Convergent Harmonious Reinforcement Learning—Lane Changing in a More Traffic Friendly Way

Ruolin Yang, Zhuoren Li, *Graduate Student Member, IEEE*, Bo Leng, Lu Xiong, and Xin Xia

***Abstract*—Lane-change is a common maneuver. However, completely egocentric lane-changing behavior may increase potential driving risks and cause oscillations in the movement of surrounding vehicles, referred to as disharmony. To improve the overall safety and efficiency of both ego vehicle and their surroundings, this paper proposes a convergent harmonious reinforcement learning (CHRL) approach to generate harmonious lane-changing strategies. It introduces a game-based model to measure the overall harmony cost. On this basis, a prosocial critic network is established to guide the policy toward harmony by decreasing the harmony cost. Meanwhile, CHRL identifies and penalizes discordant behaviors that may lead to high risk, accelerating the RL agent's learning of harmonious driving strategies through expert demonstrations of the game model. Simulation and real-data tests validate that CHRL, compared to other lane-change methods and human drivers, improves the overall harmony of lane changes for autonomous vehicles.**

***Index Terms*—Autonomous vehicles, reinforcement learning, lane change, harmony, game theory.**

I. Introduction

LANE change is a prevalent and primary driving maneuver, but accidents occurring during lane changes account for over 10% of all traffic accidents [1], [2]. Additionally, poorly executed lane changes can induce traffic congestion [2], [3] that further catalyzes accident risk [4], [5]. A recent study attributes over 90% of these incidents to human factors [6].

Autonomous vehicles (AVs) have tremendous potential to reduce traffic accidents and improve traffic efficiency [7]. However, they face complex lane-change scenarios involving human drivers whose maneuvers cannot be accurately modeled. In such situations, conventional rule-based AV algorithms often lack the adaptability required for flexible lane changes.

Received 24 July 2024; revised 24 December 2024 and 15 June 2025; accepted 10 August 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 52325212 and Grant 52372394, in part by the Fundamental Research Funds for Central Universities under Grant 22120230311, and in part by the State Key Laboratory of Intelligent Vehicle Safety Technology Open Fund Project under Grant IVSTSKL-202427. The Associate Editor for this article was B. F. Ciuffo. *(Corresponding author: Bo Leng.)*

Ruolin Yang, Zhuoren Li, Bo Leng, and Lu Xiong are with the School of Automobile Studies, Tongji University, Shanghai 201804, China (e-mail: yrljstxws@163.com; 1911055@tongji.edu.cn; leng_bo@tongji.edu.cn; xiong_lu@tongji.edu.cn).

Xin Xia is with the Department of Mechanical Engineering, University of Michigan–Dearborn, Dearborn, MI 48128 USA (e-mail: xinxia@umich.edu).

Digital Object Identifier 10.1109/TITS.2025.3598822

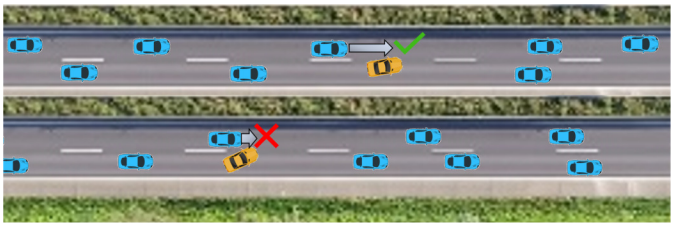

Fig. 1. Disharmonious lane change brings the risk of collision.

To enhance AV adaptability in complex scenarios, reinforcement learning (RL) methods have emerged with demonstrated capability to execute lane changes across dynamic environments [8], [9], [10].

However, existing RL's reward-oriented structure can generate "disharmonious" maneuvers, where vehicles prioritize individual rewards over collective traffic flow stability. Such maneuvers manifest as erratic lane changes that disrupt surrounding vehicles within the lane change window (LCW), as depicted in Fig. 1. While potentially avoiding immediate collisions (satisfying basic safety), these maneuvers may induce higher speed deviation and volatility across the LCW, creating unstable traffic conditions. Research [11] confirms that traffic flow speed uniformity directly correlates with overall safety, with less uniform flows exhibiting heightened accident risk [4], [12], [13].

In this paper, we define "harmony" in lane changing as the collective stability of traffic flow, characterized by gradual acceleration and deceleration patterns that maintain consistent vehicular speed profiles across all LCW participants [11]. This harmony differs fundamentally from basic safety (collision avoidance): safety concerns immediate crash prevention, while harmony addresses broader traffic stability and long-term collective safety. A technically "safe" maneuver can still disrupt flow harmony, creating ripple effects that compromise system-wide efficiency and elevate collective risk.

Current approaches to lane-changing face three fundamental limitations in addressing harmony optimization. First, safety-oriented methods [14], [15], [16], [17] focus primarily on collision avoidance without adequately considering the broader impact on traffic flow stability and coordination. Second, while Game Theory (GT) approaches [18], [19], [20], [21], [22] attempt to model vehicle interactions, they either impose rigid

interaction structures or become computationally intractable in realistic scenarios, failing to capture the dynamic, adaptive nature of harmonious traffic flow. Third, existing RL frameworks [23], [24] typically employ single-critic architectures that combine conflicting objectives into unified reward signals, creating reward confusion that fundamentally limits their capacity to develop truly harmonious maneuvers.

These limitations reveal a critical research gap: despite harmony being recognized as crucial for traffic stability, current methodologies lack the architectural design, interaction modeling, and evaluation mechanisms required to effectively promote and measure harmonious lane-changing maneuvers. Existing approaches treat harmony, if considered at all, as a secondary constraint rather than a primary optimization target, limiting their effectiveness in developing lane change strategies that actively contribute to traffic flow stability.

Our work introduces Convergent Harmony Reinforcement Learning (CHRL) to address these limitations. The key contributions of our approach include:

(1) **Dual Critics Architecture in CHRL:** We present dual critics comprising an EV-related critic (E-Critic) and a Prosocial critic (P-Critic). This design separates harmony assessment from performance evaluation, preventing reward confusion inherent in conventional single-critic systems and elevating traffic harmony from a secondary constraint to a primary optimization objective, enabling convergence toward harmonious lane-change maneuvers.

(2) **Policy Generation with Harmony Guidance in CHRL:** We develop a computationally efficient Incomplete Information Static Game (IISG) model for realistic traffic interaction modeling. This innovation serves two crucial functions: (1) generating expert demonstrations that accelerate policy learning toward harmonious maneuvers; and (2) providing reference solutions for real-time harmony-preserving action correction. Our approach captures the essence of traffic interactions while remaining practical for real-time applications, enabling policies to balance individual objectives with collective traffic stability.

(3) **Systematic Harmony Evaluation:** We establish quantitative metrics for lane-changing harmony assessment and conduct extensive comparative testing across multiple benchmarks, including human drivers, rule-based models, and existing RL approaches. Our experimental validation in both simulated and real-world traffic scenarios demonstrates CHRL's effectiveness in enhancing traffic harmony while maintaining operational efficiency.

The rest of the article is organized as follows: Section II analyzes related work in lane-change harmony. Section III describes the CHRL framework construction. Sections IV-V elaborate on the dual-critics implementation and harmony-guided policy generation. Section VI analyzes experimental results, and Section VII summarizes the harmony improvements achieved through CHRL.

II. Related Work

In autonomous lane-changing research, addressing the harmony issue between the ego vehicle (EV) and surrounding vehicles (SVs) remains a significant challenge. This section reviews existing approaches across three critical dimensions: safety-oriented constraints, interactive modeling, and architectural design limitations.

Researchers have primarily approached lane-change harmony through kinematic safety constraints. Recent works [14], [15] implemented offline safety metrics such as Time-to-Collision (TTC) and Time Difference to Merging (TDTM) to restrict RL training within safety boundaries. Zhang et al. [16] introduced an Implicit Safe Set Algorithm (ISSA) that applies post-decision constraints through hierarchical control–a structure widely adopted for ensuring operational safety [17]. Additionally, several studies incorporated safety constraints directly into critic networks: Bharadhwaj et al. [25] developed action resampling mechanisms when safety thresholds were exceeded, Wen et al. [26] constructed safety networks that conditioned policy updates on both gradient descent and safety criteria, while Ma et al. [27] redesigned critic loss functions to promote safer kinematic states. However, while these approaches enhance safety, they insufficiently address the complex interactive dynamics between vehicles that are fundamental to achieving truly harmonious lane changes.

Recognizing the limitations of purely safety-focused methods, researchers have incorporated interactive awareness through GT models. Multiple studies [18], [19], [20] applied Stackelberg game frameworks to establish leader-follower relationships between EVs and SVs during lane changes. Despite demonstrating effectiveness in controlled environments, these approaches face significant constraints in complex traffic scenarios where leader-follower distinctions become ambiguous and computational demands for multi-vehicle game modeling grow prohibitively expensive. Yang et al. [21] utilized GT as a constraint mechanism to guide decision network convergence, yet omitted game cost factors during evaluation, potentially restricting the exploration of optimal interactive strategies. Similarly, Karimi et al. [22] developed a hybrid approach that initializes decisions with GT before refining through RL iterations, but without guaranteeing strategy convergence or accurately reflecting real-world traffic behavior. Our CHRL framework addresses these limitations by integrating an Incomplete Information Stochastic Game (IISG) model that balances decision accuracy with computational efficiency, enabling more realistic harmony-oriented interactions.

