Transportation Research Part C 184 (2026) 105541

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

An integrated deep reinforcement learning-linear control strategy for longitudinal control of connected and automated vehicles

Ziwei Yi [a], Min Xu [b,*], Shuaian Wang [a]

[a] Department of Logistics and Maritime Studies, Faculty of Business, The Hong Kong Polytechnic University, Hung Hom, Hong Kong
[b] Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong

ARTICLE INFO

Keywords:
Connected and automated vehicle
Deep reinforcement learning
String stability
Linear controller
Twin delayed deep deterministic policy gradient algorithm

ABSTRACT

String stability is important to maintain the longitudinal control of connected and automated vehicles (CAVs). It prevents the amplification of the perturbations as they propagate through the platoon. A variety of methods based on the deep reinforcement learning (DRL) approach have been proposed for longitudinal control of CAVs, which show excellent performance. However, none of those methods consider string stability on theoretical grounds due to the lack of explicit mathematical models in the DRL approach. To address this problem, we integrate a novel linear controller in a DRL framework for longitudinal control of CAVs, referred to integrated DRL-linear control (IDL) strategy. It can guarantee string stability while striking a good balance among various benefits, including vehicle safety, comfort, and efficiency. We employ the twin delay depth deterministic policy gradient (TD3) algorithm, a promosing DRL, in the proposed framework for decision. Numerical simulation results demonstrate that the proposed approach ensures theoretical string stability while significantly enhancing vehicle safety, comfort, and efficiency compared to human-driven vehicles (HDVs) and a model-based cooperative adaptive cruise control (CACC) strategy. It also outperforms the deep deterministic policy gradient (DDPG) and pure TD3 strategies in terms of safety, comfort, and string stability. These results indicate that the proposed IDL strategy not only benefits from the advantages of the linear controller in analyzing theoretical string stability conditions but also retains the advantage of the DRL approach in terms of optimizing the trade-off between multiple benefits.

1. Introduction

Longitudinal control of connected and automated vehicles (CAVs) is an essential technology to achieve autonomous driving. It enables the CAVs to adjust their acceleration to maintain the desired speed and distance from the preceding vehicle, depending on various control objectives, such as safety, comfort, fuel consumption, emissions, string stability, and traffic oscillation suppression. Previous studies often incorporated multiple objectives within the framework to achieve various traffic benefits (Hart et al., 2024; Li et al., 2018). As one of the major objectives, string stability is crucial for CAV longitudinal control. It refers to whether the magnitude of a disturbance is amplified for each leader–follower pair through a vehicular string (Zhou et al., 2020). The traffic flow is stable if the perturbations, such as the deviation from equilibrium velocities, generated by the preceding vehicle are dissipated in the propagation of the following vehicles, which is important for reducing the stop-and-go waves in the traffic flow. Despite its significance, most

* Corresponding author.
E-mail address: min.m.xu@polyu.edu.hk (M. Xu).

https://doi.org/10.1016/j.trc.2026.105541
Received 16 November 2024; Received in revised form 6 October 2025; Accepted 21 January 2026
Available online 31 January 2026

existing studies did not take adequate consideration of string stability on theoretical grounds in a generic multi-objective framework.

1.1. Literature review

Over the past decade, many studies have proposed various algorithms for longitudinal control of CAVs. These methods can be classified into three categories, including model-based methods (Arem et al., 2006; Chen et al., 2018; Montanino et al., 2021; Zhang et al., 2022), model predictive control (MPC) methods (Wang et al., 2016; Zhou et al., 2019), and artificial intelligence (AI) methods (Hart et al., 2024; Jiang et al., 2022; Li et al., 2018). Each category has its limitations and advantages.

Model-based methods provide mathematical formulations that facilitate straightforward analysis of string stability through theoretical derivations (Arem et al., 2006). For example, Montanino et al. (2021) analyzed the string stability of mixed traffic flow by linearizing a nonlinear car-following model, where the CAV was controlled by the same model with different parameters. Zhang et al. (2022) employed a linear controller to regulate CAVs and derived the transfer function to assess string stability, subsequently optimizing the controller's parameters to enhance the string stability of the CAVs system. However, these model-based methods tend to focus on a single optimization objective or limit their scope to property analysis without performing optimization, largely due to the inherent difficulty in achieving multiple optimization objectives (Zhou et al., 2017, 2019).

In contrast, the MPC methods can handle multiple optimal objectives, including the string stability. For example, Zhou et al. (2019) proposed a serial distributed MPC method for CAVs, incorporating two string stability criteria. Their results verified the effectiveness of this MPC method concerning the two string stability objectives, outperforming traditional MPC methods that do not account for string stability criteria. Wang and Jin (2023) proposed an MPC-based framework to balance trade-offs among multiple objectives such as safety, efficiency, shockwave mitigation, elasticity, and comfort, further demonstrating MPC's ability to address multiple objectives. However, the MPC methods require high-performance computing and can only be applied to convex problems (Zhou et al., 2017).

Emerging AI techniques, especially Deep reinforcement learning (DRL) methods, have gained significant attention for their potential to overcome the limitations of MPC mentioned above. DRL is a machine learning approach that processes complex input data using deep neural networks and learns to make decisions by interacting with the environment to maximize cumulative reward (Lu et al., 2023). It offers advantages such as superior accuracy (Huang et al., 2019), greater error tolerance (Lin et al., 2021), reduced computational costs (Buechel and Knoll, 2018), and improved applicability and generality (Shi et al., 2021). Extensive researches were developed on DRL-based multi-objective longitudinal control algorithms (Hart et al., 2024; Jiang et al., 2022; Li et al., 2018). For example, Zhu et al. (2020) proposed a CAVs longitudinal control method based on the DRL approach, considering the safety, efficiency, and comfort, outperforming traditional MPC-based control algorithms. However, due to the lack of explicit mathematical models, it is challenging for a DRL-based multi-objective longitudinal control algorithm to guarantee string stability, which is difficult to evaluate by instantaneous metrics.

