Commun Transp Res

Research Article

Safe and Interpretable Vehicle Platoon Control via Physics-Enhanced Residual Learning

Haigen Min[1,2,✉], Zihao Mao[1], Xia Wu[1,2], Ruichen Zhang[1], Ye Li[3], Xiangmo Zhao[1,2], Yang Liu[4]

Cite this article: https://doi.org/10.26599/COMMTR.2026.9640073

ABSTRACT: Vehicle platoon control in complex dynamic environments necessitates a rigorous balance between physical interpretability safety constraints and nonlinear compensation. This paper presents PERL, a Physics-Enhanced Residual Learning framework that hierarchically integrates Distributed Model Predictive Control (DMPC) with Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm. To address the pathological residual action saturation. we propose a residual-magnitude-based action regularization mechanism. This mechanism penalizes large-amplitude residuals to stabilize the policy update and constrain the RL module within a safe, small-amplitude regime, ensuring effective compensation for unmodeled dynamics beyond the physical baseline's fidelity. Additionally, a spatiotemporal smooth handover and dynamic fusion strategy significantly enhances edge-case safety. Utilizing safety-boundary constraints and dual soft-filtering, the strategy regulates control authority to mitigate violations of Rational Driving Constraints by RL policies while preventing abrupt control mutations. Experiments across four representative scenarios show that PERL converges rapidly and consistently outperforms MPC, PPO, and TD3 in safety, comfort, and platoon stability. Specifically, PERL reduces inverse time-to-collision by 20.3% relative to PPO and improves ride comfort by over 60% compared with TD3. Ultimately, the framework attains the lowest cumulative damping ratio, demonstrating superior string stability and robust generalization capability in real dynamic traffic environments.

KEYWORDS: Connected and Automated Vehicles, Deep Reinforcement Learning, Cooperative Control, String Stability.

1 Introduction

With the increase of global urbanization, issues including traffic congestion, energy consumption, and road traffic safety are becoming increasingly severe, and have emerged as major social challenges. As a core component of intelligent transportation systems, platooning of Connected and Automated Vehicles (CAVs) has attracted considerable attention from academia and industry because it has the potential to increase road capacity, improve energy efficiency and enhance driving safety. Through vehicle-to-everything (V2X) communication technologies, multiple CAVs travel at close following distances, thereby reducing aerodynamic drag and fuel consumption and then forming stable and efficient traffic flows. However, a vehicle platoon is inherently a complex multi-agent system with strong coupling, dynamic behavior, and uncertainties. Its controller design faces multiple severe challenges, including dynamic coupling between vehicles, engine delays, model uncertainties, and external disturbances. Therefore, developing a reliable, robust, and scalable controller is essential for translating this technology from theory to practice.

Existing vehicle platoon control methods can be categorized into two approaches: physics-based approaches and data-driven approaches. Classical physics model designs have been extensively discussed in the existing literature. These methods make use of a strong theoretical foundation, which provides a solid understanding of vehicle dynamics and offer universally applicable modeling, controlling and analysis solutions. For heterogeneous vehicle platoons, Zheng et al. (2016) designed a distributed model predictive controller (DMPC) based on single-vehicle dynamics. By considering differences in physical parameters such as vehicle mass and engine responses, they used predecessor-following communication topology. The controller then predicted future states by using the physical model and optimized control inputs to achieve cooperative control. Based on traditional longitudinal vehicle dynamics, Luo et al. (2010) designed a multi-objective model predictive controller (MPC) for adaptive cruise control (ACC). Its objective function takes comfort, fuel economy, safety and car-following performance into consideration, illustrating the typical application of physical models in multi-objective optimization control. Memon et al. (2012) proposed combining the ACC model from the California PATH program (Shladover et al., 2012) with a Proportional-Integral-Derivative (PID) controller to form a cascade structure. This approach provided analytical conditions for string stability (Feng et al., 2019) under disturbances and parameter uncertainties, which can maintain uniform speed and safe distances among vehicles in a platoon. Although traditional physical models offer direct insights into systems through physical laws, their simplified structures and excessive abstraction of complex systems often lead physics-based methods to overlook the dynamic variability of vehicle platoon systems and external traffic disturbances. Such over-idealized settings pose challenges to the effective application of physics-based control methods in real-world traffic.

In recent years, data-driven vehicle platoon control methods have attracted widespread attention. It can directly study control strategies from data, so these methods are regarded as an important trend and

Just Accepted

[1] School of Information Engineering, Chang'an University, Xi'an 710064, China. [2] Collaborative Innovation Center for Western China Traffic Safety and Intelligent Cooperative Control, Xi'an 710018, China. [3] School of Traffic and Transportation Engineering, Central South University, Changsha 410083, China. [4] School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798, Singapore.

✉ Corresponding author. E-mail: **hgmin@chd.edu.cn**

Received: July 15, 2026; Revised: September 16, 2026; Accepted: November 17, 2026

© The Author(s) 2026. This is an open access article under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0, http://creativecommons.org/licenses/by/4.0/).

effective means for handling model uncertainty and complex dynamic environments in vehicle platoon systems. This approach utilizes driving trajectory data to directly design platoon control strategies, enabling them not only captures nonlinear relationships within complex systems but also handles various real-world traffic scenarios effectively. Yen et al., (2022) designed an innovative platoon control learning framework that reduces traffic delays and fuel consumption through four basic operations (e.g., splitting, accelerating, decelerating, and doing nothing). By integrating reinforcement learning with neural networks, it can simulate the nonlinear relationships between inputs and outputs, thereby adapting to complex applications. Shi et al., (2023) proposed a cooperative longitudinal control strategy of platoon based on deep reinforcement learning (DRL) for mixed traffic, dividing mixed platoon into multiple subsystems to achieve efficient centralized cooperative control. Considering that the platoon control is a high-dimensional, continuous state and action space, learning-based methods can learn complex control policies through continuous interaction with the environment. Wang et al., (2023) introduced an attention based data fusion car-following model, which constructs multi-source training dataset by fusing field data with simulation data. They also employ a sequence-to-sequence model with an attention mechanism for learning, and incorporates a safety indicator to prevent unrealistic outputs and rear-end collisions. Recently, existing researches have introduced prior knowledge into reinforcement learning to control vehicles platoons in responding to cut-in maneuvers by other vehicles (Yang et al., 2026), and using naturalistic trajectories data to train a safe speed guidance policy via offline reinforcement learning (Shi et al., 2026); reinforced imitative trajectory planning has also shown potential for urban automated driving (Zeng et al., 2026). Furthermore, the massive volumes of data generated by automated vehicles during monitoring, reporting and remote driving have further motivated the development of efficient data learning frameworks (Ponso et al., 2026). However, due to the lack of interpretability and transparency in its black-box computation, understanding the control process and dynamic mechanisms of learning-based vehicle platoons remain many challenges. In summary, for complex vehicle platoon control problems, both traditional physics-based models and learning-based approaches are very difficult to provide the most appropriate control strategies. Therefore, combining the strengths of physical models and machine learning for vehicle platoon control has become an active area of research.

In hybrid vehicle platoon control frameworks that combine data-driven models with physical models, the data-driven modules often show signs of action saturation. This phenomenon behaves as policy degradation to the boundaries of a predefined action space, whatever how environmental state change, it will output extreme control commands. In hybrid architectures, such saturation behavior is particularly pronounced, and its primary cause can be attributed to the overestimation bias inherent in off-policy RL algorithms, which drives the policy network to exploit the action boundaries rather than learning state-dependent fine-grained compensation. Ma et al., (2026) systematically identified the structural scale mismatch in residual RL as a key bottleneck, where the residual contributions are dominated by the base action, leading to ineffective value learning. To alleviate such saturation and stability issues, recent studies have explored bounded residual reinforcement learning (Yan et al., 2025) and support-preserving action rectification methods (Zhao et al., 2026). These methods take basic strategies as reference, reformulating global learning as a local residual refinement, thereby reducing the search space and preventing boundary collapse. Additionally, constraint-aware regulators that scale agent actions based on predicted constraint violation, maintaining exploration through the smooth modulation rather than directly overriding the policy. However, these methods predominantly focus on robot motion or manipulation domains, and their applicability to vehicle platoon control remains scarcely investigated, where safety-critical constraints and multi-vehicle coordination also bring additional challenges. Furthermore, most existing approaches correct actions after saturation occurs, rather than proactively introducing regularization mechanisms during training that reconcile the physical prior with the residual learning objective. Therefore, to build a more robust and reliable intelligent platoon control systems, it is necessary to investigate the conditions under which residual action saturation occurs in the hybrid control frameworks and to design targeted regularization strategies. These strategies should stabilize policy learning while preserving the compensation capability of the data-driven module.

Vehicle Platoon control systems using DRL suffer from a lack of interpretability in their decision-making process due to nonlinear black-box computations, in turn create potential safety hazards. Recent study also emphasized risk perception and quantitative evaluation for intelligent vehicles (Xiao et al., 2026), and has extended safe reinforcement learning to lane-changing decisions by using deep Q-networks with prioritized experience replay (Ran et al., 2025). Zhang and Yan (2022) pointed out that although DRL performs well on average behavior, under safety-critical cases, neural networks often exhibit unexpected abrupt changes in acceleration because they struggle to strictly satisfy physical hard constraints as traditional models (such as IDM) do. To solve this problem, they design safety constraints based on Deceleration Rate to Avoid a Crash (DRAC) and physical constraints to limit impact intensity, specifying acceleration when constraints conflict, thereby effectively preventing collisions during training. Zhang et al., (2021) further found that in complex dynamic environment, DRL agents might enter dangerous areas while pursuing accumulated rewards. For this purpose, they introduce Lyapunov function to ensure the consistent ultimate bounded stability of the system, enabling the vehicle to automatically return to the safe zone within a limited time when unexpected disturbance cause it to enter a dangerous area. This method predict the risk quantification for collision probability, significantly decrease the collision rate caused by uncertain actions during training. On this basis, Bai et al., (2024) theoretically revealed inherent safety deficiencies of DRL algorithms in specific scenarios by calculating the partial derivatives of the Actor network with respect to spacing, velocity, and relative velocity. Their research demonstrated that when the partial derivative is less than zero at small gap, the model will face severe collision risks. This work proved that native DRL algorithms cannot spontaneously satisfy Rational Driving Constraints (RDC) and the velocity response time (VRT) metric is proposed to discriminate and integrate the Intelligent Driver Model (IDM) and Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm into a hybrid control model. However, existing research mostly rely on single metrics as criterion for determination, making it difficult to comprehensively capture the

Just Accepted

complex abnormal output patterns of TD3 policies in multi-dimensional, dynamic traffic scenarios thus limits the upper bound of safety performance. Therefore, it is important for building highly reliable and interpretable intelligent platoon control systems to explore the inducing factors and external boundary conditions for abnormal actions in TD3 algorithm.

To solve the above issues, this paper introduces a Physics-Enhanced Residual Learning framework for vehicle platoon control. This method perfectly combined physical models with DRL, and effectively alleviate the boundary saturation issue of residual actions in cooperative control, while avoiding the limitations. As a physical model, distributed MPC provides interpretable baseline control outputs based on prior knowledge and physical constraints, while TD3 serves as a residual module to capture the parts that the physical model cannot accurately represent, thereby improving the model's output. This approach not only combines the interpretability of physical models, avoiding the shortcomings of DRL when RDC conditions are not satisfied, but also make accurate control decisions in complex scenarios and traffic disturbances, speeds up the convergence of DRL, and its distributed design enhances system robustness. The main contributions of this paper are as follows.

Proposing a physics-enhanced residual learning framework for vehicle platoon control that combines the strengths of physical models and data-driven methods. The framework addresses two limitations, traditional physical models struggle to capture nonlinear system dynamics and external disturbances, while learning-based methods lack interpretability. To address these limitations, use DMPC as the basic controller. It uses prior knowledge of vehicle dynamics and physical constraints to provide interpretable baseline control inputs. Introduce TD3 as a residual compensation module to learn corrections for external traffic disturbances and nonlinear dynamics that DMPC does not fully capture. Together, the two modules combine the interpretability of the physical model with the adaptability of residual learning to provide accurate and interpretable control for vehicle platoons.