A critical yet underexplored limitation in existing RL frameworks concerns architectural design for multi-objective optimization. Conventional single-critic RL frameworks integrate diverse reward components–including safety, efficiency, and social factors–into unified evaluation modules [23]. This architectural choice frequently results in reward inflation and constrained policy exploration [24], as conflicting objectives compete within the same evaluation space. Recent studies in autonomous driving RL [13], [28] demonstrate that such reward confusion particularly impacts harmony-related

objectives, which require nuanced evaluation separate from performance-oriented rewards. The fundamental issue stems from treating harmony as a secondary constraint rather than a primary optimization target, limiting the agent's ability to develop truly cooperative behaviors. The proposed CHRL framework addresses this architectural limitation through a novel dual-critic design that fundamentally transforms how interaction harmony is evaluated and optimized.

Unlike previous approaches that treat harmony as a secondary consideration or safety constraint, our CHRL framework elevates harmony to a first-order design principle through three key innovations: (1) explicit modeling of vehicle interactions using IISG that captures realistic incomplete information scenarios in traffic; (2) architectural separation of harmony evaluation through the dedicated P-Critic, preventing reward confusion and enabling focused harmony optimization; and (3) integrated perception-action mechanisms that adaptively balance individual operational objectives with collective traffic harmony. These innovations collectively enable the development of lane-changing policies that not only satisfy safety requirements but actively contribute to smoother, more cooperative traffic flow–addressing the fundamental limitations identified in existing safety-constrained, game-theoretic, and single-critic approaches.

III. Convergent Harmonious RL in Lane-Change Decision

A. Problem Formulation and Key Definitions

The problem is defined as a lane-change decision-making problem in the highway scenario, with the following key definitions:

(1) Lane change window (LCW): the range of the preceding 80 meters and the following 20 meters of the EV;
(2) Surrounding vehicles (SVs): the vehicles that are directly affected by the EV, including the preceding and following vehicles in the EV's current and target lanes within LCW.

In the described target problem, the EV operates on the highway with a decision frequency of once per second and a trajectory planning frequency of 15 times per second.

B. Construction of CHRL Lane-Change System

We model CHRL for lane-change maneuver based on the Markov Decision Process (MDP) [29]. In addition to the main modeling components of State space $\mathcal{S}$, Action space $\mathcal{A}$, Reward space $\mathcal{R}$, and discount factor (γ) for reward updates. Considering that lane changing only lasts for a while, the MDP model adds the decision horizon factor (h) and only considers the period during the lane change of the EV. So, the model can be represented as $\langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \gamma, h \rangle$.

Fig. 2 shows the process of CHRL to make harmonious lane change maneuvers. After observing the current environment, the actor network generates RL Action which is monitored by the harmony assessment module. If any RL Action is assessed as a danger, the harmony guidance module is activated to guide it to a Game Action with more harmony. During training, the dual critics are used to update the lane-change policy of the actor network based on EV-related reward and harmony-related evaluation simultaneously.

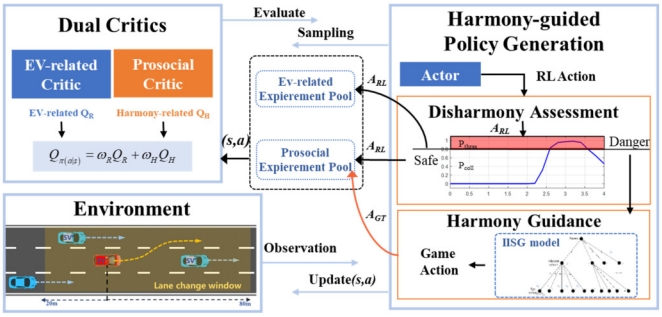

Fig. 2. The framework of convergent harmony reinforcement learning for lane-change trajectory decision. orange parts belong to the convergent harmony mechanism.

We designed key factors including the internal RL network structure, observation space, action space, and reward mechanism.

1) RL Network Structure: The architecture of the CHRL is depicted in Fig. 2, comprising one actor network and two critic networks. The Actor network's input layer processes the flattened observation space vector and has an output layer dimensioned to match the action space. The dual critics, mirroring each other's structure, accept action-state pairs from their respective experience pools as inputs and provide policy evaluations as outputs. Both types of networks feature intermediate connectivity through two 256×1 fully connected layers transitioning from input to output.

During training, the actor network produces an action A_{RL}, which is subsequently evaluated against a defined disharmony threshold. Actions exceeding this threshold are guided towards harmony using the harmonious action A_{GT}, derived from the IISG model. Once the action is executed, the actor network undergoes updates based on the EV-related reward Q_R and the harmony-related value Q_H.

2) Observation Space: The observation space serves as the input to the CHRL decision system. In our framework, the observation space is represented by a 7×5 matrix, which includes information on 7 vehicles: 6 SVs from different orientations within LCW, and the EV. Note that we model interactions solely between the ego vehicle and immediately adjacent vehicles, which excludes the indirect effects from vehicles in non-adjacent lanes. Subsequent research aims to refine this by developing a more elaborate model of multi-vehicle game-theoretic interaction chains. This advancement will facilitate a more holistic understanding of the interactions between the ego vehicle and vehicles situated beyond its immediate vicinity. Each vehicle is associated with an observation information vector, which can be denoted as: $\mathbf{S} = [S_i], i = 1, 2, \ldots 7$, where $S_i = \left[p_i, s_i, l_i, v_{s_i}, v_{l_i}\right]$, in which '$i$' refers to the i-th vehicle, p_i serves as an existence flag in the context of vehicle observation, and it is set to 0 when an observed vehicle falls outside of the observation range or if there is no vehicle present. s_i, l_i indicates the relative longitudinal and lateral distance to ego vehicle, and the v_{s_i}, v_{l_i} are correspondingly differential terms of distance.

3) Action Space: The action space A_{RL} consists of three continuous variables: x_d, a, and lane_{tar}.

The term lane_{tar} represents a discrete target lane-change action from the set (left, half-left, keep, half-right, right). It should be noted that the 'keep' action is understood as a commitment to the prevailing action. Specifically, if the vehicle is proceeding straight, 'keep' translates to maintaining this straight course; however, once a lane change is triggered, 'keep' then denotes the sustained execution of this ongoing lane change. To incentivize the agent to investigate partial lane-change maneuvers and optimize opportunities for lane changes, we introduce a "half-lane" option into the action space.

- x_d denotes the end position of the lane change on the target lane generated from action space. x_d only determines the lateral trajectory shape and is decoupled from the longitudinal speed, primarily influencing the smoothness of the EV's lane change. To better approximate real-world conditions, the range of x_d is set to [10m, 45m] along the x-axis relative to the current position of the EV.
- a represents the longitudinal acceleration, ranging from $[-3,3]m/s^2$.
- lane_{tar} means a discrete target action from (left,half-left, keep,half-right, right), we have introduced a half-lane option as a choice of actions to encourage the agent to explore possibilities in half-lane probing and seizing lane-change opportunities. A step function outlined in equation (1), is employed to map the continuous output a_{con} of the CHRL model onto one of the discrete actions lane_{tar}. As a constraint, when a vehicle is situated in the leftmost (or rightmost) lane, any action from lane_{tar} indicating a "left" ("right") lane change is automatically converted to "keep" during its application.

$$\text{lane}_{tar} = \begin{cases} \text{left}, & 0.0 \le a_{con} \le 0.5 \\ \text{half-left}, & 0.5 < a_{con} \le 1.0 \\ \text{keep}, & 1.0 < a_{con} < 2.0 \\ \text{half-right}, & 2.0 \le a_{con} < 2.5 \\ \text{right}, & 2.5 \le a_{con} \le 3.0 \end{cases} \tag{1}$$

Besides, if the vehicle goes off-road, the current episode will be terminated, and a penalty will be applied.

4) Reward: The reward setting is divided into three components. Firstly, the safety and efficiency of the ego vehicle; second, the lane-change exploration. Lastly, a penalty is set for dangerous actions. The detailed equation is as follows:

(1) ego vehicle reward

(a) safety reward

$$R_s = \Delta s_{el} \times r_{\text{TTC}} + \text{flag}_c \times p_c \tag{2}$$

The safety reward includes a constant weight r_{TTC} of the time-to-collision (TTC) reward multiplied by the distance between the ego vehicle and the leading vehicle Δs_{el}. A negative collision penalty weight p_c is applied if a collision happens.

(b) high-speed reward

$$R_e = r_e |v_{cur} - v_{tar}| \tag{3}$$

where r_e is a constant weight of efficiency reward, v_{cur}, v_{tar} are current speed and target speed respectively. The reward speed interval is [10,30]m/s.

(c) fluctuation penalty

$$R_f = \begin{cases} \deg_f \times p_f, & n_f > 3 \\ 0, & n_f \le 3 \end{cases} \tag{4}$$

where $\deg_f$ is the ratio of inconsistent actions to the historical actions in the past several seconds, p_f is a constant weight of fluctuation penalty. Considering the comfort of driving, the EV will be fined if it changes the target lane frequently.

(2) opportunity exploration reward

$$R_{ex} = \begin{cases} r_{ex}, & t_{ex} \le 1.5 \\ 0, & t_{ex} > 1.5 \end{cases} \tag{5}$$

An exploration reward weight r_{ex} is given for actions that explored a half-lane action and completed a lane change through exploration, and the exploration time was limited to 1.5 seconds.

(3) danger penalty

$$R_d = \begin{cases} \deg_d \times p_d, & \text{act = dangerous} \\ 0, & \text{act = safe} \end{cases} \tag{6}$$

where $\deg_d$ represents the degree of deviation from the reference harmonious action, the reference harmonious action is determined by the harmony guidance module. And p_d refers to a constant disharmony penalty weight.

All reward (or penalty) weights used in the reward function are set in the range $(0, 1)$. The reward (or penalty) values will be normalized to $R'_s, R'_e, R'_f, R'_{ex}, R'_d$, based on their maximum and minimum values to remove the influence of their scale. The final comprehensive reward is the sum of these normalized rewards:

$$R' = R_s{}' + R_e{}' + R_f{}' + R_{ex}{}' + R_d{}' \tag{7}$$

The comprehensive reward R' can generally be used to criticize a lane change decision.