Most DRL-based multi-objective longitudinal control algorithms evaluate the string stability by approximate measures in the time domain. One approach is to measure the string stability based on instantaneous values. For example, Chu and Kalabić (2019) designed a scheme to optimize the parameters of the optimal velocity model using the DRL approach. To ensure string stability, they restrict the velocity of the following vehicle to be less than the desired velocity. Ma et al. (2021) proposed a longitudinal control method based on the DRL approach, where the system was deemed stable if the relative velocity and relative distance of the following vehicle were both half that of the leading vehicle. Jiang et al. (2022) proposed a DRL-based method to mitigate traffic oscillations and improve string stability by constraining the following vehicle's velocity to be lower than that of the preceding vehicle. Another approach is to assess the string stability based on cumulative values. For instance, to achieve string stability, Shi et al. (2023, 2021) limited the cumulative ratio of the sum of squared accelerations between the following and the preceding vehicle to be less than one. Zhou et al. (2023) proposed a stability reward function based on the negative sum of the squared velocity differences between the following vehicle and the preceding vehicle, aiming to minimize velocity oscillations caused by preceding vehicle disturbances.

However, explicitly verifying whether the above conditions ensure string stability remains a challenge. For instantaneous values, without a clear definition of string stability criteria, assessments based on these values lack theoretical validation. For cumulative values in the time domain, they provide a general assessment of whether traffic flow remains stable over a period. The duration of the observation period can influence conclusions about string stability. However, the string stable system ensures that the variations in speed, spacing, or other parameters introduced by disturbances are bounded and do not escalate with distance or time. As a result, it is difficult to definitively confirm if these conditions ensure string stability.

Unlike time domain analysis, frequency domain analysis offers crucial insights into the system's stability and response characteristics that are not fully captured by time domain analysis (Yue et al., 2024). By obtaining the transfer function of the system in the frequency domain, string stability can be evaluated explicitly by theoretical mathematical equations independent of the duration of the observation period.

Two recent studies have investigated DRL-based longitudinal control strategies for CAVs from a frequency domain perspective. Yue et al. (2024) proposed a DRL-based longitudinal control method that integrates string stability conditions from both time and frequency domains. In their approach, time domain string stability conditions are applied to evaluate the DRL-generated output, while frequency domain string stability conditions are applied to evaluate the output of a linear controller. The outputs from the DRL and the linear controller are then combined to generate the final acceleration for the CAV. Although this method effectively reduces oscillations, it does not guarantee that the final output satisfies string stability constraints since the frequency domain stability conditions are not directly applied to the final combined output. Zhang et al. (2024) demonstrated that it is possible to analyze the string stability of DRL-based longitudinal control strategies using frequency domain methods. However, their study focused on analyzing the string stability of DRL approaches with fixed parameters, rather than applying string stability constraints to optimize the parameters of the

DRL approaches.

Overall, existing DRL-based longitudinal control strategies for string stability have not been validated by theoretical string stability conditions. Therefore, it is a critical issue to incorporate the theoretical string stability conditions into the DRL approaches to ensure strict compliance with string stability requirements while also enhancing multiple performance benefits.

1.2. Objectives and contributions

To fill the research gap, this study proposes a novel integrated DRL-linear control (IDL) strategy for longitudinal control of CAVs, ensuring theoretical string stability while achieving the trade-off solutions considering multiple benefits in vehicle safety, comfort, and efficiency. To fill the research gap, this study proposes a novel integrated DRL-linear control (IDL) strategy for longitudinal control of CAVs, ensuring theoretical string stability while achieving the trade-off solutions considering multiple benefits in vehicle safety, comfort, and efficiency. Unlike conventional DRL approaches that consist solely of environment and agent modules, the IDL framework introduces a linear controller as an embedded component within the DRL architecture and proposes a parameter tuning method to ensure string stability. This innovation allows for theoretical analysis of string stability in the frequency domain, which is not directly feasible in standard DRL-based strategies. In the IDL framework, the DRL agent does not output acceleration directly. Instead, it learns to generate three key parameters for a linear controller. The parameter tuning method revises the parameters that fail to satisfy the string stability condition. Then, the linear controller computes the final acceleration according to the revised parameters, which interact with the environment. The environment, in turn, provides feedback to update and improve the DRL policy. In this way, the DRL agent and linear controller collaboratively determine the control action, creating a structurally coupled decision-making process. The twin delayed deep deterministic policy gradient (TD3) algorithm is then employed in the proposed strategy as a DRL agent to implement longitudinal control for CAVs. Finally, numerical simulations using real-world data are conducted to evaluate the effectiveness of the proposed strategy compared to benchmark strategies.

The contributions of this study are summarized as follows:

- We propose a novel integrated DRL-linear feedback control (IDL) strategy that integrates a linear controller in the DRL approach. The strategy leverages the linear controller to theoretically guarantee string stability in the frequency domain while inheriting the advantage of DRL approaches for optimizing multiple objectives in complex and dynamic environments.
- A novel composite reward function is designed, incorporating a trade-off between theoretical string stability, safety, efficiency, and comfort. This marks the first instance where frequency-domain conditions for string stability are explicitly embedded in a reward function, providing a more balanced and theoretically grounded approach to optimizing CAV longitudinal control.
- A parameter tuning method is proposed to further improve string stability in the IDL strategy. This method adjusts any parameters that fall outside the allowable range, ensuring that the linear controller parameters generated by the DRL agent guarantee the string stability while closely approximating the original values.
- The TD3 algorithm is employed in the proposed strategy as a DRL agent to implement longitudinal control for CAVs. Numerical simulation using real-world data demonstrates that the proposed approach ensures theoretical string stability while significantly enhancing vehicle safety, comfort, and efficiency compared to uncontrolled scenarios.