We propose a reward-shaping mechanism for residual action regularization to prevent policy saturation and stabilize the hybrid learning process. Although DMPC provides physically interpretable baseline actions, overestimated Q-values can cause the TD3 agent to produce saturated residual actions rather than learn effective corrections for nonlinear dynamics. To address this problem, we add the penalty on the squared magnitude of residual actions to the reward. Unlike incentives that encourage exploration, this penalty favors smaller residual actions and guides the agent toward fine adjustments around the baseline actions within safety limits. This reduces control chattering near the action bounds and helps the residual module work more consistently with the base controller. As a result, the mechanism improves convergence during training and increases the robustness of the final control policy.

We propose a smooth takeover and dynamic fusion mechanism based on safety-boundary constraints to support safe and adaptive platoon control throughout operation. In response to the safety risks caused by the difficulty of TD3 to meet the RDC conditions, this article is different from the hard-switching limitations of traditional safety monitoring, and innovatively builds a smooth takeover strategy with dual soft filtering features in time and space. We assess safety risks in real time and gradually reduce or increase the weight of the residual action as the risk changes. This allows the physical controller to take over smoothly as risk rises and retains the benefits of residual compensation when safety conditions allow. In this way, the mechanism reduces abrupt changes in control inputs and collision risk while keeping the fusion logic interpretable and the control inputs smooth.

2 Related Work

In the vehicle platoon control field, traditional physics-based control methods have long dominated. These kind of method often models the longitudinal and lateral dynamic equations of vehicles, and use feedback control theory to achieve the stable operation of the platoon. Traditional physical model widely used in the vehicle platoon control field include the Optimal Velocity Model (OVM) (Bando et al., 1995), the Full Velocity Difference (FVD) model (Zheng et al., 2023), the Intelligent Driver Model (IDM) (Treiber et al., 2000), the Newell model (Zheng et al., 2023), and the Adaptive Cruise Control (ACC) model, as well as the Model Predictive Control (MPC) frameworks (Caruntu et al., 2016) which have received wide attention in recent years. This method directly reveals that the characteristics of the dynamic evolution process through physical laws, presenting its high interpretability and rigorous theoretical guarantees. However, physical models often rely on the oversimplification of complex system dynamics. In the actual open traffic environment, vehicle platoon system faces a high level of dynamic variability and complex external traffic interference. This over-idealized modelling hypothesis makes traditional physics-based controllers prone to performance degradation and difficult to cope with unknown dynamic perturbations.

In order to overcome the modeling bottlenecks of traditional physical models in complex environments, in recent years, data-driven approaches, especially DRL, have attracted wide attention in vehicle platoon control. Because there is no need to construct accurate mathematical model in advance, DRL can learn the optimal control policies to adapt to complex non-linear traffic flows through continuous interaction with the environment,showing the strong generalization and adaptability. For example, Sun et al., (2026) proposed the combination of the adaptive forgetting factor with Proximal Policy Optimization (PPO). By identifying the vehicle dynamic parameters in real time and inputting the DRL controllers, they realized the adaptive control of heterogeneous platoons,which is prior to the four baseline methods in reducing spacing errors, stable time, and fuel economy. Chen et al., (2025) introduced a two-layer Multi-Agent Maximum Q-value PPO (MAMQPPO) method, which extends the multi-agent PPO algorithm by introducing the maximum Q-value action selection. This method can be used for the platoon spacing control and discrete command generation, realizing the simultaneous optimization of longitudinal acceleration and the longitudinal control of merging vehicles. However, the learning- based methods are generally subject to their black-box computing nature. The high non-linearity of deep neural networks makes the whole control process lack transparency, leading it is extremely difficult for researchers to clearly explain its internal dynamic policies and decision-making procedure. More importantly, it is difficult for pure data-driven DRL to provide extreme safety constraints and system reliability

Just Accepted

guarantees during the exploration process, which is a fatal shortcoming in the field of autonomous driving and vehicle platooning with extreme high safety requirements.

In view of the limitations of the above two methods, how to take into account the interpretability of physical models and the precision and generalization of data-driven methods has become a hot topic of current research. Kashinath et al., (2021) proposed a Physics-Informed Neural Network (PINN) model. This method achieves the balance between the physical model and the neural network by adjusting the weights of the weight bias of physical model prediction and the weight loss of neural network. Although this method has significant advantages, the deviation correction of the physical model prediction is often inaccurate. It will impair the predictability of PINN in some extent, and it proves to be challenging to adjust the weight of physical model in loss function (Wang et al., 2021; Wang et al., 2022; Krishnapriyan et al., 2021). Zhang et al., (2024) utilized Bayesian optimization to automatically search for hyperparameters in the cost function of a distributed MPC to balance tracking accuracy and energy consumption, but he failed to make full use of the generalization of the learning-based method. How to design a unified framework that not only guarantee the underlying physical safety constraints but also accurately compensate the unmodeled dynamic is still challenging. In view of this, this article proposes a Physics-Enhanced Residual Learning (PERL) framework. This framework takes distributed MPC as a physical basis and generates interpretable basic actions to ensure the safety constraints. At the same time, it introduces the TD3-based DRL module as a residual module to learn the dynamic disturbances that traditional physical model can not accurately be modeled. This architecture not only gives the system strong generalization and high precision but also fundamentally solves the dual dilemma that pure learning methods are difficult to explain and the physical models are difficult to implement.

3 Preliminaries and Method

3.1. Problem Formulation

As illustrated in Fig. 1. consider a vehicle platoon consisting of $N+1$ vehicles. comprising a lead vehicle (indexed as 0) and N following vehicles (indexed as $i \in \{1,2,\dots,N\}$). The control goal of this articles is to enable all following vehicles follow the car in front at a desired inter-vehicle distance under the premise of ensuring safety,while satisfying the requirements of passenger comfort and string stability.



Fig. 1. Schematic diagram of a vehicle platoon.

At a discrete time step k, the state vector of the i-th vehicle is formulated as $x_k^i = [p_k^i, v_k^i, a_k^i]^T$. where, p_k^i, v_k^i, and a_k^i denote the longitudinal position, velocity, and acceleration of the vehicle at time k, respectively. Let Δt represent the discrete sampling interval of the system. This article assumes that the vehicles in the platoon are able to obtain the real-time state information of the front vehicle by Vehicle-to-Vehicle (V2V) communication to achieve cooperative control. In order to balance computing efficiency and the characteristic of actuator lag, this paper adopts a simplified longitudinal model with first-order inertial lag to describe the underlying dynamic behavior of the vehicle. According to the laws of kinematic principles and the first-order lag component, the state transition equations for vehicle i in discrete-time form are expressed as follows.

$$\begin{cases} p_{k+1}^i = p_k^i + v_k^i \Delta t + \frac{1}{2} a_k^i \Delta t^2 \\ v_{k+1}^i = v_k^i + a_k^i \Delta t \\ a_{k+1}^i = (1 - \frac{\Delta t}{\tau}) a_k^i + \frac{\Delta t}{\tau} u_k^i \end{cases} \quad (1)$$

Among them, u_k^i denotes the target acceleration of vehicle i at time step k, $\Delta t = 0.1s$ is the sampling interval, and $\tau = 0.5s$ represents the lag in the vehicle's longitudinal acceleration response. By introducing the first-order lag term in the kinematic model, this model approximates the low-pass response characteristics from the target acceleration to the actual acceleration. This method avoids the complex inverse dynamics calculation requirements, while retaining the sufficient dynamic fidelity for control synthesis.

3.2. Physics-Enhanced Residual Learning Framework

As shown in Fig. 2, the Physics-Enhanced Residual Learning (PERL) framework in this paper is mainly composed of two core modules: distributed model predictive control(DMPC) algorithm based on physical model and the TD3 algorithm based on DRL. In this framework, DMPC, as the underlying controller, calculates the vehicle acceleration that meets the strict safety constraints through rolling optimization based on strict prior physical knowledge and boundary constraints. This baseline action generated by the physical model lays a solid safety baseline for the whole control framework, and also gives the internal interpretability for systems.

Due to traditional physical models are difficult to accurately capture complex nonlinear characteristics, such as engine time lag, mathematical modeling inaccuracies, and stochastic traffic disturbances, the framework integrates TD3 algorithm as a residual compensation module. By using DMPC control output and instantaneous motion state as state input, this module is able to specialize for unmodeled dynamic bias. Its main control logic is to generate residual actions continuously and superimpose them on the baseline actions of physical model output, thereby improve the accuracy and generalization ability through data-driven synergies.

In addition, in order to ensure the effective training of the above-metioned hybrid control architecture, this framework designs a reward shaping policy based on residual action regularization to solve the policy collapse problems which it is easy to show in the reinforcement learning agents. This mechanism directly adds the amplitude constraints of residual action into TD3's optimization goal, and guides the policy network to learn how to compensate the local nonlinear, thereby prevents the agent from blindly outputting saturation action during the exploration process.

Furthermore, to reduce the inherent risk of the pure data-driven algorithms easily violating RDC in unknown traffic scenarios, the framework introduces a smooth takeover and dynamic fusion policy based on safety boundary constraints. The mechanism is different from the traditional hard switching logic. It adopts the spatiotemporal soft-gating strategy to adjust the weight of TD3 residual action by evaluating the safety of systems continuously. Under normal safety conditions, the controller can achieve the seamless integration of baseline actions and residual action, when system state is close to physical safety boundary, the residual weight will smoothly decay

Just Accepted

until the DMPC completely takes over the control. In this way, it can avoid the system instability caused by abrupt control change and also strengthen the safety of the framework.

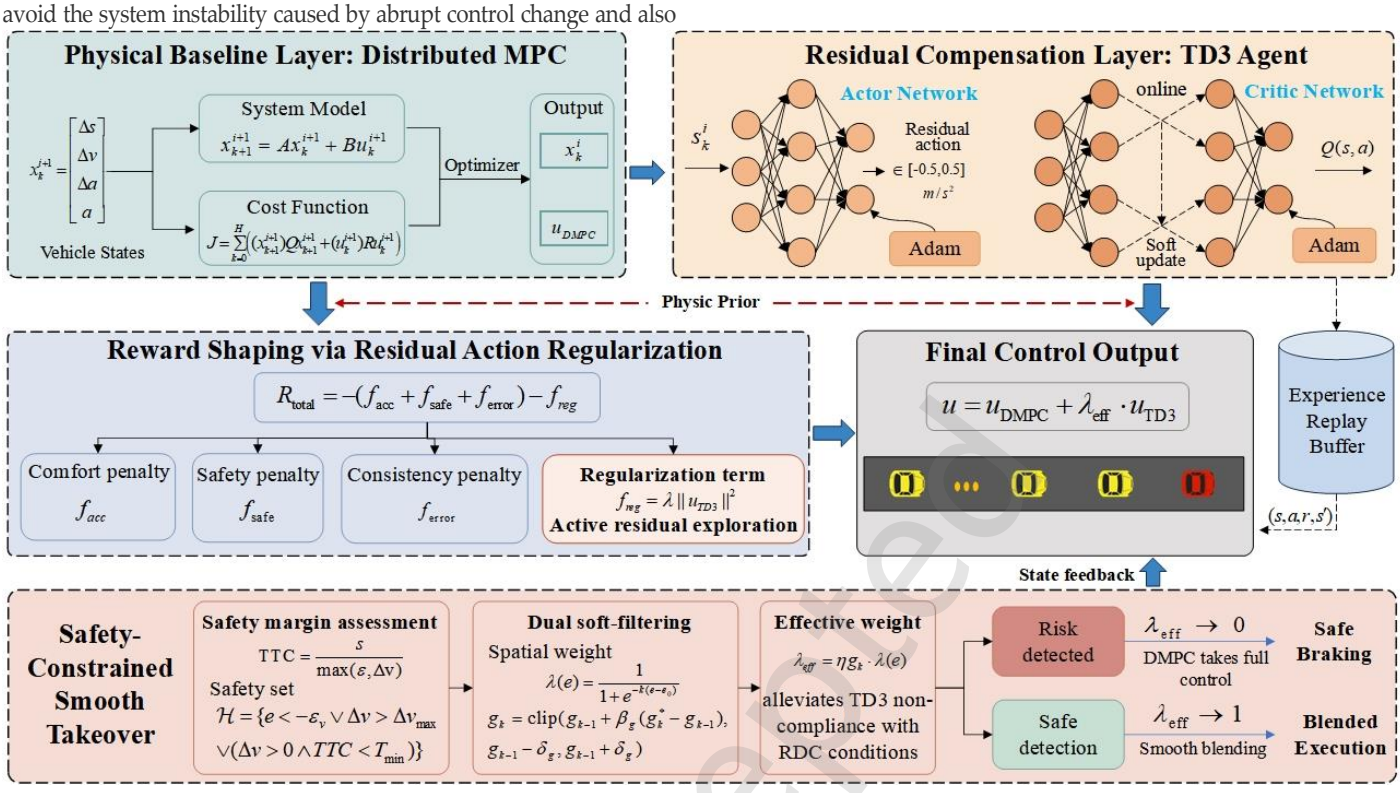


Fig. 2. Diagram of the Physics-Enhanced Residual Learning Framework.