5) Parameterized Trajectory Generation: Our method yields a trajectory as output, which is subsequently projected onto the Frenet [30] coordinate system and discretized into longitudinal and lateral coordinates (s, l) along the lane centerline. This representation facilitates trajectory tracking using the Stanley control algorithm [31] to update the environment state.

A coefficient dictionary is included specifically for generating 5th-order polynomial lane-change trajectories, as equation (8) shows.

$$\begin{aligned} & y(x) = a_5x^5 + a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0 \\ & \text{s.t.} \quad y(x_0),\ y'(x_0),\ y''(x_0) = f_{\text{init}}\ (v_0, \omega_0, \varphi_0) \\ & \qquad y(x_d),\ y'(x_d),\ y''(x_d) = f_{\text{end}}\ (v_d, \omega_d, \varphi_d) \end{aligned} \tag{8}$$

The variable x in equation (8) represents the lane change endpoint x_d during calculation. We can use the starting and ending vehicle speeds v_0, v_d, steering speeds ω_0, ω_d, and heading angles φ_0, φ_d to establish the initialization and endpoint boundary conditions of the lane change trajectory and map

the lane change endpoint x_d to the coefficients of a fifth-order polynomial.

This dictionary stores the mapping information between the trajectory and the output action of the network. By providing the endpoint x_d, the dictionary gives a unique set of coefficients corresponding to the lane-change trajectory, which is then transformed into the Frenet coordinate system for execution.

IV. Dual Critics in CHRL

A. The Structure of Dual Critics

In dual critics, the E-Critic network is used to encourage the EV-related reward based on the equation (7), and the P-Critic network is to restrain the egoism of policy.

The value of the action-state pair calculated by the E-Critic network is recorded as Q_R. Based on the Bellman equation [32], it can be written as:

$$Q_R(s_t, a_t) = \gamma \mathbb{E}_{s_{t+1}}\left[V\left(s_{t+1}\right)\right] + R(s_t, a_t) \tag{9}$$

where $R(s_t, a_t)$ is the EV-related reward designed in the MDP model in equation (7), $V_(s_{t+1})$ is calculated as below:

$$V(s_t) = \mathbb{E}_{a_t \sim \pi}\left[Q_R(s_t, a_t) - \alpha \log \pi(a_t \mid s_t)\right] \tag{10}$$

In the P-Critic network, to synchronously evaluate the harmony of action-state pairs, we summarize the harmony-related value $Q_H(s_t, a_t)$ in equation (11).

$$Q_H(s_t, a_t) = \gamma \mathbb{E}_{s_{t+1}}\left[H\left(s_{t+1}\right)\right] - C_H(s_t, a_t) \tag{11}$$

where $C_H(s_t, a_t)$ is the harmony cost derived from the IISG cost matrix, including safety, comfort, and efficiency cost components, which will be derived in the next subsection. The opposite value of $C_H(s_t, a_t)$ represents the harmony utility; $H(s_{t+1})$ represents the harmony-related value of state s_{t+1}, which is calculated as follows:

$$H(s_t) = \mathbb{E}_{a_t \sim \pi}\left[Q_H(s_t, a_t) - \alpha \log \pi(a_t \mid s_t)\right] \tag{12}$$

B. The Composition of Harmony Cost

Harmony cost aligns with the cost function in the IISG model and quantifies the "rewards" for game participants, aiming for an optimal outcome that enhances both the EV's and the trailing vehicle's interests as modeled in our lane-changing IISG framework (detailed in Section V). The "reward" of a lane change can be calculated using safety, efficiency, and comfort metrics [33], thus we define the Harmony cost in terms of these 3 key performance indicators.

1) Safety Cost: This part includes safety-related costs for ego vehicles and potential right-of-way competitors, i.e. the following SV of the target lane. The safety cost is the summary of weighted lateral and longitudinal safety costs:

$$C_{\text{safe}} = \omega_{\text{safe}_{\text{lat}}} C_{\text{safe}_{\text{lat}}} + \omega_{\text{safe}_{\text{lon}}} C_{\text{safe}_{\text{lon}}} \tag{13}$$

which includes the following components:

(1) Lateral safety cost

The lateral safety cost represents the feasibility of lane change of the vehicle.

$$C_{\text{safe}_{\text{lat}}} = \kappa_{v_{\text{lat}}} \lambda_{\text{eo}} \Delta v_{\text{eo}}^2 + \frac{\kappa_{s_{\text{lat}}}}{\Delta s_{\text{eo}}^2 + \epsilon} \tag{14}$$

in which $\kappa_{v_{\text{lat}}}, \kappa_{s_{\text{lat}}}$ are the weights for the relative speed Δv_{eo} and relative distance Δs_{eo} between the EV and the opponent vehicle (OV) in the game, λ_{eo} is flags indicating the speed deviation between EV and OV within the interaction range, ϵ is a small value to prevent division by zero.

(2) Longitudinal safety cost

The longitudinal cost mainly considers the distance and speed difference between the EV and the lead vehicle in the same lane.

$$C_{\text{safe}_{\text{lon}}} = \kappa_{v_{lon}} \lambda_{le} \Delta {v_{le}}^2 + \frac{\kappa_{s_{lon}}}{\left(\Delta {s_{le}}^2 + \varepsilon\right)} \tag{15}$$

where $\kappa_{v_{lon}}, \kappa_{s_{lon}}$ are the corresponding weights for relative speed Δv_{le} and relative distance Δs_{le} between EV and its leading vehicle.

2) Comfort Cost: The comfort cost component captures the impact of acceleration on the comfort of vehicles.

$$C_{\text{com}} = \kappa_{ax} {a_x}^2 + \kappa_{ay} {a_y}^2 \tag{16}$$

where κ_{ax}, κ_{ay} are the weights for longitudinal and lateral accelerations.

3) Efficiency Cost: The efficiency cost represents the acceleration capability of the EV.

$$C_{\text{eff}} = (v_e - v_{\text{lim}})^2 \tag{17}$$

where v_e is EV speed and v_{lim} is the speed limit of road.

The comprehensive harmony cost integrates these cost components into a function:

$$C_H = \omega_{\text{safe}} C_{\text{safe}} + \omega_{\text{com}} C_{\text{com}} + \omega_{\text{eff}} C_{\text{eff}} \tag{18}$$

The cost components within C_H will be normalized by their respective differences between maximum and minimum values to eliminate the impact of dimensional scales. After normalization, the resulting C'_H is used to assess the harmony of the policy while considering aspects of safety, comfort, and efficiency. During the training process, equal importance was assigned to these three aspects to ensure balanced optimization without bias toward any single factor. Specifically, the weighting parameters ω in Equation (18) were each configured with a value of 0.33, reflecting this balanced approach to harmony quantification. This parameterization enables the model to develop lane-changing policies that simultaneously optimize for all three dimensions of harmony rather than privileging one aspect at the expense of others.

C. The Update Algorithm With Dual Critics in CHRL

The update algorithm with dual critics during training in CHRL can be constructed in Algorithm 1:

According to Algorithm 1, once a batch of experience pool data is accumulated, the update process begins. The Q-value and harmony cost are calculated based on the action-state pairs stored in the experience pool. Then the loss of both critics is calculated. By updating the E-Critic and P-Critic with losses, we can then refine the estimations of different state-action pairs. This, in turn, enables the actor's policy to maximize expected cumulative rewards and minimize the harmony cost.

Algorithm 1 The Update Algorithm During Training

for each time step t **do**
  $ent_{t+1}R, H, a, obs_{t+1} \leftarrow \mathbb{E}[P_{log}], Env(obs_t)$
  $s, a \rightarrow buffer$
  if after a batch **then**
    $\mathbf{a}, \mathbf{s}, P_{log} \leftarrow$ buffer
    update Q_R, Q_H
    $Q_R = arg_{min}(\mu_r[\mathbf{a},\mathbf{s}] - a_{\mathrm{tar}_r}{}^*)$
    $Q_H = arg_{min}(\mu_h[\mathbf{a},\mathbf{s}] - a_{\mathrm{tar}_h}{}^*)$
    $Q_{\pi(a|s)} = \omega_R Q_R + \omega_H Q_H$
    update $critics \leftarrow loss(Q_R), loss(Q_H)$
    update $actor \leftarrow loss(\pi(a|s))$
  end if
end for

An entropy value in the form of equation (19) is incorporated as a soft part.

$$Ent(\pi(a \mid s_t)) = \mathbb{E}_{x\sim p}[-\log(\pi(a \mid s_t))] \tag{19}$$

The soft part is used to find a balance between utility and exploration.

V. Policy Generation With Harmony Guidance

A. Harmony Assessment

Dangerous actions like collisions once occur, would result in extremely terrible damage, so it should be treated more seriously in contrast to normal disharmony. To enhance the recognition of CHRL for dangerous actions, we have introduced a harmony assessment mechanism. This mechanism is used to label dangerous actions, which are then guided through IISG to improve harmony.

To assess the danger, we first predict the trajectory of SVs with the method in [34] and then calculate the collision probability between EV and SVs based on the equation (20a).

$$P_{coll}(T_{\text{plan}}, T_{\text{pre}}) = \frac{1}{X}\sum_{i=1}^{X} I_c(S_{EV}, S_{SV}) \tag{20a}$$

$$\text{Danger} = \begin{cases} True, P_{coll} \geq P_{\text{thres}} \\ False, P_{coll} \leq P_{\text{thres}} \end{cases} \tag{20b}$$

where T_{pre} stands for the predicted trajectories of SVs with Gaussian uncertainty, S_{EV} is the set of planned trajectory poses of EV at time t, S_{SV} is the correspondingly set of predicted trajectory poses of the SVs according, I_c is used to mark whether the i-th sampled pose has a collision. The collision probability is the ratio of the number of collision points number to the total number X of sampling points. The dangerous action is marked if the collision probability over a high disharmony threshold, as equation (20b) shows.