The rest of this study is organized as follows. Section 2 outlines the research problem. Section 3 introduces the IDL strategy for longitudinal control of CAVs. Section 4 presents the numerical experiment setup and evaluation metrics. Section 5 discusses the simulation results. Finally, Section 6 concludes the work.

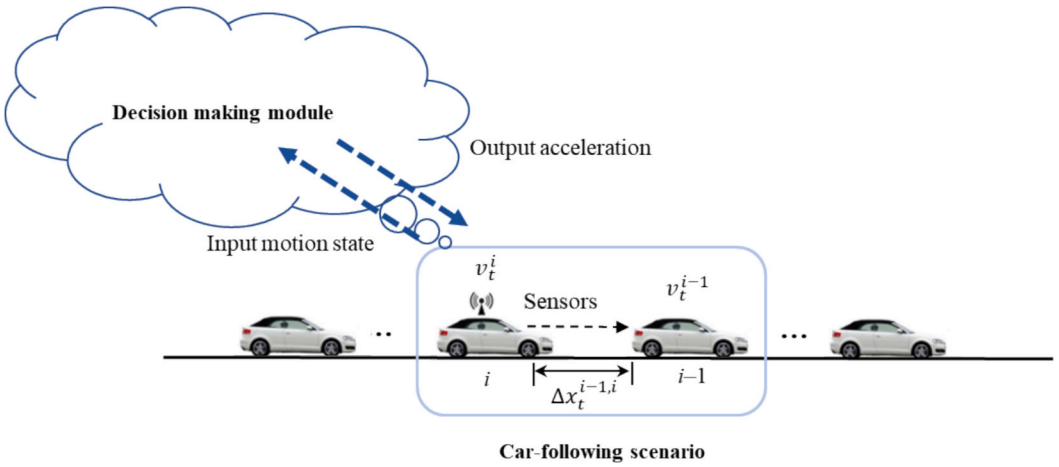


Fig. 1. Illustration of longitudinal motion of CAV in the car-following state.

2. Problem statement

CAV longitudinal motion consists of two states: free-motion state and car-following state. The free-motion state refers to a scenario where the CAV can travel at its desired speed without being influenced by any vehicles in front of it. In contrast, the car-following state imposes constraints on the CAV's motion due to the presence of a preceding vehicle, requiring the CAV to optimize its longitudinal behavior while adhering to constraints on speed and following distance. This makes the car-following state more complex than the free-motion state. The CAV car-following mission is accomplished in steps of perception, decision-making, and control. Initially, the CAV acquires information about the motion of the preceding vehicle and itself through sensors or communications. This information is then forwarded to the decision-making module of CAV to formulate appropriate decisions. Finally, decision signals are transmitted to the control center of the CAV, where controllers execute actions to drive the vehicle.

This study focuses on the car-following state of CAV on the single-line straight highway segment scenarios. In practical applications, when the CAV exits the car-following state, it can continue operating by maintaining a constant maximum speed. We consider a preceding-following information flow topology in which the CAV optimizes its motions according to the information from its preceding vehicle. The objective of this study is to develop a decision-making module to determine the next step acceleration of the CAV at each time step t, given the velocity, acceleration of the preceding vehicle, the velocity of itself, and the relative distance between them in the present time step.

Fig. 1 depicts the longitudinal motion of CAV in the car-following state. In this figure, the i^{th} vehicle is a CAV that can exchange information with other CAVs towards communication machines and detect the movement of its neighbors and itself through on-board sensors. At each time step t, the CAV detects the velocity and acceleration of the $(i\text{–}1)^{\text{th}}$ HDV, $v^{i-1}_{t-\Delta t}$, and $acc^{i-1}_{t-\Delta t}$, as well as its velocity, $v^{i}_{t-\Delta t}$, and the relative distance between them, $\Delta x^{i-1,i}_{t-\Delta t}$. Here, Δt represents the simulation time step. This information is then forwarded to the CAV's decision-making module to calculate the desired acceleration acc^{i}_{t}. Subsequently, the vehicle updates its velocity and relative distance between the preceding vehicle as follows:

$$v^{i}_{t} = v^{i}_{t-\Delta t} + acc^{i}_{t}\Delta t, \tag{1}$$

$$\Delta x^{i-1,i}_{t} = \Delta x^{i-1,i}_{t-\Delta t} + \frac{v^{i-1}_{t} - v^{i}_{t} + v^{i-1}_{t-\Delta t} - v^{i}_{t-\Delta t}}{2}\Delta t. \tag{2}$$

3. Methodology

This section introduces an integrated DRL-linear control (IDL) strategy for the longitudinal control of CAVs. First, the overall framework is presented, followed by a description of the linear controller and the string stability conditions in the frequency domain.

