration-focused and distance-focused definitions. This monotonic growth verifies that increasing α effectively enhances the RV's ability to obstruct the AV. Because this result holds whether aggressiveness is implemented via decisional intensity or anticipatory spatial engagement, we confirm that α is a reliable and controllable factor for shaping adversarial difficulty.

B. AV SUCCESS RATE

The results in Figure 3 indicate a clear inverse relationship: the AV's lane-change success rate declines with rising aggressiveness α. The success rate is high when α is low, but it diminishes significantly as α increases. This robust trend clearly demonstrates α's function as a difficulty regulator: by adopting more aggressive behaviors, either by accelerating decisively or engaging earlier, the RV systematically and successfully lowers the AV's chance of completing the maneuver.

C. AV LANE-CHANGE DURATION

Figure 4 depicts the average time required for the AV to complete successful lane changes. The data reveals an upward trend with increasing aggressiveness, meaning that even when the AV succeeds, the maneuver takes significantly longer under more aggressive adversarial conditions. This increase reflects both the RV's effectiveness in delaying the maneuver and the AV's increased caution when encountering strong opposition.

D. FAILURE CASE ANALYSIS

To provide a deeper understanding of unsuccessful lane-changing attempts, we further analyze the failure cases observed in the experiments.

In our framework, failures are not caused by collisions, as the autonomous vehicle (AV) follows a risk-averse decision-making strategy that avoids unsafe maneuvers. Instead, failures primarily occur when the AV determines that the surrounding traffic conditions are not suitable for a safe lane change and consequently postpones the maneuver until the simulation reaches the maximum time limit.

Based on this behavior, we categorize the failure cases as follows:

- **Conservative abort (timeout failure):** The AV refrains from initiating a lane change due to safety concerns and eventually fails when the time horizon is exceeded.

This result indicates that increasing the aggressiveness parameter α does not lead to unsafe interactions (e.g., collisions), but rather increases the difficulty of the scenario by forcing the AV to adopt more conservative strategies. Therefore, the observed failures reflect the AV's safety-aware decision-making under challenging conditions, rather than unsafe behavior.

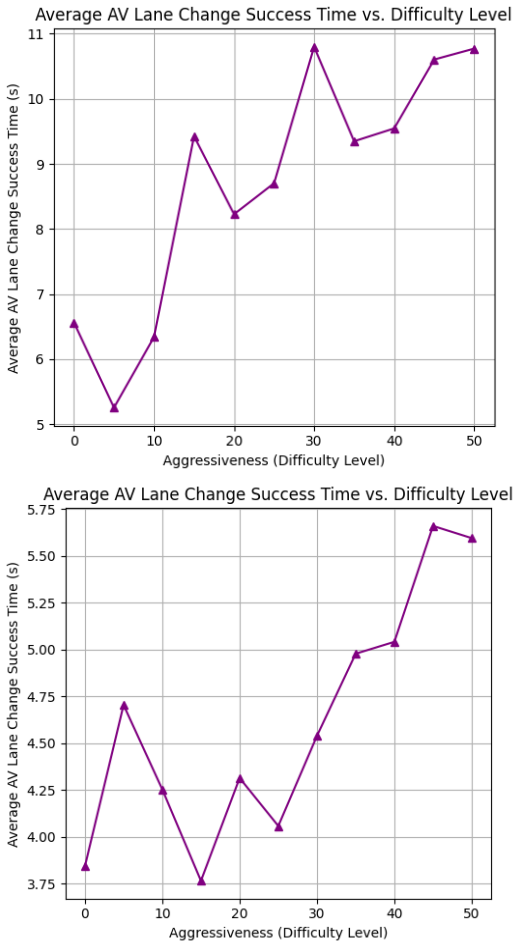

FIGURE 4. Average AV lane-change success time versus aggressiveness under acceleration-focused (top) and distance-focused (bottom) definitions.

E. COMPARATIVE ANALYSIS OF THE TWO DEFINITIONS

While both aggressiveness definitions produced the same general outcomes, increased RV payoff, reduced AV success, and longer maneuver time. However, their underlying adversarial mechanisms differed. The acceleration-focused definition results in stronger RV acceleration, resulting in sharper changes in maneuver efficiency. The distance-focused definition manifested earlier in the interaction, creating a more sustained adversarial influence that primarily affected the temporal dynamics of the AV's maneuver.

Overall, these findings confirm that aggressiveness parameter α is a robust and controllable factor for regulating scenario difficulty. The consistency of overall trends validates the generality of the framework, while the observed differences in manifestation emphasize the importance of modeling adversarial behavior using both decisional intensity (acceleration) and spatial anticipation (distance).