B. Harmony Guidance

For dangerous actions, an IISG model is established to generate a reference harmony action as expert demonstration.

TABLE I
Safety, Comfort, and Efficiency Weights for Different Characters

Type Coefficient	Aggressive	Normal	Modesty
w_s	0.3	0.5	0.7
w_c	0.2	0.25	0.3
w_e	0.5	0.25	0

1) the IISG Model: The key parts of a game are defined as follows: (1)Game players: EV and game vehicle (*GV*), i.e., the trailing vehicle in the target lane of the EV. (2) Knowledge space: consistent with the RL observation space. (3) Game matrix: The game cost of different actions is calculated using the equation (18) of harmony cost in the P-critic network to maintain consistency with the overall network's representation of harmony. Assuming that the EV is engaged in a lane-change game and the GV maintains its current trajectory, the feasible actions sampled from the action space for the EV is $\mathrm{A_{GT}} = (x_d, a, \mathrm{lane}_{tar})$, and assume that the GV has no lane changing, only longitudinal acceleration actions. In action space, x_d ranging from $[10, 45]$m with a constrain of max steering angle and acceleration, a ranging from $[-3, 3]$m/s, and lane_{tar} is an action from a discrete set (left,half-left, keep,half-right, right).

2) Game Solution: To model the drivers more realistically, we categorize the GV into three character types: aggressive, normal, and modest. Based on the historical acceleration a_{gv} of the GV, we can infer the character distribution of the GV's character with equation (21).

$$\left(P_{\text{agg}}, P_{\text{aco}}, P_{\text{mod}}\right) = \begin{cases} (0.0, 0.2, 0.8), a_{gv} < -2\text{m/s} \\ (0.1, 0.8, 0.1), -2\text{m/s} \leq a_{gv} \leq 2\text{m/s} \\ (0.8, 0.2, 0.0), a_{gv} > 2\text{m/s}. \end{cases} \tag{21}$$

Based on the utility matrix introduced in the P-critic network, we update the GV's payoff in the utility matrix with character according to the personality weights in Table I. Then a fictitious play algorithm [35], [36] is set to learn the Bayesian Nash Equilibrium.

Algorithm 2 Game Solution

input $S_{EV}, S_{GV}, \mathrm{action}_{\mathrm{sample}}$
  while not converged **do**
  for each action pair in $\mathrm{action}_{\mathrm{sample}}$ **do**
  $cost_{EV}, cost_{GV} \leftarrow CalCost(t_{\mathrm{prediction}})$
  end for
  Update game matrix
  end while
output $\mathrm{A_{GT}} \leftarrow$convergent Bayesian Nash equilibrium

As the Algorithm 2 shows, in every iteration, a Nash equilibrium solution is calculated after traversing all possible action samples. After 100 iterations, the Bayesian equilibrium solution converges to the dominant Nash equilibrium solution and becomes the reference harmonious action.

C. Expert Demonstration Experiences Enhancement

At the network level, considering the P-Critic's targeted focus on harmony, we designed a mechanism that emphasizes dangerous actions in the prosocial experience pool. We treat episodes following Harmony Guidance as "expert demonstration experiences", and redeploy them for prioritized experience replay.

The structure of the experience pool with prioritized replay of expert experiences can be represented as follows:

$$\begin{aligned}\text{EV-related Experience} &= \left\{(\mathbf{a},\mathbf{s})_{\mathbf{safe}},(\mathbf{a},\mathbf{s})_{\mathbf{danger}}\right\}\\ \text{Prosocial Experience} &= \left\{(\mathbf{a},\mathbf{s})_{\mathbf{safe}},(\mathbf{a},\mathbf{s})_{\mathbf{danger}},(\mathbf{a},\mathbf{s})_{\mathbf{danger}}\right\}\end{aligned} \tag{22}$$

This mechanism establishes a connection between the policy generation with harmony guidance and the dual critics, unifying their objectives to foster harmony.

D. The Policy Update With Convergent Harmony Mechanism

Combining dual critics with harmony-guided policy generation, we implement the convergent harmony mechanism that can update the agent's policy under the algorithm 3:

Algorithm 3 Policy Update of CHRL

for each time step t **do**
  $ent_{t+1}R, H, \mathrm{A_{RL}}, obs_{t+1} \leftarrow \mathbb{E}[P_{log}], Env(obs_t)$
  $a = \mathrm{A_{RL}}$
  Harmony assessment of current $\mathrm{A_{RL}}$.
  if $\mathrm{A_{RL}}$ is danger **then**
    Harmony guidance of the dangerous action.
    $a = \mathrm{A_{GT}}$
    Save the dangerous s, a pair to the prosocial experiment pool.
    $s, a \rightarrow harmony\_buffer$
  end if
  Save the s,a pair to the EV-related experiment pool.
  $s, a \rightarrow buffer$
  Update dual critics.
  if after a batch **then**
    update $Q_R, Q_H \leftarrow$ buffer, harmony_buffer
    Update actor network based on dual critics.
    update actor $\leftarrow \omega_R Q_R + \omega_H Q_H$
  end if
end for

The dual critics and the harmony guidance during policy generation can interact with each other and promote the agent to achieve an overall more harmonious policy.

VI. Experiments

A. Experimental Objectives and Design

This experimental study aims to verify whether CHRL consistently produces smooth, stable, and efficient harmonious lane changes across varying conditions. We will evaluate CHRL's performance in both random scenarios with different traffic densities and real-data scenarios featuring human driving behaviors, benchmarking against rule-based methods, human drivers, and other harmony-focused RL agents.

Using tailored harmony assessment metrics, we specifically examine the lane-changing harmony when CHRL interacts with both computational and human-driven vehicles in diverse traffic patterns.

To describe the experiment configuration, some key experimental parameters are set:

- Traffic density: The number of vehicles in a specific stretch of the three-lane highway.
- Gap: The distance between consecutive vehicles in the traffic flow.
- $\bar{v}$lc: The average lane-change speed of EV.
- $\bar{v}\mathrm{lc_{LCW}}$: The average lane-change speed of all vehicles in LCW.

The relationship between traffic density and the initial gap is as follows:

$$Gap = \frac{1}{\text{density}} \min dS \tag{23}$$

where $\min dS$ is the minimized default inter-vehicle distance, which is a discrete value derived from common driving practice that increases with the EV's velocity. The specific gap values corresponding to different traffic densities according to Equation 23 are presented in Table II, demonstrating how our experimental setup accounts for variations in traffic conditions while focusing on interaction harmony within the LCW.

B. Metrics Design

The differences in safety and efficiency of the EV under different lane-change methods are measured to evaluate the overall lane-change capability. Related indicators are also statistically analyzed to evaluate the overall harmony of lane changes.

1) Safety Metrics: (1) the collision number (n_{crash}); (2) the collision rate (r_{crash}) of all decisions.

2) Efficiency Metric: (1) the number of lane changes n_{lc}: It is one of the signs of lane change ability; (2) the average EV lane-change speed $\bar{v}_{\mathrm{lc}}$; and (3) the average lane change time $\bar{t}_{lc}$.

3) Harmony Metrics: (1) the Mean absolute deviation (MAD) of vehicles' lane-change speeds in the LCW ($\mathrm{MAD}_{v_{lc}}$) [11]: This index mainly measures the overall safety during lane-changing. Note that MAD focuses on the "potential risks" during the interaction, while the collision rate in the safety metric is based on the vehicle's perspective. A smaller MAD value indicates an overall safer LCW. The $\mathrm{MAD}_{v_{lc}}$ can be calculated by equation (24).

$$\mathrm{MAD}_{v_{lc}} = \frac{\sum_{t=0}^{t_{lc}} \mathrm{MAD_{LCW}}}{t_{lc}} \tag{24}$$

where,

$$\mathrm{MAD_{LCW}} = \frac{\sum_{i=0}^{n} v_i - \bar{v}_{i\mathrm{LCW}}}{n} \tag{25}$$

refers to the MAD of speed between vehicles in the LCW, which reflects the stability of the current snapshot. $\mathrm{MAD}_{v_{lc}}$ in equation (24) is the average $\mathrm{MAD_{LCW}}$ of all time steps in an episode, reflecting the overall safety of the LCW.

TABLE II
THE CONSTRUCT OF 3 SCENARIOS

Scenarios	Lane number	Lane width	Lane length	Speed limit	Target speed	Traffic density	Gap(m)	Motion model	Control Model
Random Scenario	3	3.5m	2km	33m/s	25m/s	[0.6,0.7,0.8,0.9,1.0]	[45,38.57,33.75,30,27]	EV:test methods SV:IDM [38],Mobil [39]	Stanley [31]
HighD1	From dataset		420m	From dataset		/	/	EV:test methods SV:human	EV:Stanley SV:human
HighD2	From dataset		420m	From dataset		/	/	EV:test methods SV:IDM,Mobil	Stanley

(2) Absolute lateral and longitudinal acceleration $|\bar{a}_x|, |\bar{a}_y|$: $|\bar{a}_x|$ represents the stability of the lane-change policy [37]. Harmonious lane changes should adopt a smooth policy to minimize the $|\bar{a}_x|$. Although a larger $|\bar{a}_y|$ will lead to faster lane changes, it will also reduce comfort. However, a too-small lateral acceleration will also cause an unreasonable long lane-change distance and even weaken the EV's efficiency. This may then go against the original intention of free lane changes. Therefore, $|\bar{a}_y|$ can be used to observe whether the lane-change policy can balance efficiency and comfort.