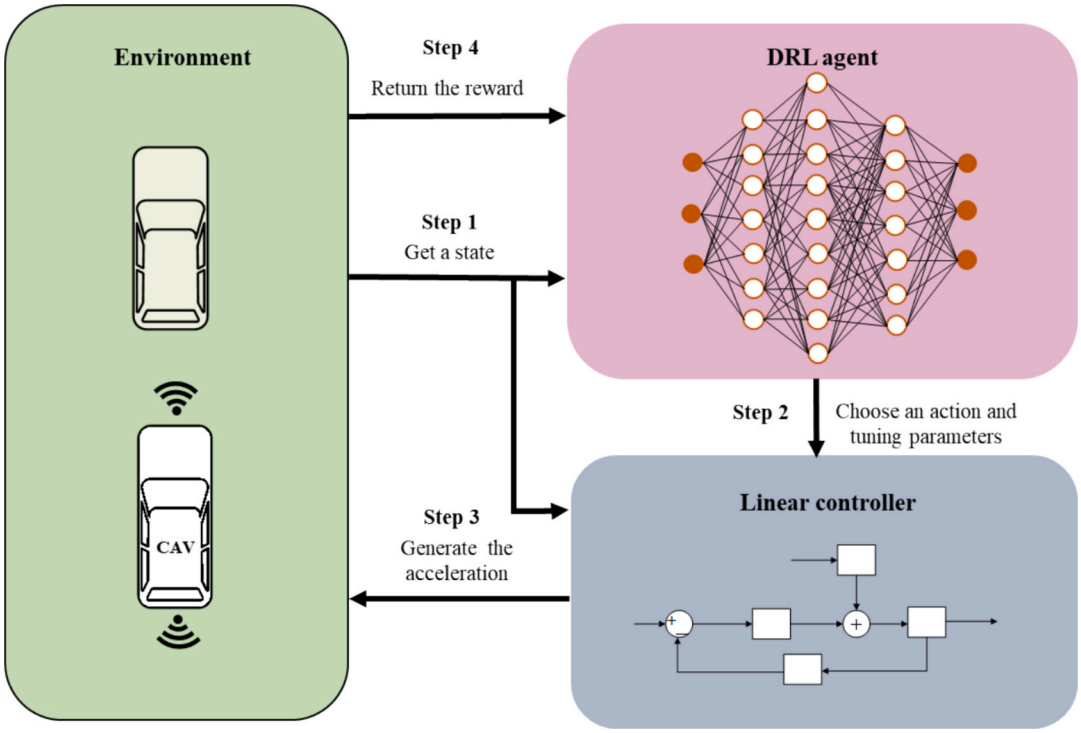


Fig. 2. The IDL strategy diagram.

Next, the critical elements of the IDL strategy are outlined, with particular emphasis on the design of the reward function, which balances multiple objectives such as vehicle safety, comfort, efficiency, and string stability. Additionally, a parameter tuning method is proposed to further improve the string stability of the IDL strategy. Finally, the twin delayed deep deterministic policy gradient (TD3) algorithm is employed in the proposed strategy as a DRL agent to implement CAV longitudinal control.

3.1. Overall framework

The IDL strategy for longitudinal control of CAVs consists of the environment (simulation platform), the DRL agent (learning algorithm for each CAV), and the linear controller. The IDL strategy diagram is shown in Fig. 2. The modules are explained as follows:

(1) **Environment**. The environment refers to the simulated traffic system, providing state information, rewards, and transition dynamics that govern the interaction between the linear controller, the DRL agent, and the system. In this module, the CAV updates its motion based on the acceleration generated by the linear controller, with feedback provided through rewards to guide the DRL agent in optimizing the linear controller's parameters.
(2) **DRL agent**. The DRL agent is a learning algorithm that perceives the environment's state to optimize the parameters of the linear controller. It continuously learns through interactions with both the environment and the linear controller, balancing multiple objectives such as safety, comfort, energy efficiency, and string stability. With its adaptive capabilities, the DRL agent ensures optimal longitudinal control under diverse traffic conditions.
(3) **Linear controller**. The linear controller is a control strategy that adjusts system inputs based on a linear combination of state variables. In the IDL strategy, the linear controller serves as the intermediary between the environment and the DRL agent. It generates the desired acceleration for the CAV based on the outputs from the DRL agent and then forwards this acceleration into the environment, altering the state of the system.

The main steps of the IDL strategy include:

Step 1: The environment's state is provided to both the DRL agent and the linear controller. The DRL agent uses this state to generate an action, while the linear controller uses it to establish a control framework.

Step 2: The DRL agent generates an action, which is then evaluated by the parameter tuning method. If the action does not satisfy the string stability condition, it is revised by the proposed search algorithm to ensure compliance, and vice versa.

Step 3: The linear controller calculates the acceleration for the CAV based on the action from the parameter tuning method and the current state. This acceleration is then transmitted to the environment.

Step 4: The acceleration is applied to the CAV to update its motion. The changed environment provides the reward, which is used by the DRL agent to continuously refine and improve its decision-making process.

3.2. Linear controller and string stability analysis

3.2.1. Linear controller

The linear controller is designed to regulate system inputs using linear relationships to achieve desired outputs. It linearly combines system state variables to generate control signals, enabling precise system regulation. The linear controller is typically structured in two layers, each responsible for different tasks and control objectives. The upper layer, referred to as the strategy layer, is responsible for developing the overall control strategy, such as defining the desired acceleration or distance of the CAV. The lower layer, known as the execution layer, carries out specific control actions, ensuring accurate control based on the instructions provided by the upper layer.

In the upper layer, the desired acceleration of the CAV is planned. In the car-following state, a CAV aims to maintain the same speed and acceleration as the preceding vehicle while achieving the equilibrium distance. Therefore, when designing its desired acceleration, it must consider these three factors simultaneously.

First, the equilibrium distance is determined based on the CAV's current speed. The expression for the equilibrium distance of the i^{th} CAV at time step t, denoted by Δx^*, is given as follows:

$$\Delta x_t^* = v_t^i T^* + l_0, \tag{3}$$

where T^* is the desired time headway, and l_0 is the standstill spacing of the i^{th} CAV.

Then, the CAV's desired acceleration is determined based on the combined effects of the difference between the current and equilibrium distances, the speed difference between the CAV and the preceding vehicle, and the acceleration of the preceding vehicle. The desired acceleration of the i^{th} CAV at time step t, denoted by u_t^i, is represented as follows:

$$u_t^i = K_i Y_t^{iT}, \tag{4}$$

where K_i is the output of the DRL agent, and $K_i = \left[k_{x,i}, k_{v,i}, k_{a,i}\right]$. $k_{x,i}$, $k_{v,i}$, and $k_{a,i}$ represent the coefficients corresponding to each term in Y_t^i, respectively. $Y_t^i = \left[\Delta p_t^{i-1,i},\ \Delta v_t^{i-1,i},\ acc_{t-\delta}^{i-1}\right]$, which is the vector composed of the vehicle motion states received from the environment. Here $\Delta p_t^{i-1,i}$ represents the difference of the vehicle distance at time step t from the equilibrium distance, calculated by $\Delta p_t^{i-1,i} = \Delta x_t^{i-1,i} - \Delta x_t^*$, $\Delta x_t^{i-1,i}$ represents the relative distance between the preceding vehicle and the following vehicle at time step t.