3.2.1 State Space

Reasonable state space design is the premise for ensuring the efficient convergence and precise decision-making of DRL algorithms. In the PERL framework of this article, in order to realize the deep fusion of the low-level physical prior knowledge and high-level data-driven algorithms, we specially reconstruct the state space of the TD3 agent. Specifically, the DMPC controller first performs receding horizon optimization based on the platoon state x_k^i at time k, utilizing a state update model and cost function to solve for the desired control input $u_{DMPC,k}^i$ as the baseline action. Here, x_k^i is defined as:

$$x_k^i = [\Delta s_k^i, \Delta v_k^i, \Delta a_k^i, a_k^i]^T \tag{2}$$

where Δs_k^i, Δv_k^i, and Δa_k^i denote the spacing error, relative velocity, and relative acceleration between vehicle i and its predecessor at time step k, respectively. a_k^i represents the acceleration of the ego vehicle. The DMPC state transition model is then defined as:

$$x_{k+1}^{i+1} = Ax_k^{i+1} + Bu_k^{i+1} \tag{3}$$

where matrices A and B are defined as follows:

$$A = \begin{bmatrix} 1 & \Delta t & 0 & -\Delta t \\ 0 & 1 & \Delta t & 0 \\ 0 & 0 & 1 & \frac{\Delta t}{\tau} \\ 0 & 0 & 0 & 1-\frac{\Delta t}{\tau} \end{bmatrix}, B = \begin{bmatrix} 0 \\ 0 \\ -\frac{\Delta t}{\tau} \\ \frac{\Delta t}{\tau} \end{bmatrix} \tag{4}$$

Eq.(4) simplify Euler discreteization is used for the state transfer model. This treatment implicitly assumes that the uniform speed assumption is that the front vehicle's acceleration remains constant and set to zero in the predicted time domain. This simplification is to create the mismatch between physical model and environment, because the actual vehicle dynamics used in the simulated environment follow the first-order inertial lag model in Eq. (1). Such mismatch is a deliberate design choice, not a defect. The residual model is specially designed to capture the differences between physical model and true environment. This design prompts the TD3 agent learning to compensate for unmodeled dynamics, including the inertial lag, external disturbances, and the accumulation of linearization errors, which finally confirms the core concept of physics-enhanced residual learning.

The cost function is formulated as a multi-objective optimization problem to balance spacing error, velocity error, control effort, and its rate of change, all while satisfying physical constraints:

$$J = \sum_{k=0}^{H} \left((x_{k+1}^{i+1})^T Q x_{k+1}^{i+1} + (u_k^{i+1})^T R u_k^{i+1} \right) \tag{5}$$

subject to the hard constraints below, which capture realistic vehicle capabilities and collision-avoidance requirements:

$$u_{min} \le u_k^{i+1} \le u_{max} \tag{6}$$

$$\Delta s_{min} \le \Delta s_k^{i+1} \le \Delta s_{max} \tag{7}$$

where $u_{min} = -3 m/s^2$ and $u_{max} = 3m/s^2$ are the acceleration bounds, $\Delta s_{min} = -1.5m$ and $\Delta s_{max} = 15m$ are the minimum and maximum allowable spacing errors to prevent rear-end collisions and platoon disengagement, respectively. The prediction horizon is set to $H = 5$, with weighting matrices assigned as $Q = diag(2,2,0.1,0)$, and R = [0.1].

This baseline action is then appended to the agent's observation vector to construct the final state input for the TD3 module. For the i-th following vehicle, the MDP state s_k^i at step k is expressed as:

$$s_k^i = \{\Delta s_k^i, \Delta v_k^i, \Delta a_k^i, a_k^i, u_{DMPC,k}^i\} \tag{8}$$

The design of introducing the action of DMPC output into the state space enables the TD3 agents not only to perceive the tracking errors, but also to directly capture the control intention of the underlying physical model, which is very crucial for learning residual action effectively. When TD3 is learning residual actions, the training goal of

Just Accepted

the agent is to generate corrections to the basic movements, instead learning the whole control strategies. By directly expose the basic movements to the state, the agent can capture the amount of control that is currently be added. This enables residual action to be learned as a function of system state and basic movement, and make the agent understand and explain the action of underlying physical model, rather than treating it as a black box. In practical applications, because the agent only needs to capture the deviation of the basic movements, and it does not need to study the whole control strategies, so the result is a significant reduction in the search space. In addition, the design of this state is very helpful for the integration between DMPC and TD3 model, making the residual model offset the error caused by simplifying the DMPC model in a targeted manner. Because DMPC are responsible for outputting the baseline actions, TD3 agent can focus on complex nonlinearity and inaccurate modeling errors caused by unknown environmental perturbations. In this way, this framework is able to output the more accurate residual compensation actions $u^i_{TD3,k}$.

3.2.2 Action Space

In the MDP, the action space defines the set of valid control instructions that can be executed by the agent in each discrete step. In this study, the control variable is the longitudinal acceleration u^i_k of the automated vehicle, which resides in a continuous action space. In order to ensure the comfort of the vehicle in the actual driving process and the real vehicle dynamic response limits, according to the guidelines proposed by (Zhu et al., 2020), the total execution acceleration must be strictly constrained within the physical boundaries of $[-3,3]\,\mathrm{m/s^2}$.

Because this article adopts the PERL framework, the final execution action is composed of the DMPC-generated baseline action $u^i_{DMPC,k}$ and the TD3-generated residual action $u^i_{TD3,k}$. In this framework, DMPC provides the main control volume to meet the basic safety and tracking performance. While TD3 algorithms are served as a residual module, it is specially used for compensating the unmodeled dynamic and minor deviations that the physical model fails to accurately capture. In order to prevent the instability caused by abrupt command change during DRL exploration, and to ensure that the total action does not break through the absolute physical limitations, we need to strictly limit the output from the agent. Therefore, this paper will restrict the TD3 agent's action space to a narrow continuous interval: $u^i_{TD3,k} \in [-0.5,0.5]\,\mathrm{m/s^2}$. This limited residual action space design not only provides the ability for system to high-precision fine-tuning for unknown disturbances, but also minimized the risk of destroying underlying physical safety constraints to the biggest extent.

3.2.3 Reward Function

In DRL, reword function can directly decides the optimization objectives of agent and upper limits of the policy convergence. To guide TD3 agent from study to safety, stable and high-precision vehicle platoon coordination strategies, this article first designs a multi-dimensional base reward function R_{base}. This base reward function comprehensively considers the three core indicators, including ride comfort, strict safety, and platoon consistency, and applies the multi-dimensional penalty to regulate the agent's exploration spaces.

Comfort Penalty: we need to constraint the drastic changes in the longitudinal dynamics of the vehicle in order to restraint velocity oscillations and enhance the passenger experience. A jerk penalty is introduced to limit the sharp fluctuations of acceleration and force the agent to output continuous and smooth control commands:

$$f_{acc} = \alpha_{acc}(a^i_k - a^i_{k-1})^2 \tag{9}$$

where $\alpha_{acc} = 2$ is the jerk penalty coefficient.

Safety Penalty: When the car collides with the preceding vehicle, a penalty is imposed, because we need collision data to help agent study when training. And during the simulation training process, the distance between vehicles might be negative. When this situation occurs and the value is smaller, it demonstrated that the collision is more serious, at this time, the penalty is severer. In other cases, the reward is zero. The penalty is formulated as:

$$f_{safe} = \begin{cases} -10 + \alpha_{safe}(\Delta s^i_k), & \Delta s^i_k < 0 \\ 0, & otherwise \end{cases} \tag{10}$$

where $\alpha_{safe} = 5$ is the collision penalty coefficient.

Consistency Penalty: through the punishment of spacing penalties, the agent can maintain the precise following distance from the preceding vehicle in details. If the distance is too close or too far, it will be punished. The punishment for velocity and acceleration error allows agent to match the state of preceding vehicle, so that it can make sure the state of platoon is consistent and reduce the velocity oscillations. Through three penalties can make agent learn how to follow the front vehicle and maintain the consistency of the vehicle platoon.

$$f_{error1} = \alpha_{error1}(\Delta s^i_k)^2 \tag{11}$$

$$f_{error2} = \alpha_{error2}(\Delta v^i_k)^2 = \alpha_{error2}(v^{i-1}_k - v^i_k)^2 \tag{12}$$

$$f_{error3} = \alpha_{error3}(\Delta a^i_k)^2 = \alpha_{error3}(a^{i-1}_k - a^i_k)^2 \tag{13}$$

$$f_{error} = f_{error1} + f_{error2} + f_{error3} \tag{14}$$

where $\alpha_{error1} = 4$, $\alpha_{error2} = 0.5$, and $\alpha_{error3} = 0.1$ are the penalty coefficients for spacing, velocity, and acceleration errors, respectively. The total base reward R_{base} is defined as the negative sum of these individual penalties:

$$R_{base} = -(f_{acc} + f_{safe} + f_{error}) \tag{15}$$

3.3. Residual Action Regularization Mechanism

In the proposed PERL framework, the DMPC provides a high-level reliable baseline action for system. However, in the early stage of joint training in this framework, TD3 often encounters a policy collapse issue that manifests as action saturation. If it isn't properly restrained, the strategy of agent will often quickly push the residual action to the predefined interval boundaries. This saturation often caused by the inflated Q-value estimation bias from DRL, which make agent outputs the extreme instructions, but cannot generate the small compensation state-dependent. That kind of saturation is behavior not only introduces the hazardous control chattering but also prevent the residual module' meaningful nonlinear correction to unmodeled dynamics.

In order to alleviate this saturation problem and stabilize the training process, we introduce the residual action regularization mechanism. Unlike the intrinsic exploration incentives, this mechanism plays the role in a prior shrinkage operator. By penalizing the quadratic magnitude of the residual action, the optimization objective clearly restricts the agent strategy to stay within the small-amplitude range near the DMPC reference line. The modified total reward function is reformulated as:

$$R_{total} = -(f_{acc} + f_{safe} + f_{error}) - \lambda\|u_{TD3}\|^2 \tag{16}$$

where $\lambda = 190$ is the regularization coefficient. In order to find this optimal values, we conducted many experiments within $\lambda \in [100,250]$. By comparing the action from the action output of the first following vehicle, we can evaluate whether the choice of λ is reasonable. When λ is too small, for example, when $\lambda = 100$ and 170,

Just Accepted

the residual action will lack sufficient restraint. This will cause unnecessary interference to the baseline action of DMPC output and cause oscillation in the compensation action. On the contrary, when λ is too large, for example, when $\lambda = 210$ and 250, the residual action will become overly constrained. This will lead to insufficient compensation during braking, cannot learn residual action effectively and emerge sharp mutations in the recovery stage. It has be verified by many experiments that when $\lambda = 190$, the effect is the best. At this time, the residual action smoothly converges to zero before the event occurs, providing the small action compensation during emergency braking, and there is no shaking in the recovery process. Finally, while suppressing the saturation action, it keeps the small action compensation for residual module. The penalty item $\lambda\|u_{TD3}\|^2$ imposes a monotonically increasing cost on the large residual action instructions. Its gradient naturally pulls the saturated action back to near zero to prevent the strategy from being stuck at the action boundary. By implementing this soft constraint, the TD3 agent must allocate its limited policy capacity to those scenarios where residual compensation can significantly improve the tracking error.