4) Human-Likeness: we additionally analyzed the Human-likeness in the HighD scenario. The metric is speed deviation $dev(v)$ between the tested methods and human drivers, in equation (26).

$$\mathrm{MD}(v) = \frac{\sum_{i=0}^{n} \left| v_m^{t_i} - v_h^{t_i} \right|}{n} \tag{26}$$

We analyzed the human-likeness of EV speed and LCW speed during lane-changing, to observe whether a lane-change method can meet human efficiency needs while enhancing harmony.

C. Scenario Construction

We devised both random and real-data scenarios for our experimentations, applying random scenarios during training and ablation experiments, while leveraging real-data scenarios for cross-comparison experiments with other mainstream methods.

For randomly generated scenarios, the inter-vehicle initial *Gap* (detailed in Table II) was derived by applying the predetermined traffic density to Equation 23.

The random scenario is based on the Highway-Env platform. As for the real data scenario, we generated a collection of test scenarios based on the HighD dataset.

The construction of scenarios is as Table II. In the random scenario, the agents were tested 150 times under the traffic densities range in Table II, resulting in a total of 33,750 decision numbers. While real-data scenarios include nearly 33,000 decisions in 1,417 cases from the HighD dataset. These cases contain lane changes with a minimum TTC of less than 6 seconds. We constructed 2 real-data scenarios, both scenarios chose only one human-driven vehicle that performed lane changes in the original dataset as the EV in each episode, and different decision-making methods were used for the EV. SVs in HighD1 are driven under the origin dataset to analyze the human-likeness of tested methods, while SVs in HighD2 drive under the IDM and Mobil methods to interact with the EV so that can measure the impact of the EV's lane-change policy on the LCW.

TABLE III
STRUCTURES OF SAC, SACP, SACH, CHRL

Structures	SAC	SACP	SACH	CHRL
Soft part	✓	✓	✓	✓
Actor-Critc Network	✓	✓	✓	✓
Harmony guidance	×	×	✓	✓
Harmony cost	×	×	✓	✓
P-Critic Network	×	✓	×	✓

D. Models Structure

In ablation experiments with the random scenario, the SAC (Soft Actor-Critic, SAC) [40] model is set as the baseline, To more specifically analyze the contribution of the P-Critic network and harmony guidance module to CHRL, we carry out ablation experiments with two other agents: the lane-change agent based on the SACP (Soft Actor-Critic-Prosocial, SACP) network and based on the SACH (Soft Actor Critic with the harmony guidance module, SACH) network, respectively. Table III shows the architecture comparison of the SAC-based agents in ablation experiments. Hyperparameters setting follows the expressions: learning rate = $\max\left(10^{-3}, 0.999^{n_{\text{eps}}} 10^{-2}\right)$, Batch Size = 256, Discount Factor(γ) = 0.99, Entropy Coefficient(τ) = 0.005. Harmony weight = $\min\left(0.5, 0.05 + 1.0004^{n_{\text{eps}}}\right)$ in SACP and CHRL, and it's set to be 0 in SAC and SACH.

In cross-comparison experiments with HighD scenarios, the following mainstream methods are included:

(1) Rule-based method: The Intelligent Driver Model (IDM) [38] is utilized as a representative baseline approach.
(2) RL agent with disharmony penalty: As opposed to the dual critics structure, we introduce a Disharmony-Punished SAC (DPSAC) model. The disharmony penalty in DPSAC is computed using equation (18) and incorporated into the comprehensive reward R'. Subsequently, the rewards are evaluated by a single critic. By comparing with the DPSAC model, we can identify the enhancements achieved by the dual critics structure.
(3) RL agents with rule constraints: three additional comparative methods that employ rule constraints to enhance the lane-change capabilities of the RL agent are introduced. i) The first is a Collision-Risk Constrained RL

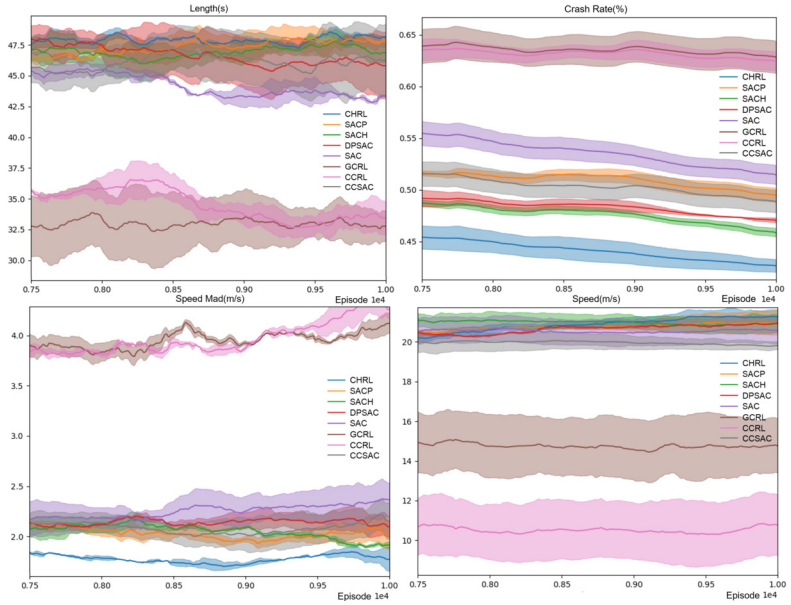

Fig. 3. The training result of learning-based models.

(CCRL) model leveraging the Double Deep Q-Network (DDQN) framework [41], [42], which includes collision risk as a safety constraint to refine lane-change strategies; ii) To further enhance the structural control over the network, this study also compared the SAC network with safety rule constraints, referred to as Constrained-Safety SAC (CCSAC); iii) The third is a Game Theory constrained RL (GCRL) method based on the DDQN framework [21], which incorporates a GT model to influence policy during training. By comparing with the rule-constraint RL method, we can demonstrate the advantages of the CHRL framework in promoting policy exploration.

All models utilized the same EV-related reward, network architecture, and observation space, and were trained using identically designed Random Scenarios to ensure a consistent training environment.

E. Training Rusults

We trained each model five times and conducted statistical analyses on four performance metrics: episode lengths, crash rate, average speed, and mean speed deviation within the LCW.

Fig. 3 and Table IV show the training metrics for 7 methods, focusing on the last 2000 episodes. In Fig, 3, the colored lines indicate the average metric values for each method, while the color blocks around them represent the range of results from 5 training runs, with Gaussian smoothing applied.

The training metrics show that compared to SAC, CHRL achieves 8.6% longer episodes, 27.6% fewer collisions, 1.5% higher speeds, and 23.5% more consistent LCW speed, highlighting its safe, efficient, and harmonious policy. CHRL outperforms other learning-based methods by finding more balanced strategies and a better understanding of interaction harmony, leading to enhanced lane-change safety and stability–fulfilling our design goals.

TABLE IV
TRAINING RESULTS OF LAST 2000 EPISODES

Models	$\overline{\text{length}}$	$\bar{r}_{\text{crash}}(\%)$	MAD_v	$\bar{v}$
CHRL	**47.70**	**0.42**	**1.92**	**21.69**
SACP	46.49	0.52	2.05	21.57
SACH	46.17	0.48	2.07	21.38
DPSAC	45.99	0.48	2.31	21.24
SAC	43.94	0.58	2.51	20.72
CCSAC	44.47	0.48	2.06	20.15
GCRL	32.87	0.66	4.51	14.86
CCRL	35.27	0.61	4.12	11.25

The ablation models SACP (only with P-Critic) and SACH (only with Harmony Guidance) outperform SAC in all metrics. With an average around 17.9% less MAD and a 5.6% longer episode duration than SAC, these features show that convergent harmony module improves both lane-change performance and LCW stability. SACP enhances efficiency over SAC and SACH, indicating that interaction guidance during evaluation aids optimal policy exploration. However, SACH achieves a superior crash rate, decreasing by 17.3% compared to SAC, benefiting from its harmony constraints during execution. The ablation models also demonstrate the benefits of the P-Critic and Harmony Guidance for safe, stable and efficient LCW.

F. Test Results of Random Scenario

We conducted ablation experiments using random scenarios and analyzed the results based on corresponding metrics.

1) Safety: As safe metrics in Table V show CHRL has a 0-collision result, while SAC has the highest crash number. SACP, which introduces the P-Critic network, will pay attention to the safety cost of interaction, so its collision rate is less. SACH conducts harmonious guidance when making decisions, replacing dangerous actions with safer IISG solutions, which