$\Delta v_t^{i-1,i}$ represents the relative velocity between the preceding vehicle and the following vehicle at time step t, and $acc_{t-\delta}^{i-1}$ is the acceleration of the preceding vehicle delayed by δ seconds.

In the lower layer, the widely used linearized third-order state-space (Yue et al., 2024) is adapted to represent the longitudinal dynamics for each CAV. Vehicle dynamics are typically nonlinear, especially in complex traffic environments. Through linearization, the third-order state-space model approximates the nonlinear system as a linear one, simplifying the design and analysis of the controller. This linearized model is effective in capturing the vehicle's dynamic behavior under normal driving conditions. It is expressed as

$$\dot{x}_t^i = v_t^i, \tag{5}$$

$$\dot{v}_t^i = acc_t^i, \tag{6}$$

$$\dot{acc}_t^i = -\frac{1}{\tau_i} acc_t^i + \frac{1}{\tau_i} u_t^i, \tag{7}$$

where x_t^i, v_t^i, and acc_t^i represent the position, velocity, and acceleration of the i^{th} CAV at time step t, respectively. τ_i represents the internal dynamics.

According to the vehicle longitudinal equilibrium and dynamics defined previously, we can represent vehicle movement following the continuous linear time-invariant system as follows:

$$\dot{x}_t^i = M_1 x_t^i + M_2 u_t^i + M_3 acc_t^{i-1}, \tag{8}$$

where $\dot{x}_t^i$ represents the rate of change of the state x_t^i over time in the continuous state space system, $x_t^i = \left[\Delta p_t^{i-1,i}, \Delta v_t^{i-1,i}, acc_t^i\right]^{\mathrm{T}}$ is defined as the system state, and M_1, M_2, and M_3 are the coefficient matrices defined as follows:

$$M_1 = \begin{bmatrix} 0 & 1 & -T^* \\ 0 & 0 & -1 \\ 0 & 0 & -1/\tau_i \end{bmatrix}, \; M_2 = \begin{bmatrix} 0 \\ 0 \\ 1/\tau_i \end{bmatrix}, \; M_3 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}. \tag{9}$$

String stability analysis

The L_p-stability proposed by Ploeg et al. (2014) is a commonly used method in the stability analysis of linear systems, where p represents the norm used (e.g., $p = 1, 2, \infty$.). Among them, the L_2-stability is widely used to analyze the string stability of a system (Montanino et al., 2021; Zhou et al., 2020). The string stability condition ensures that the disturbance experienced by a following vehicle is smaller than that of its preceding vehicle. This principle applies to every pair of adjacent vehicles for any CAV follower. As a result, in CAV platoons of any size, disturbances are progressively attenuated along the vehicle string, since each pair of adjacent CAVs further reduces the magnitude of the disturbance.

Definition 1. (*(Ploeg et al., 2014). A system is L_2 stable if it satisfies the following condition:*)

$$\sup_i \|\Gamma_s^i\|_{H_\infty} \leq 1, \; s \in \mathrm{C}, \tag{10}$$

Where s represents the complex frequency variable in the Laplace transform, and C denotes the set of complex numbers. $\|\Gamma_s^i\|_{H_\infty}$ is the H-infinity norm of the transfer function Γ_s^i. The transfer function of the linear controller introduced above can be calculated according to the study of Zhou et al. (2020) and is expressed as follows:

$$\Gamma_s^i = \frac{k_{x,i} + k_{v,i} s + k_{a,i} s^2 e^{-\delta s}}{\tau_i s^3 + s^2 + \left(\tau_i k_{x,i} + k_{v,i}\right) s + k_{x,i}}. \tag{11}$$

Proposition 1. (*The CAV system exhibits string stability if $\Psi_1 \geq 0$, $\Psi_2 \geq 0$, and $\Psi_3 \geq 0$, or if the discriminant $\Lambda = {\Psi_2}^2 - 4\Psi_1\Psi_3 \leq 0$, according to the study of Zhou et al. (2020) and Ma et al. (2022), where*)

$$\Psi_1 = {\tau_i}^2 + \frac{k_{v,i} k_{a,i} \delta^3}{3}, \tag{12}$$

$$\Psi_2 = 1 - 2\tau_i\left(k_{x,i} + k_{v,i} T^*\right) + k_{a,i}\left(-2k_{v,i}\delta + k_{x,i}\delta^2 - k_{a,i}\right), \tag{13}$$

$$\Psi_3 = k_{x,i}\left(k_{x,i} {T^*}^2 + 2k_{v,i} T^* + 2k_{a,i} - 2\right), \tag{14}$$

$$\Lambda = {\Psi_2}^2 - 4\Psi_1\Psi_3. \tag{15}$$

3.2.2. Critical elements in IDL strategy

The critical elements in the IDL strategy include agent, state, action, and reward.

Agent: In this study, the decision-making module of each CAV is set up as an agent, which can provide different longitudinal motion decision commands to the CAVs. A_i denotes the i^{th} agent, where $1 \le i \le N$, and N represents the number of CAVs in the segment.

State: State is defined as the instantaneous motion state of the CAV and its interaction with the preceding vehicle in the car-following process. The state of the i^{th} CAV at time step t, denoted by s_t^i, includes $\Delta p_t^{i-1,i}$, $\Delta v_t^{i-1,i}$, and $acc_{t-\delta}^{i-1}$. Since the three state variables are on different scales, each element is normalized by the scaling factors α_p, α_v, and α_{acc}, respectively, to bring all variables to a consistent scale. This normalization enhances the DRL agent's ability to identify the optimal action across varying states. The state representation is given as follows:

$$s_t^i = \left[\frac{\Delta p_t^{i-1,i}}{\alpha_p}, \frac{\Delta v_t^{i-1,i}}{\alpha_v}, \frac{acc_{t-\delta}^{i-1}}{\alpha_{acc}} \right]. \tag{16}$$

Action: Action is defined as a decision signal generated for each control cycle that determines the parameters of the linear controller. It is represented as

$$a_t^i := K_i \tag{17}$$

where a_t^i denotes the action of the i^{th} CAV at time step t, and $K_i = \left[k_{x,i}, k_{v,i}, k_{a,i}\right]$ is a vector of the linear controller parameters.