The global exploration ability of TD3 agent does not come from regularization term. Instead, it depends on intrinsic Gaussian action noise and the clipped Double-Q learning mechanism of the TD3 algorithm. The proposed of regularization only narrows the action space to a safe and interpretable neighborhood defined by the physical prior. This design allows the residual module to focus on the high-precision local compensation, and avoid the typical instability and saturated outputs of pure data-driven strategies.

3.4. Safety-Constrained Smooth Takeover and Dynamic Fusion Mechanism

Although using TDC as a residual module can significantly improve the overall tracking accuracy of the system, due to the uncertainty of the data-driven algorithms, its output action is very difficult to meet the RDC condition. In extreme or unseen traffic scenarios, this may cause the aggregated control action break the physical safety constraints, leading to the collision risks. To overcome this inherent defect, this article proposes a smooth switching and dynamic integration mechanism based on safety constraint boundary. This mechanism realizes the seamless coordination between underlying physical module and upper learning algorithm through the real-time assessment of the safety margin of the platoon and employs dynamic weight allocation combined with spatiotemporal dual soft-filtering.

The core of this mechanism is first it accurately define the vehicle's safety state set. Based on the kinematic principles, this article integrates key indicators such as inter-vehicle distance, relative velocity, and Time-to-Collision (TTC), and defines the following safety assessment variables of the collaborative control laws:

$$\begin{cases} e = s - s_{des} = s - (Tv^i + s_0) \\ \Delta v = v^i - v^{i-1} \\ TTC = \dfrac{s}{max(\epsilon, \Delta v)} \end{cases} \tag{17}$$

where e is the spacing error, s is the actual distance , s_{des} is the desired distance defined by a constant time-gap policy with $T = 1s$ and $s_0 = 1m$, v^i and v^{i-1} represent the velocities of the ego and lead vehicles respectively, and $\epsilon = 10^{-3}$ is a small positive constant to prevent division by zero. Accordingly, the safety boundary set $\mathcal{H}$ is defined as:

$$\mathcal{H} = \{e < -\varepsilon_s \vee \Delta v > \Delta v_{max} \vee (\Delta v > 0 \wedge TTC < T_{min})\} \tag{18}$$

where the safety thresholds are set to $\varepsilon_s = 0.7m$, $\Delta v_{max} = 1m/s$, and $T_{min} = 1s$. when any condition in set $\mathcal{H}$ is met, it indicates that the system is approaching a physical safety boundary and has entered a potential risk mode. To avoid the acceleration jumps and passenger discomfort caused by abrupt shifts in control authority during mode switching, we introduce a continuous spatial fusion weight function $\lambda(e)$ and a temporal safety soft-gate g_k. The spatial weight $\lambda(e)$ is defined as:

$$\lambda(e) = \frac{1}{1 + e^{-k(e-e_0)}} \tag{19}$$

where $k = 6$ adjusts the steepness of the weight function and $e_0 = -1$m is the equilibrium point at which the MPC and TD3 contributions are equally weighted. This design ensures that $\lambda(e)$ approaches 0 when the spacing is severely insufficient ($e \ll e_0$) and approaches 1 when the spacing is ample ($e \gg e_0$).

Because risk evolution in traffic environments unfolds continuously rather than jumping discretely, we design a dynamic soft-gating strategy with hysteresis to maximize the precision benefits from reinforcement learning:

$$g_k = clip(g_{k-1} + \beta_g(g_k^* - g_{k-1}), g_{k-1} - \delta_g, g_{k-1} + \delta_g) \tag{20}$$

where g_k^* is an indicator variable defined with respect to the unsafe state set $\mathcal{H}$:

$$g_k^* = \begin{cases} 0, & if\ x_k \in \mathcal{H} \\ 1, & if\ x_k \notin \mathcal{H} \wedge e > \varepsilon_s \\ g_{k-1}, & if\ x_k \notin \mathcal{H} \wedge e \le \varepsilon_s \end{cases} \tag{21}$$

In this formulation, $g_0 = 1$ initializes g_{k-1}, allowing the TD3 residual module to be fully active initially. The coefficient $\beta_g = 0.15$ then governs the system's dynamic response rate, while $\delta_g = 0.2$ caps the single-step change rate of control authority. Here, $\mathcal{H}$ denotes the unsafe state set from Eq. (18), and x_k is the system state at step k. When system detects a potential threat $\mathcal{H}$, this soft-gating mechanism controller does not instantly deprive the control of the TD3 algorithm, but will gradually reduce the proportion of the residual action step by step, using δ_g as the maximum decay rate. If the dangerous state persists in the subsequent time steps, the weight will gradually converge to zero, and finally realize the independent safety decision completely taken over by the DMPC. This smooth degradation mechanism greatly in synchronization with the physical objective law of the progressive traffic risk evolution. It not only maximizes the compensation benefits of the DRL in the critical state to the greatest extent, but also fundamentally eliminates the control chattering caused by mode switching.

Finally, the effective weight of the TD3 residual module is calculated as the product of the spatiotemporal constraints, $\lambda_{eff} = \eta g_t \cdot \lambda(e)$. The final control command sent to the low-level actuator is:

$$u = u_{DMPC} + \lambda_{eff} \cdot u_{TD3} \tag{22}$$

In summary, this fusion strategy eases the difficulty vanilla TD3 faces in satisfying RDC conditions. It also improves anti-chattering and control smoothness by combining continuous spatial mapping with temporal soft-gating. Ultimately, this strikes an efficient balance between self-adaptive cooperative optimization and the absolute safety floor of interpretable physical models.

4 Numerical Simulation Experiments

4.1. Ablation Study

We conduct ablation experiments under the Emergency Braking and Recovery Scenario to evaluate the residual action regularization and

Just Accepted

the spatiotemporal soft-gating takeover mechanism, the two core components of the proposed framework.

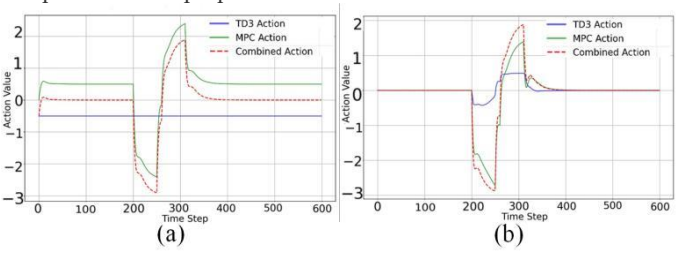


Fig. 3. Output of the TD3 residual module during training. (a) without regularization. (b) with regularization.

We first evaluate the effectiveness of the residual action regularization mechanism. Fig. 3(a) shows the line chart of the baseline action , residual action and combination action of the first following vehicle without regularization. In the whole simulation process, the actions of TD3 are saturated in the lower bounds of the action space, i.e., $u_{TD3} = -0.5$. This indicates that there is a gradient collapse and the agent cannot provide any state-dependent compensation for the baseline action. Due to the lack of L2 penalty with large-amplitude residuals, the overestimate Q-values make policy network to utilize the action boundary. Subsequently, the policy falls into a state of the saturation and loses the ability to explore the effective residual strategies. By contrast, Fig. 3(b) shows that TD3 agent is able to learn the residual action that can compensate output of DMPC actively when the regularization is enabled. This residual action presents the clear time structure and are related to the action of the DMPC. This confirms that agent has captured the mismatch between module and environment, and provides the effective nonlinear corrections. This comparison demonstrates that the regularization mechanism is essential for the residual learning module. Without this mechanism, the TD3 agent will not be able to make a contribution to the whole control performance.

Next, we evaluate the role of the smooth takeover and dynamic fusion mechanism. Fig. 4 compares and evaluates the performance of the PERL framework with and without using this mechanism based on three evaluation indicators. In the variant that does not use this mechanism, TD3' s residual action is directly superimposed onto the DMPC baselines without any safety filtering. When agent outputs the unsafe action, it will introduce dangerous compensation instructions. As shown in the figure, compared with the variants without this mechanism, the use of the mechanism leads to the slightly increase on Dam_{acc}. The reason why happens this kinds of phenomenon is because when the system is close to the safety constraints, soft-gating control will weaken the contribution of the residual, thus partially reduce the compensation effect that is beneficial for the stability of the platoon. However, the big improvement of other two indicators effectively make up for the small cost, thus this trade-of is absolutely justified. Because this mechanism effectively suppressed the acceleration mutation caused by the unsafe TD3 action, the ride comfort has been significantly enhanced. More importantly, the significant reduction of $iTTC$ presents the better safety performance, because this mechanism can detect potential dangerous residual instructions in time and gradually weaken vehicles before they endanger the safety. These results confirm that the smooth takeover and dynamic fusion mechanism effectively prevent the TD3 agent form injecting the harmful compensation into the DMPC baselines, achieving a good balance between the safety and comfort, and ensuring the smooth and collision-free platoon operation.

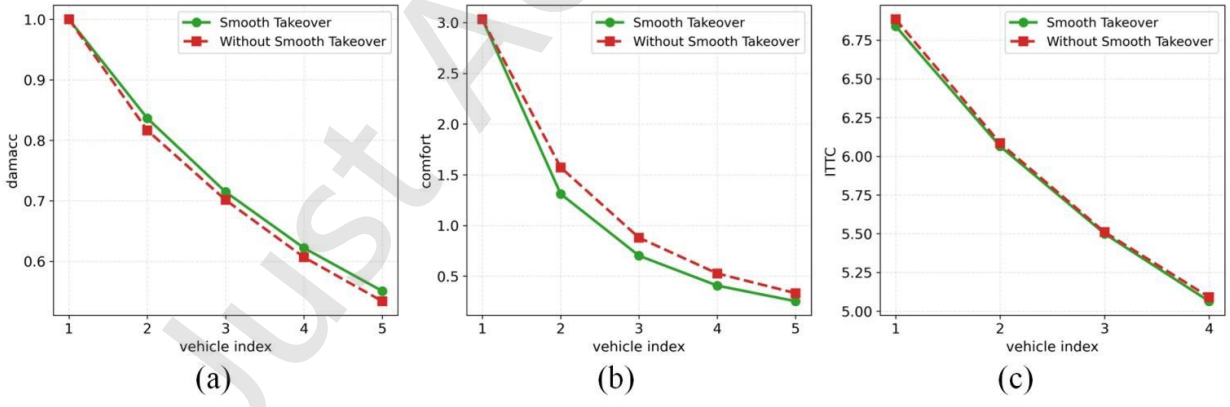


Fig. 4. Ablation study on the soft-gating mechanism. (a) Cumulative damping ratio. (b) Ride comfort. (c) Inverse time-to-collision.

4.2. Model Training Strategy

For the PERL framework, we augment the experience replay buffer so the TD3 agent can better capture and compensate for unmodeled dynamics. Alongside standard Markovian transitions, the buffer also records the baseline control sequence u_{DMPC} and its subsequent state transitions. This treats the physical model's output as an explicit heuristic prior. As a result, the TD3 agent can focus its learning capacity on the nonlinear residual action u_{TD3} atop this baseline. During parameter updates, the policy network optimizes the combined performance of the composite action $u_{target} = u_{DMPC} + u_{TD3}$. This preserves the interpretability and boundary safety of the underlying physical model while narrowing the RL search space. This improves convergence stability and disturbance rejection.