TABLE V
STATISTICS TABLE OF EFFICIENCY METRICS & HARMONY METRICS IN RANDOM SCENARIOS

	n_{crash}				$r_{crash}(\%)$				n_{lc}				$\bar{v}_{lc}$			
Density	SAC	SACH	SACP	CHRL	SAC	SACH	SACP	CHRL	SAC	SACH	SACP	CHRL	SAC	SACH	SACP	CHRL
0.6	2	0	0	**0**	0.03	0.00	0.00	**0.00**	175	163	195	**230**	17.72	15.53	16.99	**16.49**
0.7	3	0	0	**0**	0.04	0.00	0.00	**0.00**	175	150	228	**222**	17.61	15.08	16.09	**16.33**
0.8	3	0	0	**0**	0.04	0.00	0.00	**0.00**	164	161	204	**217**	17.25	14.8	15.50	**15.95**
0.9	8	2	1	**0**	0.12	0.03	0.01	**0.00**	160	134	200	**208**	16.94	14.64	15.49	**15.51**
1.0	9	2	1	**0**	0.07	0.01	0.004	**0.00**	155	155	196	**191**	16.62	13.3	14.27	**15.50**
mean	5	0.8	0.4	**0**	0.06	0.01	0.003	**0.00**	165.8	152.6	204.6	**213.6**	17.23	14.67	15.67	**15.96**
	$\bar{t}_{lc}$				$\text{MAD}_{v_{lc}}$				$\lvert\bar{a}_x\rvert$				$\lvert\bar{a}_y\rvert$			
Density	SAC	SACH	SACP	CHRL	SAC	SACH	SACP	CHRL	SAC	SACH	SACP	CHRL	SAC	SACH	SACP	CHRL
0.6	6.07	5.60	5.78	**5.48**	1.05	0.95	0.89	**0.55**	1.31	0.78	0.90	**0.98**	0.24	0.2	0.22	**0.22**
0.7	6.17	5.90	5.87	**5.55**	1.05	0.95	0.82	**0.54**	1.68	0.80	0.84	**0.91**	0.23	0.18	0.22	**0.21**
0.8	6.32	6.14	5.96	**5.71**	0.95	0.85	0.70	**0.53**	1.24	0.95	0.87	**0.94**	0.22	0.18	0.20	**0.21**
0.9	6.31	6.07	6.06	**5.84**	0.86	0.8	0.67	**0.52**	1.22	0.89	0.81	**1.00**	0.21	0.21	0.19	**0.20**
1.0	6.93	6.43	6.11	**6.01**	0.85	0.73	0.59	**0.48**	1.21	0.79	0.89	**1.00**	0.22	0.18	0.19	**0.20**
mean	6.36	6.03	5.96	**5.72**	0.95	0.86	0.73	**0.52**	1.33	0.84	0.86	**0.97**	0.22	0.19	0.2	**0.21**

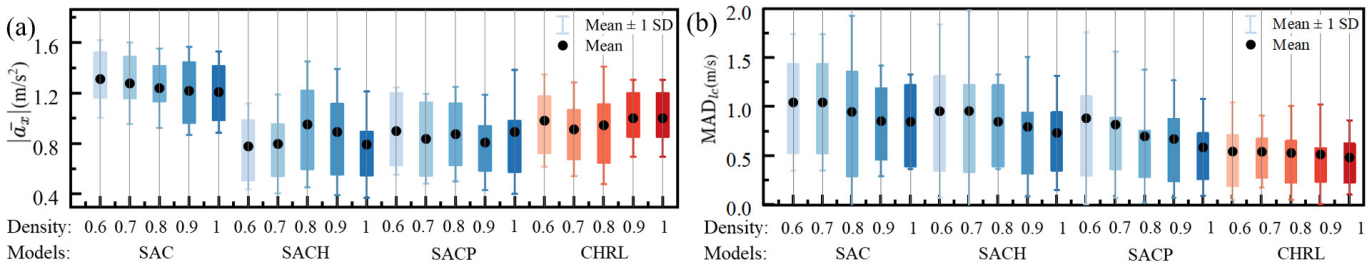

Fig. 4. The boxplot of lane-change speed MAD of vehicles in LCW, with SAC, SACP, and SACH methods in blue and CHRL method in red.

directly strengthens interaction safety of decision, so the collision rate of SACH is only 0.002%. SACH still had a crash, it is because the game-based harmonious guidance mechanism would only intervene when the collision risk was higher than 0.8. However, during the crash episode, the vehicle speed was too high, and the front vehicle was slower with a narrow distance. Despite SACH's decision to slow down with the maximum acceleration, an unavoidable collision still occurred. This problem shows that the risk threshold and the longitudinal action space can be further optimized. Our group will further consider related issues. In CHRL, the harmony guidance and the P-Critic can mutually promote each other to enhance the harmony of the lane-change decision, making CHRL have the highest interaction safety and completely avoid collisions in the test set.

2) Efficiency: As can be seen from Table V, CHRL takes the harmony of interactions into account during action evaluation and generation, which improves the efficiency of EV based on overall interaction safety. The n_{lc} of CHRL is significantly increased by 39.97% compared to SAC, and the $\bar{t}_{lc}$ is shortened by 7.73% compared to SAC, achieving the maximum number of lane changing while ensuring the lowest collision. It is proved that by introducing the lane-change IISG interaction model, the safety and efficiency of lane changes can be improved simultaneously.

3) Harmony Metrics: The harmony metrics in V demonstrate CHRL's superior performance with the lowest MAD*vlc* and average smallest $|\bar{a}x|$. Compared to standard SAC, CHRL reduces speed discreteness in LCW by 45.30% and $|\bar{a}x|$ by 17.17%, with only a slight increase in $|\bar{a}_y|$, confirming its optimal harmony-safety balance.

Fig. 4 shows that CHRL maintains consistently lower MAD*vlc* distribution across all traffic densities. While congested conditions naturally reduce MAD*vlc* as vehicles maintain similar speeds to avoid collisions, CHRL outperforms all comparison methods. SACP, guided by harmony cost, shows improved performance over SAC. SACH, with harmony guidance during decision generation, further reduces MAD*vlc*. CHRL, integrating both mechanisms, achieves the best results by monitoring dangerous actions and optimizing for harmony during training.

CHRL demonstrates more stable acceleration profiles in congested traffic, avoiding sudden braking or rushing during lane changes. The slight increase in $|\bar{a}_y|$ enables faster lane-change completion, reducing the duration that might otherwise require more longitudinal adjustments, thereby enhancing overall LCW harmony.

4) Comprehensive Analysis: Comprehensive analysis reveals CHRL's superior performance across multiple dimensions. CHRL reduces potential risks through improved harmony guidance and enhanced speed consistency during lane changes. Compared to SAC, CHRL achieves 10% shorter lane-changing time while maintaining the lowest average $|\bar{a}_x|$, demonstrating an optimal balance between efficiency and comfort. This superior vehicle spacing control prevents risks from sudden speed variations, confirming that the convergent

TABLE VI

STATISTICS TABLE OF SAFETY, EFFICIENCY, AND HARMONY METRICS OF CROSS-COMPARISON METHODS IN HIGHD SCENARIOS

	Methods	n_{crash}	r_{crash}	MD($\bar{v}$)	MD($\bar{v}_{\text{LCW}}$)	n_{lc}	$\bar{t}_{\text{lc}}$	$\bar{v}_{\text{lc}}$	$\bar{v}_{\text{lc}_{\text{LCW}}}$	$\text{MAD}_{v_{lc}}$	$\lvert\bar{a}_x\rvert$	$\lvert\bar{a}_y\rvert$
HighD1	Human	0	0	0	0	1417	5.55	29.23	29.81	2.81	0.42	0.27
	IDM	3	0.01	-4.10	-2.61	1413	1.89	28.07	27.71	2.16	1.25	2.78
	GCRL	13	0.04	-17.92	-16.22	1089	9.61	20.31	22.58	2.87	0.94	0.16
	CCRL	12	0.04	-17.91	-16.2	1057	9.72	19.74	22.72	2.89	0.93	0.16
	CCSAC	9	0.03	-7.67	-4.08	953	8.99	27.03	25.18	1.93	0.90	0.21
	SACH	4	0.01	-6.81	-3.59	949	8.61	26.01	27.36	1.79	1.11	0.21
	SACP	3	0.01	-6.98	-3.79	958	8.59	25.8	27.16	1.73	0.83	0.19
	DPSAC	5	0.02	-5.03	-2.87	1519	7.05	27.22	28.24	1.23	1.09	0.23
	CHRL	**0**	**0.00**	**-4.81**	**-2.63**	**1592**	**6.81**	**28.08**	**28.20**	**0.69**	**0.94**	**0.20**
HighD2	Human	0	0	0	0	1417	5.55	29.23	29.81	2.81	0.42	0.27
	IDM	2	0.01	-3.93	-2.49	1339	1.99	28.03	28.03	2.15	1.30	2.70
	GCRL	4	0.01	-17.93	-16.44	1049	9.77	20.44	22.71	3	0.93	0.17
	CCRL	4	0.01	-17.93	-16.41	1046	9.81	19.72	27.80	3.05	0.94	0.16
	CCSAC	7	0.02	-7.64	-4.08	966	8.95	27.22	25.20	1.93	0.89	0.21
	SACH	8	0.02	-6.81	-3.21	999	8.67	26.59	27.93	1.78	1.11	0.21
	SACP	4	0.01	-6.76	-3.75	978	8.64	25.92	27.53	1.86	0.81	0.2
	DPSAC	3	0.01	-4.55	-2.66	1542	7.06	27.56	27.79	1.23	1.09	0.22
	CHRL	**0**	**0.00**	**-4.31**	**-2.43**	**1602**	**6.54**	**28.56**	**28.76**	**0.66**	**0.87**	**0.20**

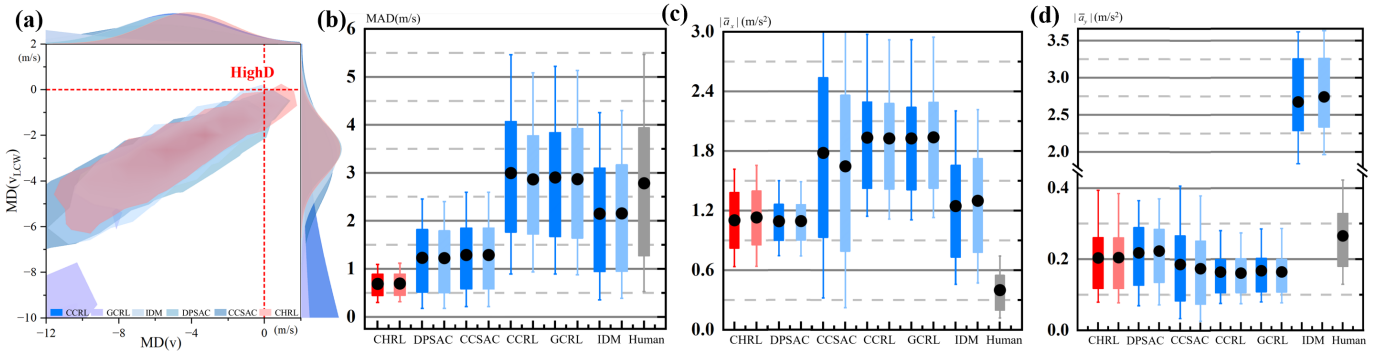

Fig. 5. (a) shows the distribution of lane-change speed deviation of EV and LCW between human-driven and cross-comparison methods; (b), (c), (d) are boxplots of the MAD(v_{lc}), $|\bar{a}_x|, |\bar{a}_y|$ for cross-comparison methods in HighD scenarios. The black dot represents the average value, the box indicates the 25th to 75th percentile distribution, and the vertical line denotes the 10th to 90th percentile distribution.

harmony mechanism significantly enhances both safety and lane-change efficiency.