Reward: Reward is a scalar signal provided by the environment to the agent to indicate the immediate outcome of the action the agent has taken. It serves as a form of feedback that reflects how good or bad the agent's action was in achieving the task's objective. In the IDL strategy, the reward function consists of vehicle safety, efficiency, and comfort. The expression for the composite reward function is as follows:

$$r_t^i = \omega_1 r_{f,t}^i + \omega_2 r_{j,t}^i + \omega_3 r_{e,t}^i \ , \tag{18}$$

where $r_{f,t}^i$, $r_{j,t}^i$, and $r_{e,t}^i$, are the safety reward, comfort reward, and efficiency reward of the i^{th} vehicle at time step t, respectively. ω_1, ω_2, and ω_3 are corresponding coefficients, which sum to 1.

(1) Safety reward function

The time to collision (TTC) was proposed by Hayward (1972), and has been widely used to assess the safety risks in microscopic traffic flow (Li et al., 2017). It is defined as the time that a collision would occur if two adjacent vehicles continued to travel at their current velocities. TTC with larger values implies a safer condition. The TTC is calculated as follows:

$$TTC_t^i = \begin{cases} \dfrac{x_t^{i-1} - x_t^i}{v_t^i - v_t^{i-1}} & , \text{if} v_t^{i-1} < v_t^i, \\ \infty & , \text{otherwise}. \end{cases} \tag{19}$$

Since a larger value of TTC indicates a higher level of safety, we should maintain the desired TTC. Therefore, penalties are applied when TTC is lower than a specific value. And the penalty increases as TTC decreases. If TTC ≤ 0, it indicates that a collision has occurred, and a maximum penalty should be applied. The value range of $r_{f,t}^i$ is $[-1, 0]$. The piecewise function is as follows:

$$r_{f,t}^i = \begin{cases} 0, & \text{if } TTC_t^i \in (\text{TTC}^*, \infty), \\ \dfrac{TTC_t^i}{\text{TTC}^*} - 1, & \text{if } TTC_t^i \in (0, \text{TTC}^*], \\ -1, & \text{otherwise}, \end{cases} \tag{20}$$

where TTC^* denotes the safety threshold used for calculating safety metrics.

(2) Comfort reward function

Passenger comfort is commonly evaluated through physical motion parameters, particularly during vehicle acceleration and deceleration. A key metric for this evaluation is the jerk, the rate of change of acceleration. Fluctuations in jerk can strongly influence the perceived abruptness of motion. Higher jerk values can cause sudden jolts, which negatively impact comfort. The square of jerk is adopted to evaluate driving comfort. A smaller jerk squared indicates smoother vehicle motion and thus better comfort; therefore, larger penalties are assigned to higher jerk squared values. The metric is normalized by its maximum value so that its range remains within $[-1, 0]$. During normalization, the term Δt^2 is canceled out and hence is not explicitly included. The comfort reward function is expressed as follows:

$$r_{j,t}^i = \frac{-\left(acc_t^i - acc_{t-1}^i\right)^2}{\left(acc_{\max} - acc_{\min}\right)^2}, \tag{21}$$

where $acc_{\max}$ and $acc_{\min}$ are the maximum and minimum acceleration of CAV, respectively, and $acc_0^i = 0$.

(3) Efficiency reward function

Time headway is a fundamental concept in traffic flow theory, used to quantify the time interval between two consecutive vehicles and commonly applied to evaluate traffic density (Brackstone and McDonald, 1999). It refers to the time difference between when the rear of the preceding vehicle passes a specific point and when the front of the following vehicle reaches that same point. In intelligent transportation systems, reducing time headway can increase road capacity and improve traffic flow efficiency (Arem et al., 2006). The

time headway is defined as follows:

$$h_t^i = \frac{x_t^i - x_t^{i-1}}{v_t^i}, \tag{22}$$

where h_t^i represents the time headway of the i^{th} CAV at time step t.

To enhance efficiency, we aim to minimize the time headway, although excessively small time headway may pose safety risks. To balance this, we adopted the efficiency reward function from the study of Zhu et al. (2020) as the efficiency reward function in this research. Let $F(x)$ denote the probability distribution of the time headway, which is expressed as follows:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-\left(\ln h_t^i - \mu\right)^2}{2\sigma^2} \right), \tag{23}$$

where μ, and σ are the mean and log standard deviation of the h_t^i, and they equal 0.4226 and 0.4365 according to the empirical data, respectively. The efficiency reward is largest when h_t^i reaches 1.26 s, and decreases when it exceeds or falls below this value. The maximum value of $F(x)$ is $F(h^*) \approx 0.6588$. $F(x)$ is then normalized so that its range falls within $[-1, 0]$, and the efficiency reward function is defined as follows:

$$r_{e,t}^i = \begin{cases} \dfrac{F(x)}{\max(F(x))} - 1, \text{if} h_t^i \in (0, \infty), \\ -1, \text{ otherwise.} \end{cases} \tag{24}$$

3.3. Parameter tuning method

To enhance string stability in the IDL strategy, we propose a parameter tuning method. This method assesses the actions generated by the IDL strategy and modifies any actions that do not meet the string stability conditions. The modified actions are then forwarded into the linear controller as the actual output of the DRL agent, ultimately influencing the environment. During the parameter tuning process, the modified actions are kept as close as possible to the original ones while quickly eliminating infeasible regions. To achieve this, we implement the region-growing search method for parameter tuning, which enforces strict string stability constraints on the IDL strategy.