VII. DISCUSSION

The experimental results presented in Section VI provide strong evidence that the aggressiveness parameter α is a principled and effective mechanism for regulating the difficulty of adversarial lane-changing scenarios. Beyond the quantitative performance data, these findings prompt a deeper discussion regarding the interpretability, generalizability, and practical implications of using an aggressiveness-based approach for scenario design.

A. SIGNIFICANCE OF A UNIFIED DIFFICULTY PARAMETER

A key contribution of this work is demonstrating that the single parameter, α, can consistently modulate both the adversarial incentives and AV outcomes, even when applied to distinct behavioral definitions. This approach is superior to previous method that often relied on ad hoc or complex multi-dimensional factors, which tend to compromise both interpretability and reproducibility. The monotonic relationships found in the RV payoffs, AV success rates, and maneuver durations confirm α as a reliable difficulty regulator, enabling controlled, systematic stress-testing of AV decision-making.

B. INTERPRETABILITY OF AGGRESSIVENESS DEFINITIONS

The comparision between acceleration-focused and distance-focused formulations highlights two complementary aspects of aggressiveness. The acceleration-focused approach stresses decisional intensity, mirroring typical concepts of driving style and risk. In contrast, the distance-focused approach embodies anticipatory engagement, showing how early and sustained adversarial responses control the interaction dynamics. The fact that the outcome trends were consistent across both definitions suggests that aggressiveness is a generalizable construct. However, the qualitative differences underscore the necessity of explicitly modeling both instantaneous (acceleration) and anticipatory (distance) behavioral dimensions in adversarial traffic scenarios.

C. IMPLICATIONS FOR AV TESTING AND VALIDATION

From a practical standpoint, the ability to controllably modulate scenario difficulty directly benefits the systematic evaluation of AVs. By simply adjusting the parameter α, developers and regulators can subject AV systems to increasingly challenging conditions, ensuring robustness against not only nominal situations but also under strategically adversarial circumstances. This method fills a crucial gap between purely random scenario generation and custom-made stress cases, offering a scalable and reproducible pathway for comprehensive testing. Furthermore, separating aggressiveness into its decisional and spatial components allows for targeted stress tests, marking it possible to isolate and fix specific weakness in AV planning or prediction modules.

VIII. CONCLUSION

This paper introduced a game-theoretic framework that overcomes the limitations of current AV testing methods by generating adversarial lane-changing scenarios with controllable difficulty. By introducing a single adjustable parameter, α, the model systematically regulates aggressiveness at both

the action level and the interaction level. Two complementary definitions of aggressiveness were formalized, offering distinct yet consistent mechanisms for shaping adversarial behavior.

Simulation experiments validated that increasing α consistently enhanced the rear vehicle's payoff, reduces the AV's success rate, and extended maneuver duration. By formalizing and testing two complementary definitions of aggressiveness, we not only confirmed the framework's effectiveness but also highlighted the necessity of modeling both decisional intensity and anticipatory engagement in adversarial traffic.

The resulting framework offers a unified, interpretable, and reproducible mechanism for scenario-based AV testing, filling the critical gap between stochastic and handcrafted scenarios. It enables progressive stress evaluation and sets a principled foundation for scalable and systematic safety validation under strategic adversarial conditions. Future research will build upon this foundation by incorporating multi-lane, multi-agent, and lateral dynamics.

REFERENCES

[1] P. Koopman and M. Wagner, "Challenges in autonomous vehicle testing and validation," *SAE Int. J. Transp. Saf.*, vol. 4, no. 1, pp. 15–24, Apr. 2016.

[2] Z. Wang, J. Ma, and E. M.-K. Lai, "A survey of scenario generation for automated vehicle testing and validation," *Future Internet*, vol. 16, no. 12, p. 480, Dec. 2024.

[3] Z. Ghodsi et al., "Generating and characterizing scenarios for safety testing of autonomous vehicles," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Nagoya, Japan, Jul. 2021, pp. 157–164.

[4] R. Krajewski, J. Bock, L. Kloeker, and L. Eckstein, "The highD dataset: A drone dataset of naturalistic vehicle trajectories on German highways for validation of highly automated driving systems," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 2118–2125.

[5] W. Zhan et al., "INTERACTION dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps," 2019, *arXiv:1910.03088*.

[6] F. M. Heuer, "Scenario generation for testing of automated driving functions based on real data," Ph.D. dissertation, Inst. Automotive Eng., Braunschweig, Germany, 2022, p. 200.

[7] H. Qi, "Real world observations, maneuver estimation and behavioral predictability," in *Stochastic Two-Dimensional Microscopic Traffic Model, Theory Appl.*, 2024, pp. 27–61.

[8] J. Zhou, L. Wang, and X. Wang, "Online adaptive generation of critical boundary scenarios for evaluation of autonomous vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 6, pp. 6372–6388, Jun. 2023.