G. Test Results of HighD Scenarios

Cross-comparison experiments are performed on the HighD Scenarios to evaluate the methods described in the model structure section, with subsequent analyses generated.

1) Safety and Human-Likeness: Through the statistics in Table (VI), it can be seen that the CHRL method can achieve the same 0 collision result as the Origin dataset, and has the smallest absolute values of MD($\bar{v}$) and MD($\bar{v}_{\text{LCW}}$) relative to human drivers. DPSAC resulted in a few collisions due to insufficient prediction of SVs during lane-change interactions and less comprehensive consideration for overall safety. CCRL, CCSAC and GCRL methods, which did not incorporate harmony cost, considered SVs even less during lane changes, producing more disharmonious trajectories that sometimes led to scrapes.

Meanwhile, IDM and DPSAC often do not pay enough attention to SVs, making collisions occur. It shows that CHRL can not only take into account the overall safety and avoid collisions in time but also strengthen the consideration of interactions in LCW so that CHRL pays more attention to the distance and speed differences between vehicles in LCW and avoids collisions in advance.

Comparing the human-likeness of the EV speed and the speed of the overall LCW in Fig. 5(a), both the MD of LCW and EV speed of CHRL are closer to the Human in the red dashed line. This demonstrates that CHRL is more human-like in efficiency compared to other agents.

2) Efficiency: CHRL outperforms in efficiency metrics n_{lc}, $\bar{t}_{lc}$, $\bar{v}_{lc}$, and $\bar{v}_{lc_{\text{LCW}}}$ as seen in Table VI, enabling faster and safer lane changes. Although IDM's lane change time is minimal, its aggressive policy overlooks SV awareness and increases collision risks. In HighD2, CHRL boosts average LCW speed $\bar{v}_{lc_{\text{LCW}}}$ by 2%(from 28.20m/s to 28.76m/s) with its harmonious strategy, showcasing the efficacy of combining Harmony Guidance and P-Critic for balanced safety and efficiency.

3) Harmony: Fig. 5 (b) shows that CHRL maintains consistently lower MAD_{vlc} across all scenarios, improving this metric by 45.5-77.1% compared to Human, DPSAC, CCRL, CCSAC, GCRL, and IDM models as detailed in Table VI. This confirms CHRL achieves the highest speed uniformity in LCW, creating more harmonious traffic flow.

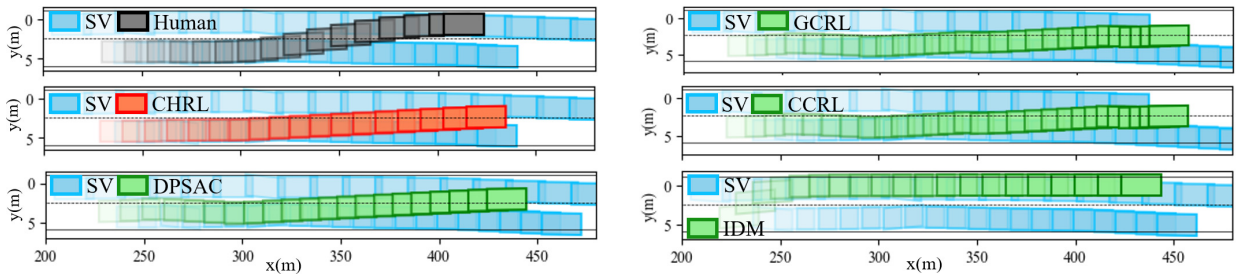

Fig. 6. A snapshot of a lane-change example under the same HighD scenario during testing.

Fig. 5 (c)(d) reveals distinct acceleration patterns. Human drivers exhibit small $|\bar{a}x|$ with large $|\bar{a}y|$, prioritizing personal efficiency while considering longitudinal harmony. CHRL adopts a similar strategy but with more balanced parameters–smaller $|\bar{a}y|$ and moderate $|\bar{a}x|$–optimizing both harmony and efficiency. In contrast, methods without harmony consideration show larger $|\bar{a}_x|$, indicating poor handling of surrounding vehicles and tendency toward aggressive maneuvers. CHRL's coupling of interaction mechanisms with P-Critic enables deeper understanding of lane-change harmony dynamics, resulting in superior performance across metrics.

Cross-comparison experiments confirm CHRL's ability to enhance both ego vehicle and overall traffic efficiency while maintaining optimal harmony in diverse traffic scenarios.

4) Case Discussion: Taking an example of test cases, we can further illustrate how CHRL achieves lane-change harmony.

The example in Fig. 6 illustrates that human drivers often choose more aggressive lane changes than learning-based models. CHRL, supported by an IISG model, provides a balance between the EV's and SVs' benefits, yielding trajectories that are smoother than both humans and more efficient than other learning-based approaches. CHRL's deeper understanding of interactions is evident as it successfully executes lane changes on the first try, while Other learning-based methods initiate premature lane changes near the 300m mark and have to correct back after encountering faster-trailing vehicles, leading to a secondary attempt.

This illustrates that integrating interactive guidance mechanisms and the P-Critic can significantly improve the CHRL's interaction awareness and its ability to optimize the LCW benefits.

5) Comprehensive Analysis: Concurrently, leveraging dependency decoupling within its dual critics architecture, our approach demonstrably enhances key harmony performance indicators.

VII. Conclusion

This paper implements a convergent harmony reinforcement learning framework in lane change decision-making, introduces a P-Critic network and a harmony guidance module, and promotes the harmony of the policy through disharmony criticism and IISG-based overall optimization. The P-Critic network evaluates the benefits of the entire surrounding traffic by considering the safety, comfort, and efficiency of the ego vehicle and surrounding vehicles to improve the agent's consideration of traffic harmony when making lane-change decisions. The role of the harmony guidance module is to further enhance the focus on traffic safety, it monitors the actions provided by the actor network, marks dangerous actions, and replaces them with IISG-based overall safety actions, which will improve the ability of collision-avoidance of the agent, enabling a more harmonious policy in an overall safe way.

Both the ablation experiments and the cross-comparison experiments show that CHRL has improved the harmony of the ego vehicle's lane-change actions and the surrounding traffic. Specifically, the contributions can be classified as follows:

1) Harmony in Ego Vehicle:

CHRL lowers the collision rate to 0% while showing higher speeds and shorter lane-change times. This indicates that CHRL strikes a better balance between safety, comfort, and efficiency.

2) Harmony in the Surrounding Traffic:

CHRL makes its surrounding traffic the smallest lane-change speed deviation, which is significantly lower than that of human drivers by 75.4%. This demonstrates a higher overall safety. Additionally, the lane-change trajectory of CHRL is gentler and the longitudinal acceleration is lower, which reduces interference of surrounding vehicles and improves traffic harmony.

In future work, we aim to enhance CHRL's adaptability to target speeds during training and expand its application scenarios to urban or low-speed scenarios. Additionally, we plan to further enhance the transportability of the CHRL, to develop a more compatible frame that can be integrated conveniently with different networks.

References

[1] H. Jula, E. B. Kosmatopoulos, and P. A. Ioannou, "Collision avoidance analysis for lane changing and merging," *IEEE Trans. Veh. Technol.*, vol. 49, no. 6, pp. 2295–2308, Jun. 2000.

[2] H. Deng, Y. Zhao, Q. Wang, and A.-T. Nguyen, "Deep reinforcement learning based decision-making strategy of autonomous vehicle in highway uncertain driving environments," *Automot. Innov.*, vol. 6, no. 3, pp. 438–452, Aug. 2023.

[3] Z. Li, J. Hu, B. Leng, L. Xiong, and Z. Fu, "An integrated of decision making and motion planning framework for enhanced oscillation-free capability," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 6, pp. 5718–5732, Jun. 2024.