By analyzing Proposition 1, we can establish the relationship between Ψ_1, Ψ_2, Ψ_3, Λ, and the string stability of the system. Since $k_{x,i}$, $k_{v,i}$, $k_{a,i}$, τ_i, and δ are all non-negative, it can be concluded that $\Psi_1 \geq 0$. Referring to the study by Ma et al. (2022), string stability can be classified into the following four types:

Type I: Stability, satisfying the string stability condition of $\Psi_2 \geq 0$ and $\Psi_3 \geq 0$.

Type II: Stability, satisfying the string stability condition of $\Psi_3 \geq 0$, $\Psi_2 < 0$, and $\Lambda \leq 0$.

Type III: Instability, due to $\Psi_3 < 0$.

Type IV: Instability, due to $\Psi_3 \geq 0$, $\Psi_2 < 0$, and $\Lambda > 0$.

If the initially generated values fall into type III or type IV instability, the parameter tuning method works. For type III instability, it is necessary to increase Ψ_3 until $\Psi_3 > 0$. For type IV instability, it is necessary to reduce Λ until $\Lambda < 0$. Therefore, it is essential to analyze the influence of parameter adjustments on Ψ_2, Ψ_3, and Λ. Firstly, the partial derivatives of Ψ_2 and Ψ_3, concerning $k_{x,i}$, $k_{v,i}$, and $k_{a,i}$ are calculated, yielding the following equations. The partial derivative of Λ is too complex, making it difficult to analyze trends based on its partial derivative, so it will not be calculated.

$$\nabla_K \Psi_2 = \begin{pmatrix} \partial\Psi_2/\partial k_{x,i} \\ \partial\Psi_2/\partial k_{v,i} \\ \partial\Psi_2/\partial k_{a,i} \end{pmatrix} = \begin{pmatrix} 1 - 2\tau_i + k_{a,i}\delta^2 \\ -2\tau_i T^* - 2k_{a,i}\delta \\ k_{x,i}\delta^2 - 2k_{v,i}\delta - 2k_{a,i} \end{pmatrix}, \tag{25}$$

$$\nabla_K \Psi_3 = \begin{pmatrix} \partial\Psi_3/\partial k_{x,i} \\ \partial\Psi_3/\partial k_{v,i} \\ \partial\Psi_3/\partial k_{a,i} \end{pmatrix} = \begin{pmatrix} 2k_{x,i}T^{*2} + 2k_{v,i}T^* + 2k_{a,i} - 2 \\ 2k_{x,i}T^* \\ 2k_{x,i} \end{pmatrix}. \tag{26}$$

Based on partial derivatives, two definitive conclusions can be made: increasing $k_{v,i}$ leads to an increase in Ψ_3 and a decrease in both Ψ_2 and Λ, while increasing $k_{a,i}$ consistently increases Ψ_3, irrespective of the values of other parameters.

To stabilize the system, we implement the region-growing search method to explore new values for $k_{x,i}$, $k_{v,i}$, and $k_{a,i}$ when the initially generated parameters result in type III or type IV instability. The region-growing search method is a global search algorithm that progressively expands the search region (Blum and Roli, 2003), enabling a comprehensive exploration of the parameter space. The main steps of the method are outlined as follows:

Step 1: Initialization

– Set the initial search points $k_{x0,i}$, $k_{v0,i}$, and $k_{a0,i}$, which are the original values generated by the DRL agent that lead to type III or type IV instability.

- Define the initial search region R_1, a rectangular area with ranges for $k_{x,i}$, $k_{v,i}$, and $k_{a,i}$ as $\left[k_{x0,i}-r_0, k_{x0,i}+r_0\right]$, $\left[k_{v0,i}-r_0, k_{v0,i}+r_0\right]$, and $\left[k_{a0,i}-r_0, k_{a0,i}+r_0\right]$, respectively, where r_0 represents the initial search step size.
- Set the stopping conditions, including parameter boundaries and the maximum number of searches.

Step 2: Local search.

- Within the initial region R_1, a grid search is conducted.
- For type III instability, the subregion where $k_{x,i} = k_{x0,i}$, $k_{v,i} \in \left[k_{v0,i}-r_0, k_{v0,i}\right]$, and $k_{a,i} \in \left[k_{a0,i}-r_0, k_{a0,i}\right]$ is excluded from the search range and is placed in the region R_N, which represents the area where no further search will be conducted. The rest area of the region R_1 is searched, and the value of Ψ_3 is evaluated. If $\Psi_3 < 0$, type III instability is confirmed, and the search continues. If $\Psi_3 \geq 0$, Ψ_2 and Λ will be calculated. If the solution meets the criteria for type I or type II string stability, the search stops and the solution is output. Otherwise, the search continues.
- For type IV instability, the entire region R_1 is searched, and the value of Λ is assessed. If $\Lambda > 0$, type IV instability is confirmed, and the search continues. If $\Lambda \leq 0$, then Ψ_2 and Ψ_3 will be calculated. If the solution satisfies the conditions for type I or type II string stability, the search stops and the solution is output; otherwise, the search continues.

Step 3: Feasibility check.

- If no feasible solution is found within the region R_1, check whether the maximum search region has been reached. If stopping conditions are met, the search is terminated. Otherwise, continue to expand the search region.

Step 4: Expansion of the search region

- Expand the search region. The k^{th} expanded region R_k is defined as $k_{x,i} \in \left[k_{x0,i}-kr_0, k_{x0,i}+kr_0\right]$, $k_{v,i} \in \left[k_{v0,i}-kr_0, k_{v0,i}+kr_0\right]$, and $k_{a,i} \in \left[k_{a0,i}-r_0, k_{a0,i}+r_0\right]$. If this region exceeds the upper or lower bounds of $k_{x,i}$, $k_{v,i}$, and $k_{a,i}$, the corresponding bounds are enforced.
- The region R_{k-1} is added to R_N, excluding it from further searching.
- Continue the local search in the expanded region R_k to find a feasible solution.