[9] S. C. Schnelle et al., "Review of simulation frameworks and standards related to driving scenarios," Nat. Highway Traffic Saf. Admin. (NHTSA), Washington, DC, USA, Tech. Rep. DOT HS 812 737, 2019.

[10] F. Khan, M. Falco, H. Anwar, and D. Pfahl, "Safety testing of automated driving systems: A literature review," *IEEE Access*, vol. 11, pp. 120049–120072, 2023.

[11] W. Ding, C. Xu, M. Arief, H. Lin, B. Li, and D. Zhao, "A survey on safety-critical driving scenario generation—A methodological perspective," *IEEE Trans. Intell. Transp. Syst.*, pp. 1–18, Jul. 2023.

[12] Q. Song, "Critical scenario identification for testing of realistic autonomous driving systems," Ph.D. dissertation, Dept. Comput. Sci., Lund Univ., Lund, Sweden, 2024.

[13] D. Krajzewicz et al., "Recent development and applications of SUMO—Simulation of urban mobility," *Int. J. Adv. Syst. Meas.*, vol. 5, no. 3, pp. 128–138, 2012.

[14] M. Fellendorf and P. Vortisch, "Microscopic traffic flow simulator VISSIM," in *Fundamentals of Traffic Simulation*. New York, NY, USA: Springer, 2010, pp. 63–93.

[15] S. Fang, P. Hang, C. Wei, Y. Xing, and J. Sun, "Cooperative driving of connected autonomous vehicles in heterogeneous mixed traffic: A game theoretic approach," *IEEE Trans. Intell. Vehicles*, early access, May 13, 2024, doi: 10.1109/TIV.2024.3399694.

[16] P. Hang, C. Lv, Y. Xing, C. Huang, and Z. Hu, "Human-like decision making for autonomous driving: A noncooperative game theoretic approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 4, pp. 2076–2087, Apr. 2021.

[17] C. Neurohr, L. Westhofen, T. Henning, T. de Graaff, E. Möhlmann, and E. Böde, "Fundamental considerations around scenario-based testing for automated driving," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Las Vegas, NV, USA, Oct. 2020, pp. 121–127.

[18] B. Toth and Z. Szalay, "Minimum critical test scenario set selection for autonomous vehicles prior to first deployment and public road testing," *Appl. Sci.*, vol. 15, no. 13, p. 7031, Jun. 2025, doi: 10.3390/app15137031.

[19] B. Kim and E. Kang, "Toward large-scale test for certifying autonomous driving software in collaborative virtual environment," *IEEE Access*, vol. 11, pp. 72641–72654, 2023.

[20] M. Bäumler, F. Linke, and G. Prokop, "Categorizing data-driven methods for test scenario generation to assess automated driving systems," *IEEE Access*, vol. 12, pp. 52030–52050, 2024.

[21] J. Cai, W. Deng, H. Guang, Y. Wang, J. Li, and J. Ding, "A survey on data-driven scenario generation for automated vehicle testing," *Machines*, vol. 10, no. 11, p. 1101, Nov. 2022.

[22] X. Zhang et al., "Finding critical scenarios for automated driving systems: A systematic mapping study," *IEEE Trans. Softw. Eng.*, vol. 49, no. 3, pp. 991–1026, Mar. 2023.

[23] H. X. Liu et al., "Behavioral safety assessment towards large-scale deployment of autonomous vehicles," 2025, *arXiv:2505.16214*.

[24] P. Goudarzi and B. Hassanzadeh, "Collision risk in autonomous vehicles: Classification, challenges, and open research areas," *Vehicles*, vol. 6, no. 1, pp. 157–190, Jan. 2024.

[25] H. U. Ahmed, S. Ahmad, X. Yang, P. Lu, and Y. Huang, "Safety and mobility evaluation of cumulative-anticipative car-following model for connected autonomous vehicles," *Smart Cities*, vol. 7, no. 1, pp. 518–540, Feb. 2024.

[26] Z. Cheng, J. Zhu, Z. Feng, M. Yang, W. Zhang, and J. Chen, "Driving safety risk analysis and assessment in a mixed driving environment of connected and non-connected vehicles: A systematic survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 26, no. 5, pp. 5747–5781, May 2025.

[27] J. Jiang et al., "Interactive risk (IR): An omnidirectional safety metric of CAVs based on multimodal trajectory prediction and driving risk field," *Accident Anal. Prevention*, vol. 222, Nov. 2025, Art. no. 108228.

[28] H. Iqbal, H. Sadia, A. Al-Kaff, and F. Garciéa, "Novelty detection in autonomous driving: A generative multi-modal sensor fusion approach," *IEEE Open J. Intell. Transp. Syst.*, vol. 6, pp. 799–812, 2025.