To reduce the risk of sub-optimal convergence or overfitting in high-dimensional state spaces, we use a curriculum learning strategy. This strategy guides the agent through an incremental task gradient. We begin with a basic car-following task, where the lead vehicle holds a constant velocity. This allows the agent to establish basic control logic and reward mapping. We then add low-amplitude sinusoidal disturbances to build up its handling of continuous traffic fluctuations. After that, complexity increases through a strong disturbance rejection stage. In this stage, high-amplitude oscillations and emergency braking maneuvers push the agent to output safe, compensatory actions under abrupt state transitions. The final stage pushes the agent toward extreme case generalization with high-frequency acceleration-deceleration sequences. Across all curriculum stages, the kinematic

Just Accepted

parameters of the disturbances (frequencies, amplitudes, and trigger timings) are continuously randomized within predefined physical bounds. This domain randomization prevents the agent from overfitting to specific trajectories. It also supports its zero-shot generalization when facing deterministic benchmarks or unseen real-world datasets in later evaluations.

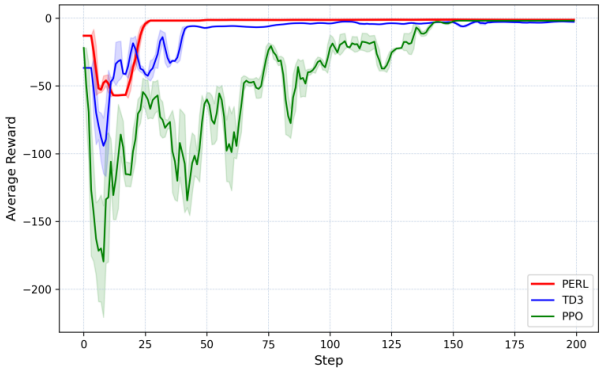


Fig. 5. The average reward of the training episode.

Fig. 5 presents the average reward curve of the PERL compared with PPO and TD3 algorithm. This line chart evaluates the effectiveness of the proposed framework and training mechanism. Pure data-driven PPO and TD3 agent perform the high variance and need to go through a long and ineffective exploration stage to establish the feasible strategies. In contrast, PERL convergence is much faster and can achieve stability in less episodes. It should be noticed is that the PERL rewarding function contains an extra residual regularization that is not available in TD3 and PPO. Therefore, the absolute reward values in Fig. 5 cannot directly compare between the different algorithm. Conversely, this chart mainly demonstrates the convergence speed, that is, the round numbers required to achieve a stable policy. This rapid convergence confirms the advantage of introducing the DMPC prior. The baseline action of DMPC based on the physical module provides a safe and interpretable control. It protects the TD3 residual module and limits the search scope to the narrow small value action space. In this way, the residual agent does not need to explore the whole unbounded action space. It just needs to study the mismatch between the module and environment, thus bypassing the large number of trial-and-error process brought by the standard reinforcement learning. The core hyperparameters for the TD3 modules are detailed in Table 1.

Table 1. Hyperparameters of TD3 and the Corresponding Values

Hyperparameter	Value
Learning rate of the actor network	0.0001
Learning rate of the critic network	0.0001
Discount factor	0.99
Optimization Algorithm	Adam
Batch Size	256
Soft Update Rate	0.01
Policy Frequency	4
Experience Replay Size	50000

4.3. Experimental Design

We build a simulation environment with a five-vehicle platoon and design four test scenarios of increasing complexity. Scenario 1 applies sinusoidal acceleration and deceleration to the lead vehicle, which checks basic car-following, whether the framework keeps spacing error and velocity consistent under ordinary traffic fluctuations. Scenario 2 adds stochastic oscillations through an Ornstein-Uhlenbeck process, so the controller faces continuous, irregular disturbances rather than smooth waves. Scenario 3 is emergency acceleration and deceleration: sudden maneuvers triggered by external disturbances, where the smooth takeover mechanism must keep the platoon safe and comfortable when collision risk appears. Scenario 4 draws on the NGSIM dataset for a Sim-to-Real check. We feed the lead vehicle with real acceleration trajectories from the I-80 interstate highway in California, sampled at 10 Hz, to see how the model behaves under real-world long-tail distributions.

4.4. Comparative Experiments

4.4.1 Sinusoidal Smooth Acceleration and Deceleration Scenario

We design a smooth acceleration and deceleration scenario with sinusoidal fluctuations to evaluate the dynamic response and string stability of the controller under standard car-following conditions. The leading vehicle runs through a complete acceleration-cruise-deceleration cycle using a continuous and regular sinusoidal acceleration profile. Its velocity smoothly rises from low to high speed before finally decelerating. This scenario mainly tests how well each model fits and tracks the vehicle's nonlinear dynamics.

Just Accepted

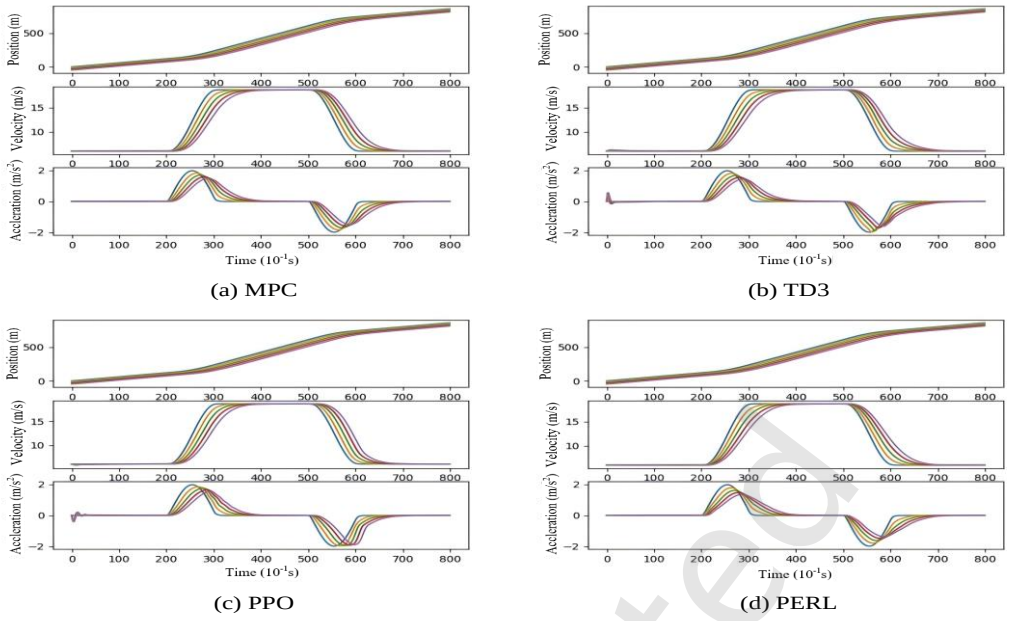


Fig. 6. Comparison of kinematic state trajectories of vehicle platoons under four control strategies in smooth sinusoidal acceleration and deceleration.

Fig. 6 shows that during the smooth transition from low to high speeds, all algorithms handle the basic longitudinal car-following task. However, they differ notably in acceleration smoothness and multi-vehicle cooperative tracking precision. We evaluate the performance of these four methods quantitatively using three metrics: the cumulative damping ratio (Shi et al., 2022), inverse time-to-collision ($iTTC$) (Balas and Balas, 2006), and ride comfort (Liao et al., 2025).

The cumulative damping ratio evaluates string stability by measuring how well the platoon attenuates traffic oscillations. It is defined as:

$$d_{p,i} = \frac{\left(\sum_{k=0}^{m}\left|a_k^i - a_{i,mean}\right|^2\right)^{\frac{1}{2}}}{\left(\sum_{k=0}^{m}\left|a_k^0 - a_{0,mean}\right|^2\right)^{\frac{1}{2}}} \tag{23}$$

where i ($1 \le i \le n$) is the vehicle index within the platoon, m represents the total traveling steps, a_k^0 is the leading vehicle's acceleration at time k, a_k^i is vehicle i's acceleration at time k, and $a_{i,mean}$ denotes the average acceleration over the m time steps. If $d_{p,i} < 1$ and strictly decreases as i increases, it indicates that traffic oscillations decay as they propagate upstream, confirming that the vehicle platoon is string stable.

We use the inverse $iTTC$ as the primary safety metric. Unlike standard TTC, $iTTC$ calculates collision conflicts more precisely and avoids the mathematical singularity (infinity) when the leading and following vehicles travel at identical speeds. The $iTTC$ at a specific time step is computed as:

$$TTC_{i,k}^{-1} = max\left(0, \frac{v_k^i - v_k^{i-1}}{p_k^{i-1} - p_k^i}\right) \tag{24}$$

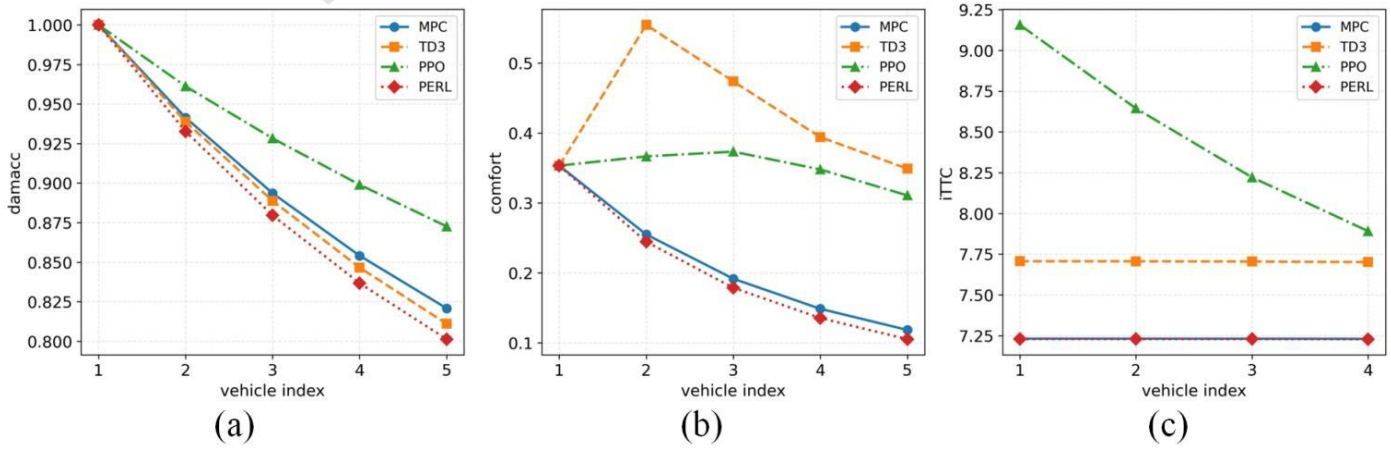
where p_k^i and v_k^i denote the position and velocity of vehicle i at time k, respectively. A $TTC_{i,k}^{-1}$ value of 0 indicates the absence of collision risk (maximum safety). generally, a smaller value signifies higher safety. Integrating $TTC_{i,k}^{-1}$ over the entire travel duration yields the aggregate inverse time-to-collision, TTC_i^{-1}, which reflects the overall safety of vehicle i throughout the trip:

$$TTC_i^{-1} = \sum_{k=0}^{m} max\left(0, \frac{v_k^i - v_k^{i-1}}{p_k^{i-1} - p_k^i}\right) \tag{25}$$

Ride comfort assesses vehicle stability by measuring acceleration variation. It is calculated as:

$$comfort_i = \sum_{k=0}^{m-2}\left(a_{k+1}^i - a_k^i\right)^2 \tag{26}$$

where $comfort_i$ represents the ride comfort of vehicle i. A smaller value indicates a higher level of comfort.



Just Accepted

Fig. 7. Evaluation metrics comparison. (a) Cumulative damping ratio. (b) Ride comfort. (c) Inverse time-to-collision. i=1 and i=N represent the first and last follower, respectively.