[4] X. He, H. Yang, Z. Hu, and C. Lv, "Robust lane change decision making for autonomous vehicles: An observation adversarial reinforcement learning approach," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 1, pp. 184–193, Jan. 2023.
[5] J. Yao, G. Chen, and Z. Gao, "Target vehicle selection algorithm for adaptive cruise control based on lane-changing intention of preceding vehicle," *Chin. J. Mech. Eng.*, vol. 34, no. 1, pp. 1–18, Dec. 2021.
[6] A. J. Khattak, N. Ahmad, B. Wali, and E. Dumbaugh, "A taxonomy of driving errors and violations: Evidence from the naturalistic driving study," *Accident Anal. Prevention*, vol. 151, Mar. 2021, Art. no. 105873.
[7] B. Leng et al., "Multi-mode evasion assistance control method for intelligent distributed-drive electric vehicle considering human driver's reaction," *Chin. J. Mech. Eng.*, vol. 38, no. 1, p. 102, Jun. 2025.
[8] A. Irshayyid, J. Chen, and G. Xiong, "A review on reinforcement learning-based highway autonomous vehicle control," *Green Energy Intell. Transp.*, vol. 3, no. 4, Aug. 2024, Art. no. 100156.
[9] Y. Yang, N. M. Negash, and J. Yang, "Recent advances in interactive driving of autonomous vehicles: Comprehensive review of approaches," *Automot. Innov.*, vol. 8, no. 2, pp. 304–334, May 2025.
[10] H. Li, G. Yu, P. Chen, Y. Li, and Q. Xia, "A human-like parking trajectory planning approach for autonomous vehicle load tasks in mining site," *Automot. Innov.*, vol. 2025, pp. 1–22, Jul. 2025.
[11] X. Wang, Q. Zhou, M. Quddus, T. Fan, and S. Fang, "Speed, speed variation and crash relationships for urban arterials," *Accident Anal. Prevention*, vol. 113, pp. 236–243, Apr. 2018.
[12] J. Nilsson, M. Brännström, E. Coelingh, and J. Fredriksson, "Lane change maneuvers for automated vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 5, pp. 1087–1096, May 2017.
[13] Y. Chen, H. Yu, J. Zhang, and D. Cao, "Lane-exchanging driving strategy for autonomous vehicle via trajectory prediction and model predictive control," *Chin. J. Mech. Eng.*, vol. 35, no. 1, p. 71, Dec. 2022.
[14] T. Wei and C. Liu, "Safe control with neural network dynamic models," in *Proc. Learn. Dyn. Control Conf.*, 2021, pp. 739–750.
[15] G. Li, W. Zhou, S. Lin, S. Li, and X. Qu, "On-ramp merging for highway autonomous driving: An application of a new safety indicator in deep reinforcement learning," *Automot. Innov.*, vol. 6, no. 3, pp. 453–465, Aug. 2023.
[16] L. Zhang, Q. Zhang, L. Shen, B. Yuan, X. Wang, and D. Tao, "Evaluating model-free reinforcement learning toward safety-critical tasks," in *Proc. AAAI Conf. Artif. Intell.*, 2023, vol. 37, no. 12, pp. 15313–15321.
[17] H. Lu, C. Lu, Y. Yu, G. Xiong, and J. Gong, "Autonomous overtaking for intelligent vehicles considering social preference based on hierarchical reinforcement learning," *Automot. Innov.*, vol. 5, no. 2, pp. 195–208, Apr. 2022.
[18] H. Shao, M. Zhang, T. Feng, and Y. Dong, "A discretionary lane-changing decision-making mechanism incorporating drivers' heterogeneity: A signalling game-based approach," *J. Adv. Transp.*, vol. 2020, pp. 1–16, Jan. 2020.
[19] 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.
[20] P. Hang, C. Lv, C. Huang, Y. Xing, and Z. Hu, "Cooperative decision making of connected automated vehicles at multi-lane merging zone: A coalitional game approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 4, pp. 3829–3841, Apr. 2022.
[21] R. Yang, Z. Li, B. Leng, and L. Xiong, "Safe reinforcement learning for autonomous vehicles to make lane-change decisions: Constraint based on incomplete information game theory," in *Proc. 7th CAA Int. Conf. Veh. Control Intell. (CVCI)*, Oct. 2023, pp. 1–6.
[22] S. Karimi, A. Karimi, and A. Vahidi, "Level-*K* reasoning, deep reinforcement learning, and Monte Carlo decision process for fast and safe automated lane change and speed management," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 6, pp. 3556–3571, Jun. 2023.
[23] G. Wang, J. Hu, Z. Li, and L. Li, "Harmonious lane changing via deep reinforcement learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 5, pp. 4642–4650, May 2022.
[24] S. Mysore, G. Cheng, Y. Zhao, K. Saenko, and M. Wu, "Multi-critic actor learning: Teaching RL policies to act with style," in *Proc. Int. Conf. Learn. Represent.*, 2022, pp. 1–23.
[25] H. Bharadhwaj, A. Kumar, N. Rhinehart, S. Levine, F. Shkurti, and A. Garg, "Conservative safety critics for exploration," 2020, *arXiv:2010.14497*.
[26] L. Wen, J. Duan, S. E. Li, S. Xu, and H. Peng, "Safe reinforcement learning for autonomous vehicles through parallel constrained policy optimization," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2020, pp. 1–7.
[27] H. Ma, C. Liu, S. E. Li, S. Zheng, and J. Chen, "Joint synthesis of safety certificate and safe control policy using constrained reinforcement learning," in *Proc. 4th Annu. Learn. Dyn. Control Conf.*, vol. 168, 2022, pp. 97–109.
[28] W. Zhao, T. He, R. Chen, T. Wei, and C. Liu, "State-wise safe reinforcement learning: A survey," 2023, *arXiv:2302.03122*.
[29] R. S. Sutton et al., *Introduction to Reinforcement Learning*, vol. 135. Cambridge, MA, USA: MIT Press, 1998.
[30] M. Werling, J. Ziegler, S. Kammel, and S. Thrun, "Optimal trajectory generation for dynamic street scenarios in a Frenet frame," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2010, pp. 987–993.
[31] S. Jeong and J. Choi, "Differentiable moving horizon estimation for vehicle kinematics via learning covariance matrices," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 9, pp. 5955–5969, Sep. 2024.
[32] T. Nishi, P. Doshi, and D. Prokhorov, "Merging in congested freeway traffic using multipolicy decision making and passive actor-critic learning," *IEEE Trans. Intell. Vehicles*, vol. 4, no. 2, pp. 287–297, Jun. 2019.
[33] H. Wang, W. Wang, S. Yuan, X. Li, and L. Sun, "On social interactions of merging behaviors at highway on-ramps in congested traffic," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 11237–11248, Aug. 2022.
[34] D. Zeng et al., "A novel robust lane change trajectory planning method for autonomous vehicle," in *Proc. IEEE Intell. Vehicles Symp.*, Paris, France, Jun. 2019, pp. 486–493.
[35] P. K. Dutta, *Strategies and Games: Theory and Practice*. Cambridge, MA, USA: MIT Press, 1999.
[36] T. Roughgarden, "Algorithmic game theory," *Commun. ACM*, vol. 53, no. 7, pp. 78–86, 2010.
[37] F. Zong, M. Wang, J. Tang, and M. Zeng, "Modeling AVs & RVs' car-following behavior by considering impacts of multiple surrounding vehicles and driving characteristics," *Phys. A, Stat. Mech. Appl.*, vol. 589, Mar. 2022, Art. no. 126625.
[38] K. Hao et al., "Adversarial safety-critical scenario generation using naturalistic human driving priors," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 9, pp. 5392–5406, Sep. 2024.
[39] A. Kesting, M. Treiber, and D. Helbing, "General lane-changing model MOBIL for car-following models," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1999, no. 1, pp. 86–94, Jan. 2007.
[40] T. Haarnoja et al., "Soft actor-critic algorithms and applications," 2018, *arXiv:1812.05905*.
[41] H. V. Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 30, 2016, pp. 2094–2100.
[42] Z. Li, L. Xiong, B. Leng, P. Xu, and Z. Fu, "Safe reinforcement learning of lane change decision making with risk-fused constraint," in *Proc. IEEE 26th Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2023, pp. 1313–1319.

Ruolin Yang received the B.E. degree in vehicle engineering from Chang'an University, Xi'an, China, in 2021. She is currently pursuing the M.Eng. degree with the School of Automotive Studies, Tongji University, Shanghai, China. Her research interests include decision-making, motion-planning, and reinforcement learning methods in autonomous vehicles.

Zhuoren Li (Graduate Student Member, IEEE) received the B.E. degree in engineering mechanics from Tongji University, Shanghai, China, in 2019, where he is currently pursuing the Ph.D. degree with the School of Automotive Studies. His current research interests include interaction decision-making, motion planning, and the safe reinforcement learning of autonomous vehicles.

Bo Leng received the Ph.D. degree in vehicle engineering from Tongji University, Shanghai, China. He is currently an Associate Professor with the School of Automotive Studies, Tongji University. His current research interests include the dynamic control of distributed drive electric vehicles and motion planning and control of intelligent vehicles. He has won the First Prize in China Automobile Industry Technology Invention Award and the First Prize in Shanghai Science and Technology Progress Awards in 2020 and 2022. He has been selected into the Young Elite Scientists Sponsorship Program of China Association for Science and Technology in 2022.

Lu Xiong received the Ph.D. degree in vehicle engineering from Tongji University, Shanghai, China, in 2005. He is currently the Vice President and a Professor with the School of Automotive Studies, Tongji University. His current research interests include the dynamic control of distributed drive electric vehicles, motion planning and control of intelligent vehicles, and all-terrain vehicles. He won the First Prize in Shanghai Science and Technology Progress Awards in 2013, 2020, and 2022. He was a recipient of the National Science Fund for Distinguished Young Scholars.

Xin Xia received the B.E. degree in vehicle engineering from the School of Mechanical and Automotive Studies, South China University of Technology, Guangzhou, China, in 2014, and the Ph.D. degree in vehicle engineering from the School of Automotive Studies, Tongji University, Shanghai, China, in 2019. He was a Post-Doctoral Fellow associated with Dr. Amir Khajepour at the Department of Mechanical and Mechatronics Engineering, University of Waterloo, Waterloo, ON, Canada, from January 2020 to March 2021. After that, he was an Assistant Professional Researcher with the Department of Civil and Environmental Engineering, University of California at Los Angeles, Los Angeles, CA, USA, from March 2021 to August 2024. He is currently an Assistant Professor with the Department of Mechanical Engineering, University of Michigan–Dearborn, Dearborn, MI, USA. His research interests include state estimation, cooperative localization, cooperative perception, and dynamics control of the autonomous vehicle.