[29] T. Zhang, A. H. S. Chan, H. Xue, X. Y. Zhang, and D. Tao, "Driving anger, aberrant driving behaviors, and road crash risk: Testing of a mediated model," *Int. J. Environ. Res. Public Health*, vol. 16, no. 3, p. 297, Jan. 2019.

[30] R. Chandra, M. Wang, M. Schwager, and D. Manocha, "Game-theoretic planning for autonomous driving among risk-aware human drivers," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, Philadelphia, PA, USA, May 2022, pp. 2876–2883.

[31] J. Xie et al., "Incomprehensible but intelligible human logics: Toward a data-knowledge-driven trajectory prediction model," *IEEE Open J. Intell. Transp. Syst.*, vol. 7, pp. 412–433, Dec. 2025.

[32] P. V. Gent, H. Farah, N. Van Nes, and B. Van Arem, "The persuasive automobile: Design and evaluation of a persuasive lane-specific advice human machine interface," *IEEE Open J. Intell. Transp. Syst.*, vol. 1, pp. 93–108, 2020.

[33] Z. Fu, B. Chai, D. Zhao, B. Ma, S. Rakheja, and J. Hu, "Motion sickness-oriented cooperative control in mixed traffic: A hierarchical MPC framework with multi-objective optimization," *IEEE Open J. Intell. Transp. Syst.*, vol. 6, pp. 1133–1142, 2025.

[34] Z. Du, M. Deng, N. Lyu, and Y. Wang, "A review of road safety evaluation methods based on driving behavior," *J. Traffic Transp. Eng.*, vol. 10, no. 5, pp. 743–761, Oct. 2023.

[35] W. Hu et al., "Formulating vehicle aggressiveness towards social cognitive autonomous driving," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 3, pp. 2097–2109, Mar. 2023.

[36] H. Niu, Q. Chen, Y. Li, Y. Zhang, and J. Hu, "Stackelberg driver model for continual policy improvement in scenario-based closed-loop autonomous driving," 2023, *arXiv:2309.14235*.

[37] Q. Chen, D. Zhao, C. Liu, M. Yang, and Y. Shi, "Autonomous vehicles in mixed-autonomy traffic: Game theoretic human-like decision making countermeasures," *Complex Eng. Syst.*, vol. 4, no. 4, Dec. 2024, Art. no. 100085.

[38] Y. Cui, J. Tang, Q. Luo, Z. Feng, and T. Huang, "A Nash-stackelberg game theoretic planner for many-to-few multi-vehicle racing," *IEEE Trans. Intell. Vehicles*, early access, Oct. 11, 2025, doi: 10.1109/TIV.2024.3478391.
[39] Z. Sun, Y. Liu, J. Wang, C. Anil, and D. Cao, "Game theoretic approaches in vehicular networks: A survey," 2020, *arXiv:2006.00992*.
[40] M. A. Ramos, M. C. Moura, I. D. Lins, and F. S. Ramos, "The use of game theory for autonomous systems safety: An overview," in *Proc. 31st Eur. Saf. Rel. Conf. (ESREL)*, Angers, France, 2021, pp. 2494–2501.
[41] Z. Lin and Z. Tian, *Scenario-based decision-making using game theory for interactive autonomous driving: A survey*,arXiv preprint arXiv:2509.05777, Sep. 2025.
[42] T. Başar and G. J. Olsder, *Dynamic Noncooperative Game Theory*, 2nd ed., Philadelphia, PA, USA: SIAM, 1998.
[43] D. Reeves and M. P. Wellman, "Computing best-response strategies in infinite games of incomplete information," 2012, *arXiv:1207.4171*.

ZIYU WANG received the master's degree in computer and information sciences from The University of Sydney, Sydney, Australia, in 2022. She is currently pursuing the Ph.D. degree with Auckland University of Technology, Auckland, New Zealand. Her research interests include the design and validation of algorithms for autonomous vehicles.

JING MA (Member, IEEE) received the Ph.D. degree in computer sciences from Auckland University of Technology, Auckland, New Zealand, in 2019. She is currently a Senior Lecturer with the Department of Data Science and AI, Auckland University of Technology. Her research interests include intelligent transportation systems, artificial intelligence for autonomous vehicles, and robotics.

EDMUND M.-K. LAI (Life Senior Member, IEEE) received the B.E. (Hons.) and Ph.D. degrees in electrical engineering from The University of Western Australia, Australia, in 1982 and 1991, respectively. He is currently a Professor of information engineering with Auckland University of Technology, New Zealand. He has more than 30 years of academic experience, having previously held faculty positions at universities, Australia, Hong Kong, and Singapore. He has published more than 150 international refereed journal and conference papers in signal processing, intelligent control, computational intelligence, and artificial neural networks. He is a fellow of the Institution of Engineering and Technology (IET) and the Engineers Australia (IEAust).