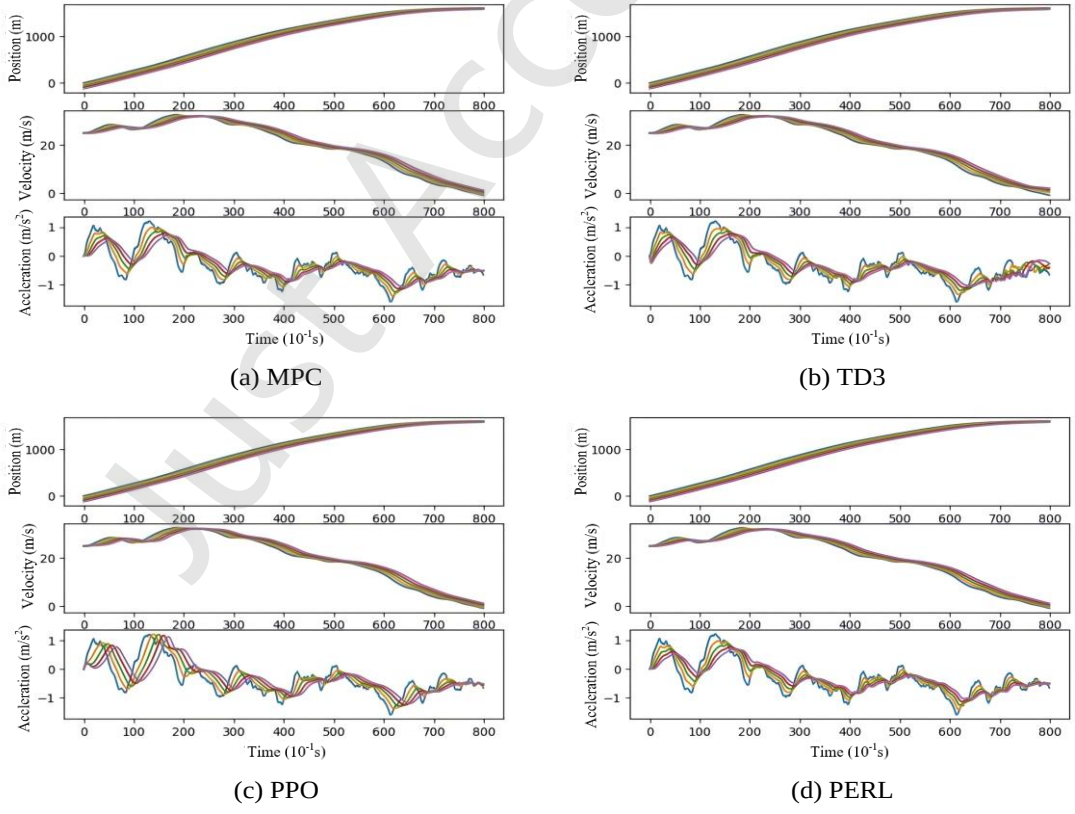
Fig. 7 shows how the four benchmarked methods perform under the sinusoidal acceleration-deceleration profile. For string stability, both PERL and TD3 achieve cumulative damping ratios notably lower than PPO. As the vehicle index rises, damping values clearly attenuate. This confirms our framework effectively suppresses the upstream propagation of traffic oscillations. By contrast, PPO suppresses oscillations poorly. Without physical-constraint guidance, it fails to maintain the string stability required for platoon operation.

Analyzing the trade-off between safety and comfort highlights key differences among these algorithms. Fig. 7(b) and (c) show that although PPO slightly outperforms vanilla TD3 in ride comfort, its $iTTC$ values are much higher, indicating a greater collision risk during dynamic following. TD3 offers better safety and tracking than PPO, but its early-stage exploration tends to cause high-frequency acceleration fluctuations. This, in turn, hurts the comfort metric. This explains the abrupt spikes in the TD3 comfort curve (Fig. 7(b)), where Gaussian exploration noise temporarily pushes the policy away from optimal actions.

By contrast, PERL uses MPC for a safe baseline control effort, with TD3 residual actions refining it further. This design curbs the aggressive control typical of RL during exploration. As a result, PERL eliminates these spikes. The smooth takeover mechanism limits the rate of change in control authority, and the residual regularization penalizes large-amplitude commands. Together, they prevent abrupt acceleration jumps. Overall, PERL successfully combines the interpretability of physical models with the adaptability of reinforcement learning. Under sinusoidal fluctuations, the framework compensates for the precision loss of traditional MPC when handling nonlinear time delays, while overcoming the safety-constraint limitations of purely data-driven approaches. Ultimately, PERL achieves the highest safety and cumulative damping while maintaining ride comfort comparable to MPC. This confirms the advantage of the PERL architecture in enhancing overall platoon control performance.

4.4.2 OU Stochastic Oscillation Scenario

After the baseline verification, we add irregular acceleration disturbances using an OU stochastic process. This scenario tests the robustness of the control strategies in complex dynamic environments. It simulates the high-frequency velocity fluctuations caused by driver distraction or subtle pedal adjustments in real-world driving. The mean-reverting nature of the OU process generates continuous, small-scale stochastic oscillations around a predefined mean velocity (set to 25m/s), using a mean-reversion rate of $\theta = 0.2$ and volatility of $\sigma = 0.5\text{m/s}^2$. This creates a significant challenge for the algorithms' high-frequency noise filtering and dynamic adaptability.



(a) MPC (b) TD3

(c) PPO (d) PERL

Fig. 8. Comparison of kinematic state trajectories of vehicle platoons under four control strategies in OU stochastic oscillation conditions.

Just Accepted

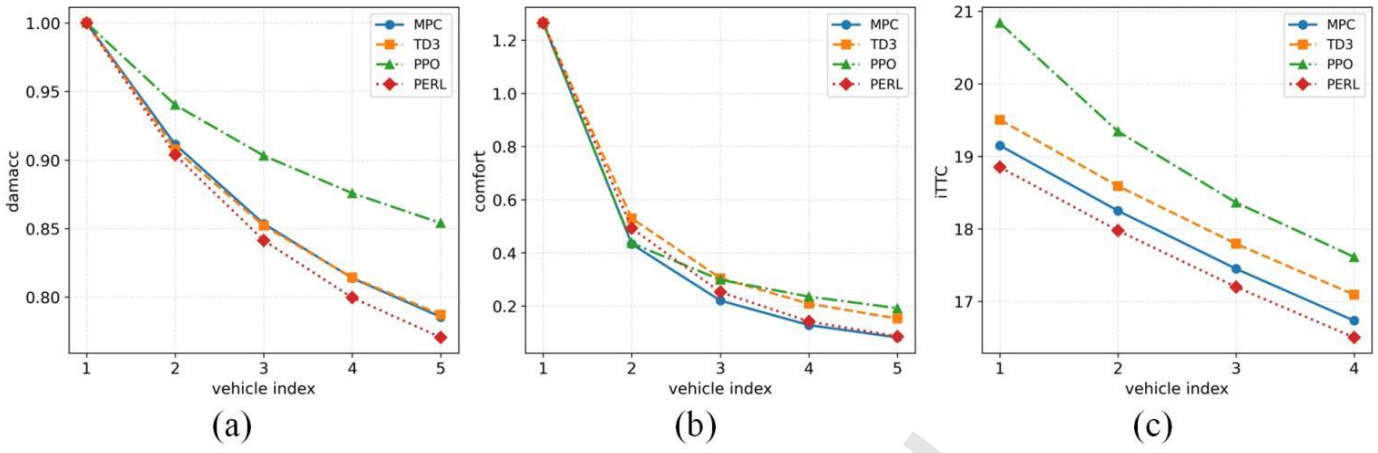


Fig. 9. Evaluation metrics comparison. (a) Cumulative damping ratio. (b) Ride comfort. (c) Inverse time-to-collision.

Fig. 9 further validates the rationality and advantages of the PERL framework. When handling high-frequency stochastic disturbances, PERL's hierarchical cooperative control strategy divides the task: the physics-driven MPC handles vehicle dynamics constraints and sets safety boundaries for robust low-level control, while the data-driven TD3 residual module generates small compensatory actions to fine-tune the baseline control output from MPC. This mechanism combines a physical model with reinforcement learning to smooth the high-frequency oscillatory noise from the OU process. Comparisons show that PERL achieves the lowest values across three core metrics compared to the other three baseline algorithms: cumulative damping ratio (Dam_{acc}), ride comfort ($Comfort$), and $iTTC$. These results show that the framework filters the stochastic fluctuations of the preceding vehicle and improves platoon following smoothness. It also maintains high system safety and better error suppression under demanding dynamic conditions.

4.4.3 Emergency Braking and Recovery Scenario

To test how the control framework intervenes and avoids collisions in safety-critical states, we design an extreme test scenario with rapid deceleration and subsequent acceleration. Here, the lead vehicle brakes aggressively and then recovers quickly to simulate severe traffic disturbances. These boundary conditions demand high transient performance from the controller. It must respond quickly to maintain inter-vehicle spacing and guarantee safety near the system's physical limits.

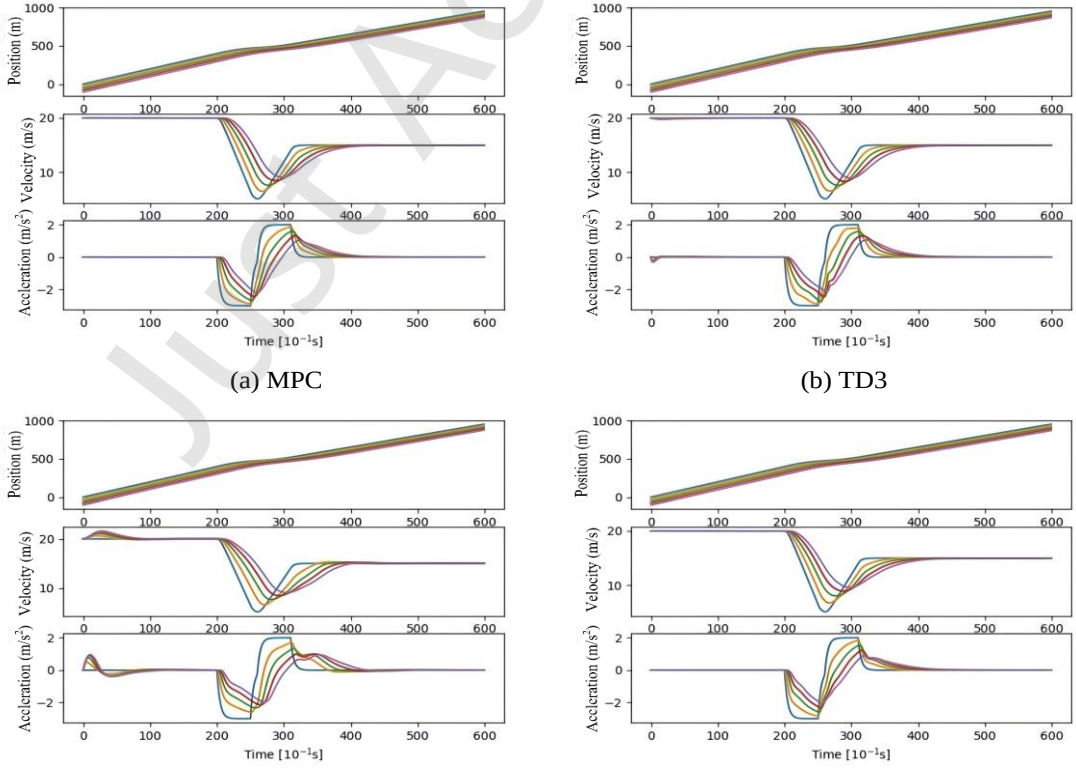


Fig. 10. Comparison of kinematic state trajectories of vehicle platoons under four control strategies in emergency braking and recovery conditions.

Just Accepted

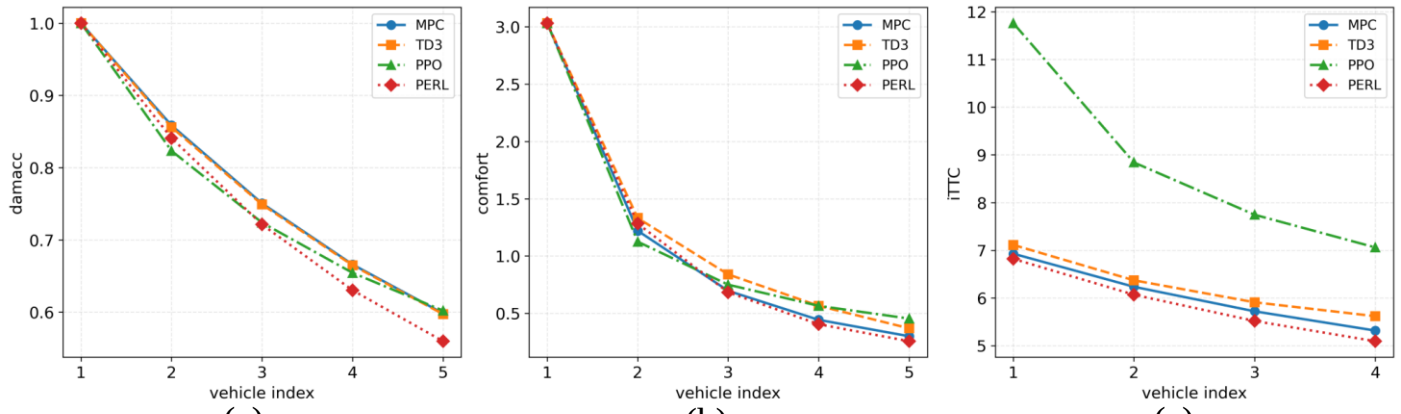


Fig. 11. Evaluation metrics comparison. (a) Cumulative damping ratio. (b) Ride comfort. (c) Inverse time-to-collision.

Fig. 11 provides the quantitative comparison results, confirming the advantages of the PERL framework in terms of safety and stability. As shown in Fig. 11(c), PERL maintains a consistently low iTTC, outperforming the other algorithms. This confirms the effectiveness of our safety-boundary-constrained smooth switching mechanism.

When sharp deceleration triggers potential collision risks, the smooth handover mechanism progressively reduces the weight of the TD3 residual module in real-time. As a result, the MPC module—which provides recursive feasibility guarantees—takes over safety-critical decision-making to ensure robust operation. PERL also improves ride comfort. As shown in Fig. 11(b), PERL uses the smooth guidance of the physical baseline to filter out the high-frequency acceleration oscillations that plague pure TD3 during policy exploration under extreme conditions.

For string stability, the PERL framework maintains excellent cumulative damping ratios, preventing disturbances from being amplified as they propagate upstream through the platoon. Our analysis shows that by combining physical models with reinforcement learning, PERL surpasses the tracking precision of traditional physical models under extreme conditions. It also addresses the safety unreliability of black-box algorithms. This architecture ultimately achieves a balance among safety, stability, and passenger comfort.

4.4.4 Empirical Validation via NGSIM Dataset

After evaluating idealized simulations and edge-case scenarios, we use the Federal Highway Administration's NGSIM dataset to bridge the simulation-to-reality gap and test the generalization of our framework. We extract a typical longitudinal acceleration trajectory as the motion input for the lead vehicle. This trajectory spans 500 time steps. It captures real acceleration and deceleration patterns of human drivers in complex traffic flows. Unlike the synthetic sinusoidal or OU processes, this empirical data contains hidden factors like road gradients, nonlinear human driving styles, and sensor measurement noise. These elements are not explicitly defined in the simulation models.

Just Accepted

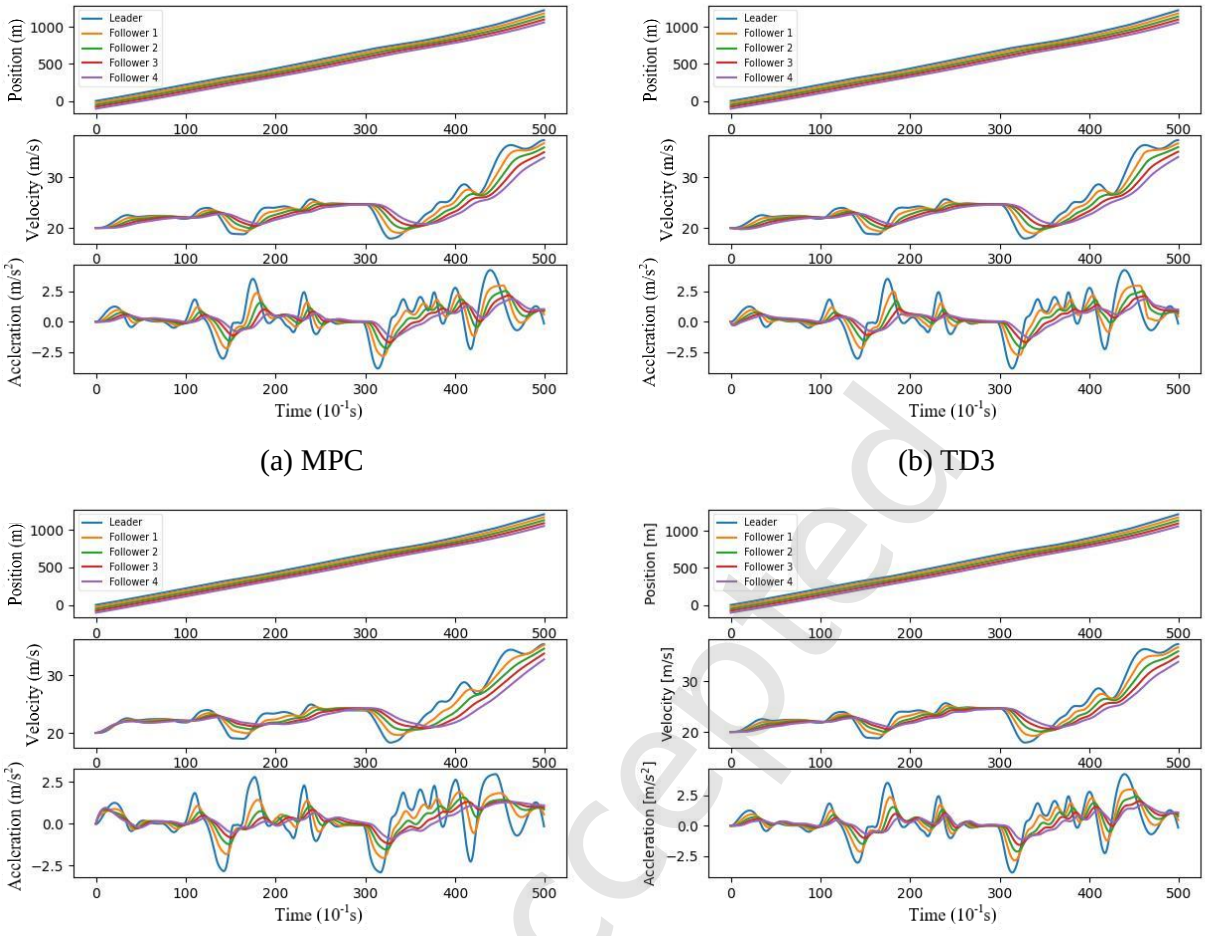


Fig. 12. Comparison of kinematic state trajectories of vehicle platoons under four control strategies in empirical validation using the NGSIM dataset.

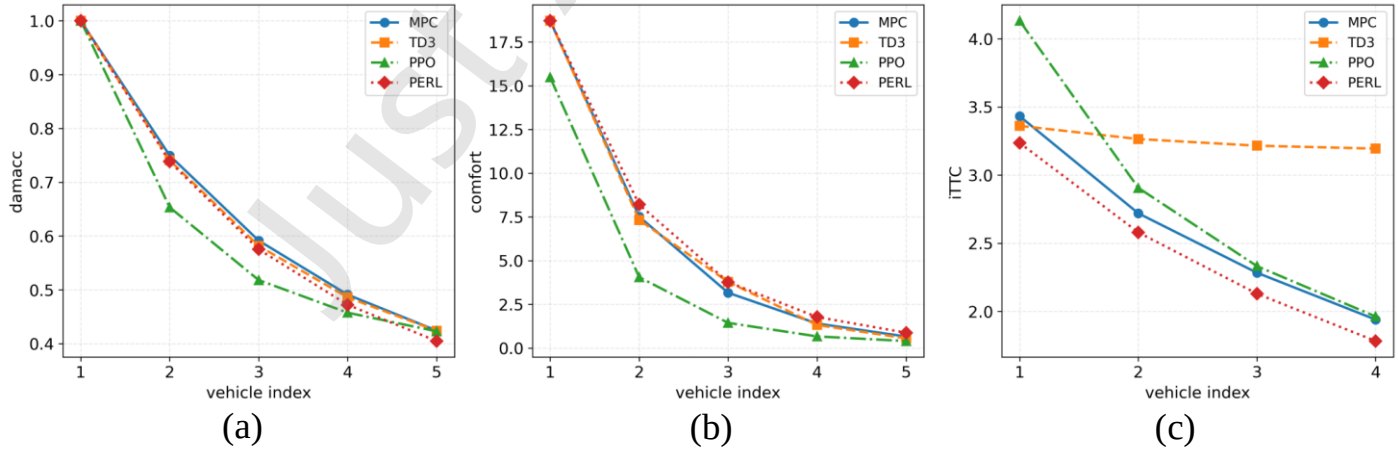


Fig. 13. Evaluation metrics comparison. (a) Cumulative damping ratio. (b) Ride comfort. (c) Inverse time-to-collision.

Fig. 13 shows that PERL achieves strong generalization without extra fine-tuning on the empirical dataset. Under the unpredictable acceleration behaviors of human drivers, PERL maintains lower cumulative Dam_{acc} values than conventional MPC and baseline RL algorithms. This proves it can effectively absorb and dissipate stochastic oscillations in real-world traffic.

In terms of safety, as shown in Fig. 13(c), baseline RL algorithms lack explicit physical boundary constraints. Under real-world noise environments, they tend to output high-risk or aggressive control commands, thereby degrading driving safety. In contrast, PERL relies on hard-constraint shielding from the physical model and a smooth switching mechanism based on safety boundaries, which guarantees safety. However, due to this switching mechanism, the improvement in comfort might be limited. Overall, these results confirm that the Physics-Enhanced Residual Learning architecture can overcome the safety blind spots and generalization hurdles of deploying RL in real-world environments, highlighting its engineering value in complex automated vehicle platoon control.

Just Accepted

5 Concluding Remarks

This study addresses the PERL framework to cope with this collaborative control difficulties of heterogeneous vehicle platoons in complex dynamic environment. This framework integrates the DMPC and TD3 algorithm, achieving the deep integration of physical prior knowledge and data-driven adaptability. To alleviate the widespread problem of the strategy collapse in the mixed framework, we design a reward shaping mechanism based on residual action regularization to guide the network to learn the complex nonlinear compensation strategies at the target function level. Besides, in order to overcome the inherent difficulty of DRL in the inability to strictly meet safety constraints, we introduce the dynamic fusion and smooth handover mechanism based on safety-boundary constraints. This mechanism effectively guarantees the recursive feasibility of the system and ensures the seamless transfer of the control authority under the safety-critical conditions.

Under four different test scenarios: sine fluctuation, OU stochastic processes, emergency acceleration/deceleration, and real-world NGSIM empirical datasets-substantiate, we make extensive simulation experiments. The results show that the PERL framework perform well in control performance and generalization ability. Quantitative analysis confirms that, compared with the traditional MPC methods and vanilla data-driven algorithms (TD3 and PPO), the proposed framework greatly decreases the cumulative suppression rate and improves the ride comfort and safety. And when facing the actual traffic data of the human driving characteristic, PERL can effectively alleviates the common control collision and strategy collapse in pure reinforcement learning, thus verifying the robustness of the physical enhancement framework when dealing with the long-tail traffic scenarios.

In summary, the PERL framework proposed by this article provided a new solution for autonomous vehicle platoon control, taking into the account the needs of safety, interpretability, and high precision. By organically combine with the physical baseline action and nonlinear residual compensation, this method can not only compensate the loss of the accuracy caused by traditional mathematical modelling errors and actuator time delays, but also provide a fusion framework that balances the robustness and safety for the application of the DRL in safety-critical tasks. In future research will explore the multi-constraint coupling, and extend the framework to the collaborative scheduling scenario involving the large-scale heterogeneous traffic flows.

Author contributions

Haigen Min: Writing-Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing - review & editing. Zihao Mao: Writing-original draft, Methodology. Xia Wu: Formal analysis, Writing - original draft, Writing - review & editing. Ruichen Zhang: Validation, Writing - review & editing. Ye Li: Resources, Writing - review & editing. Xiangmo Zhao: Data curation, Resources. Yang Liu: Validation, Writing - review & editing.

Replication and data sharing

The source codes and replication package are available on ETS-Data at https://doi.org/10.26599/ETSD.2026.9190126.

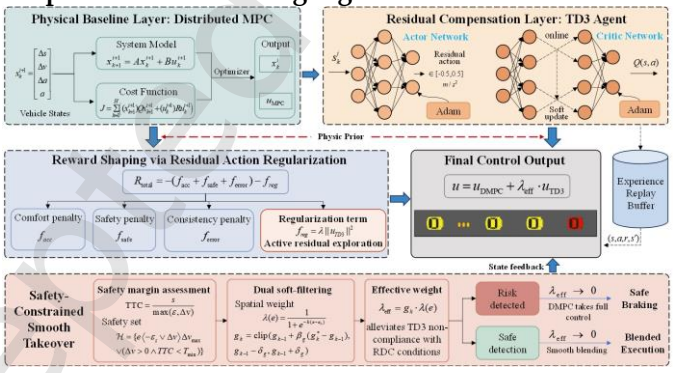
Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (52372426, 52302399, U25B20136), Key Research and Development Program of Shaanxi Province (2026CY-ZDGG-53), and Fundamental Research Funds for the Central Universities, Chang'an University (300102246304, 300119246501).

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Graphical abstract and Highlights



References

Bai, L., Zheng, F., Hou, K., Liu, X., Lu, L., Liu, C., 2024. Longitudinal control of automated vehicles: A novel approach by integrating deep reinforcement learning with intelligent driver model. IEEE Transactions on Vehicular Technology, 73, 11014 – 11028.

Balas, V. E., Balas, M. M., 2006. Driver assisting by inverse time to collision. In: 2006 World Automation Congress, 1 – 6.

Bando, M., Hasebe, K., Nakayama, A., Shibata, A., Sugiyama, Y., 1995. Dynamical model of traffic congestion and numerical simulation. Physical Review E, 51, 1035.

Caruntu, C. F., Braescu, C., Maxim, A., Rafaila, R. C., Tiganasu, A., 2016. Distributed model predictive control for vehicle platooning: A brief survey. In: 2016 20th International Conference on System Theory, Control and Computing (ICSTCC), 644 – 650.

Chen, J., Zhu, B., Zhang, M., Ling, X., Ruan, X., Deng, Y., et al., 2025. Multi-agent deep reinforcement learning cooperative control model for autonomous vehicle merging into platoon in highway. World Electric Vehicle Journal, 16, 225.

Feng, S., Zhang, Y., Li, S. E., Cao, Z., Liu, H. X., Li, L., 2019. String stability for vehicular platoon control: Definitions and analysis methods. Annual Reviews in Control, 47, 81 – 97.

Kashinath, K., Mustafa, M., Albert, A., Wu, J.-L., Jiang, C., Esmaeilzadeh, S., Azizzadenesheli, K., Wang, R., Chattopadhyay, A., Singh, A., et al., 2021. Physics-informed machine learning: case studies for weather

Just Accepted

and climate modelling. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 379, 20200093.

Krishnapriyan, A., Gholami, A., Zhe, S., Kirby, R., Mahoney, M. W., 2021. Characterizing possible failure modes in physics-informed neural networks. Advances in Neural Information Processing Systems, 34, 26548 – 26560.

Liao, L., Xing, P., Gan, Z., Wei, X., Liu, Y., 2025. Optimizing safety and comfort for distracted drivers via adaptive shared control. Digital Transportation and Safety, 4, 89 – 100.

Luo, L.-H., Liu, H., Li, P., Wang, H., 2010. Model predictive control for adaptive cruise control with multi-objectives: Comfort, fuel-economy, safety and car-following. Journal of Zhejiang University SCIENCE A, 11, 191–201.

Memon, Z. A., Jumani, S., Larik, J., 2012. Longitudinal control of a platoon of road vehicles equipped with adaptive cruise control system. Mehran University Research Journal of Engineering & Technology, 31, 475 – 494.

Ma, G., Li, L., Wang, H., Liu, Z., Bacon, P.-L., Tao, D., 2026. What Makes Value Learning Efficient in Residual Reinforcement Learning. https://arxiv.org/abs/2510.12717.

Ponso, A., Bonfitto, A., Amati, N., Ciuffo, B., Donà, R., 2026. Analysis of data volumes generated by automated vehicles for monitoring, reporting and remote driving. Automotive Innovation, 1 – 16.

Ran, Q., Liang, C., Liu, P., 2025. A safe lane-changing strategy for autonomous vehicles based on deep Q-networks and prioritized experience replay. Digital Transportation and Safety, 4, 170 – 174.

Shi, H., Chen, D., Zheng, N., Wang, X., Zhou, Y., Ran, B., 2023. A deep reinforcement learning based distributed control strategy for connected automated vehicles in mixed traffic platoon. Transportation Research Part C: Emerging Technologies, 148, 104 – 019.

Shi, H., Zhou, Y., Wang, X., Fu, S., Gong, S., Ran, B., 2022. A deep reinforcement learning-based distributed connected automated vehicle control under communication failure. Computer-Aided Civil and Infrastructure Engineering, 37, 2033 – 2051.

Shi, X., Lu, Y., Sheng, Z., Zhang, J., Chen, S., Huang, H., et al., 2026. Learning to drive from naturalistic trajectories: Offline reinforcement learning for safe speed guidance at signalized intersections. Journal of Intelligent and Connected Vehicles.

Shladover, S. E., Su, D., Lu, X.-Y., 2012. Impacts of cooperative adaptive cruise control on freeway traffic flow. Transportation Research Record, 2324, 63 – 70.

Sun, J., Xiong, C., Zhai, C., Li, Y., Liu, X., Chen, C., et al., 2026. Adaptive eco-cooperative adaptive cruise control for heterogeneous vehicle platoons using online identification-informed deep reinforcement learning. Engineering Applications of Artificial Intelligence, 165, 113517.

Treiber, M., Hennecke, A., Helbing, D., 2000. Congested traffic states in empirical observations and microscopic simulations. Physical Review E, 62, 1805.

Wang, L.-Z., Li, J., Zhang, H., Me, X.-P., Zhao, X.-M., Xie, D.-F., 2023. An attention-based car-following model based on fused data. IEEE Access, 11, 51368 – 51381.

Wang, S., Teng, Y., Perdikaris, P., 2021. Understanding and mitigating gradient flow pathologies in physics-informed neural networks. SIAM Journal on Scientific Computing, 43, A3055 – A3081.

Wang, S., Yu, X., Perdikaris, P., 2022. When and why PINNs fail to train: A neural tangent kernel perspective. Journal of Computational Physics, 449, 110768.

Xiao, D., Liu, Y., Xu, X., Šarić, Ž., 2026. Risk perception and quantitative assessment for intelligent vehicles: A systematic review of methodologies, policy implications, and transportation systems integration. Digital Transportation and Safety, 5, 136–143.

Yan, Y., Mascaro, E. V., Egle, T., Lee, D., 2025. I-CTRL: Imitation to control humanoid robots through bounded residual reinforcement learning. IEEE Robotics & Automation Magazine, 32, 59–67.

Yen, C.-C., Gao, H., Zhang, M., 2022. Deep reinforcement learning based platooning control for travel delay and fuel optimization. In: 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), 737 – 742.

Yang, J., Zheng, S., Sun, C., Li, H., Xu, L., 2026. Prior knowledge-assisted reinforcement learning for vehicle platoon control in cut-in scenarios. Journal of Intelligent and Connected Vehicles.

Zeng, D., Zheng, L., Yang, X., Li, Y., 2026. Reinforced imitative trajectory planning for urban automated driving. Automotive Innovation, 1 – 17.

Zhang, L., Zhang, R., Wu, T., Weng, R., Han, M., Zhao, Y., 2021. Safe reinforcement learning with stability guarantee for motion planning of autonomous vehicles. IEEE Transactions on Neural Networks and Learning Systems, 32, 5435 – 5444.

Zhang, N., Nazari, S., Chen, J., Xu, X., Zhang, H., 2024. Bayesian optimization for automatic tuning of an energy aware 2-d vehicle platoon. IEEE Transactions on Intelligent Vehicles, 9, 7380 – 7391.

Zhang, Y., Yan, R., 2022. Safe car-following strategy with multi-constraints based on deep reinforcement learning for autonomous driving vehicles. In: 2022 IEEE International Conference on Unmanned Systems (ICUS), 1151 – 1156.

Zhao, J., Pan, W., Liang, X., Lin, B., 2026. SPAR: Support-Preserving Action Rectification. https://arxiv.org/abs/2605.27877.

Zheng, Y., Li, S. E., Li, K., Borrelli, F., Hedrick, J. K., 2016. Distributed model predictive control for heterogeneous vehicle platoons under unidirectional topologies. IEEE Transactions on Control Systems Technology, 25, 899–910.

Just Accepted

Author biography



Haigen Min (Senior Member, IEEE) received the Ph.D. degree in traffic information engineering and control from Chang'an University, China, in 2018. He is currently a Professor at Chang'an University. His research interests include high-precision localization, environment perception, and cooperative control for connected and autonomous vehicles, and their fault detection and diagnosis.



Zihao Mao received the B.E degree in Internet of Things engineering from Yangtze University, China, in 2024. He is currently pursuing the M.S. degree at Chang'an University. His research interests include cooperative control of intelligent connected vehicle platoons.



Xia Wu received the Ph.D. degree in traffic information engineering and control from Chang'an University, China, in 2020. Her research interests include mixed traffic flow control, trajectory optimization in a mixed traffic flow with connected and automated vehicles, and human-driven vehicles.



Ruichen Zhang is currently pursuing the B.E. degree in artificial intelligence at the School of Information Engineering form Chang'an University, China,in 2023. Her research interests include cooperative control of intelligent connected vehicle platoons.



Ye Li received the B.Eng. and Ph.D. degrees from Southeast University, Nanjing, China, in 2014 and 2019, respectively, and was a joint doctoral student at the University of Michigan, Ann Arbor. He is currently a Professor with the School of Traffic and Transportation Engineering, Central South University, Changsha, China, and serves as Deputy Director of the Hunan Key Laboratory of Intelligent Transportation. His research interests include traffic safety and intelligent transportation systems. He has authored over 80 SCI/SSCI papers, including 6 ESI Highly Cited Papers. Prof. Li is an Associate Editor of Accident Analysis & Prevention.



Xiangmo Zhao received the Ph.D. degree from Chang'an University, Xi'an, China, in 2003. He is currently a Professor with the School of Information Engineering, Chang'an University, and also is the President of Xi'an University of Architecture and Technology. He currently serves as the Vice Chairman of the China Society of Automotive Engineers. His current research interests include testing of connected vehicles, automated vehicles, and intelligent transportation systems.



Yang Liu graduated in 2005, the B.S. degree of Computing (Hons.) from the National University of Singapore (NUS), and the Ph.D. degree in 2010. He started his post doctoral work with NUS, MIT, and SUTD. In 2012 fall, he joined Nanyang Technological University (NTU) as a Nanyang assistant professor. He is currently a full professor and the director of the Cybersecurity Lab, NTU. His research has

Just Accepted

bridged the gap between the theory and practical usage of formal methods and program analysis to evaluate the design and implementation of software for high assurance and security. His work led to the development of a state-of-the-art model checker, Process Analysis Toolkit (PAT). With more than 20 million Singapore dollar funding support, he is leading a large research team working on the state-of-the-art software engineering and cybersecurity problems. He has authored or coauthored more than 300 publications and six best paper awards in top-tier conferences and journals. His research interests include software verification, security, and software engineering. In 2011, Dr. Liu was the recipient of the Temasek Research Fellowship at NUS to be the Principal Investigator in the area of cyber securit

Just Accepted