Step 5: Iterative expansion

- Continue expanding the search region and repeat steps 2 to 4 until a feasible solution is found or stopping conditions are met.

Step 6: Termination criteria.

- When a feasible solution is found, output the solution and terminate the search.
- If the search exceeds the predefined maximum range or number of iterations without finding a solution, the initial solution is output along with the message "No optimized solution found".

This approach ensures that the new replacement parameters $k_{x,i}$, $k_{v,i}$, and $k_{a,i}$ remain as close as possible to the original values, minimizing unnecessary changes that could affect system performance. It allows for the fast elimination of invalid regions based on the type of instability, significantly improving the efficiency of the search process.

3.4. TD3 algorithm

The twin delayed deep deterministic policy gradient (TD3) is a policy-based DRL method designed to address policy optimization challenges within continuous action spaces. It has been widely utilized in the autonomous driving domain. TD3 improves upon the deep deterministic policy gradient (DDPG) algorithm (Lillicrap et al., 2015) to overcome the overestimation bias observed in DDPG. The main innovations of TD3 are as follows: (1) double Q-networks: TD3 uses two independent Q-networks and selects the smaller of the two Q-values when calculating the target value. This approach reduces overestimation bias and enhances the stability of the learning process. (2) delayed policy updates: TD3 reduces the frequency of policy network updates, performing updates only after several Q-network updates. This delayed updating strategy reduces instability caused by frequent policy changes. (3) target policy smoothing: TD3 introduces noise into the target policy by adding noise to the target actions and constraining them within the action bounds. This reduces sensitivity to minor environmental changes, improving the robustness of the learned policy.

The TD3 algorithm is employed in this study to optimize the parameters of a linear controller in the IDL strategy. The input to the TD3 agent is the state, as defined in subsection 3.1, representing the observed information from the car-following task. The output is the action, which defines the parameters of the linear controller. The TD3 agent is trained by alternating between updates of the policy network (actor) and the Q-value networks (critics) to improve car-following decisions. The key steps are as follows:

(1) Experience sampling: In a simulated environment, the CAV interacts with the environment based on the current policy, which is controlled by the policy network. During this interaction, the agent collects data on the state, action, reward, and next state, (s_t^i, a_t^i, r_t^i, s_{t+1}^i), which is stored in an experience replay buffer.
(2) Batch update: A batch of samples is randomly selected from the experience replay buffer. The two Q-networks in TD3 estimate the Q-values for the current policy, and the minimum Q-value is used to reduce overestimation bias and compute the target value.
(3) Policy update: The parameters of the Q-value networks are updated by minimizing the loss function. The updated Q-values are then used to optimize the policy network, ensuring that the agent outputs optimal acceleration or speed actions.
(4) Delayed policy updates: The policy network is updated less frequently than the Q-value networks, typically after several Q-network updates, to improve the stability of the training process.
(5) Target policy smoothing: To prevent the policy network from converging to suboptimal solutions, TD3 adds noise to the target actions. This noise reduces the likelihood that the policy overfits to minor environmental fluctuations.

After training, the TD3 agent is applied in real-time, where it outputs control actions based on the current state. These actions are executed via the linear controller to regulate the vehicle's longitudinal behavior.

4. Numerical experiments

This section describes the experimental setup and evaluation metrics. The experimental setup includes the data preparation, simulation environment setup, parameters setting of the IDL strategy, and the selection of the benchmark strategies. Then, the evaluation metrics for string stability, vehicle safety, efficiency, and comfort are defined from both microscopic and macroscopic perspectives to assess the performance of individual vehicles in car-following behavior over short time scales, and the overall performance of all simulated vehicles throughout the entire simulation duration.

4.1. Experimental setup

4.1.1. Data preparation

This study utilizes vehicle trajectory data from the next generation simulation (NGSIM) project on the I-80 freeway in the San Francisco Bay area. The data were collected between 4:00 p.m. and 5:30 p.m. on April 13, 2005, and consisted of vehicle positions and velocities sampled at a frequency of 10 Hz. The car-following pairs of vehicles with a duration longer than 15 s were selected from all trajectory data for the experiments. In total, 1286 car-following pairs were used in the experiments, of which 900 were used for the training dataset and 386 for the testing dataset.

4.1.2. Simulation environment

A digital simulation environment was established on a computer equipped with an Intel (R) Core (TM) i7-9900 CPU @ 3.20 GHz and 16 GB RAM, and compiled using Python 3.8 and Tensor Flow 1.1.0.

During the training process, each following event in the training dataset is used as input in sequence. The real leading vehicle in the

Table 1
Parameters for the DRL agent in the IDL strategy.

Type	Parameter	Description	Value/ Formula
Network hyperparameters	Learning rate A	The learning rate in the actor network	0.001
	Learning rate C	The learning rate in the critic network	0.001
	Discount factor	The discount factor is used in the Bellman equation to prioritize immediate rewards over future rewards	0.90
	Exploration noise	It encourages sufficient exploration of the state–action space and prevents premature convergence	0.15
	Memory capacity	The number of training samples in the replay memory	10,000
	Minibatch size	Number of training cases over which each stochastic gradient descent update is computed	128
	Soft update rate	The rate at which the target networks are updated towards the actor network.	0.01
Architecture parameters	State dimension	Dimension of the state	3
	Action dimension	Dimension of the action	3
	Action bound	Maximum absolute action value	[−7.6,3]
Supplementary parameters	Policy target update interval	Frequency for updating actor and target networks	3
	Output activation	Activation function at the actor output layer	tanh
	Initializer	Weight initialization method	Glorot uniform
	Optimizer	Optimizer used for both actor and critic	Adam
	Loss critic	The loss function for the critic	Mean squared error
	Loss actor	The loss function for the actor (maximize Q)	−mean (Q(s, a))
	Update counter	Internal counter for delayed actor updates	Internal variable

training dataset was used to simulate the leading vehicle for the CAV. The initial position and velocity of the CAV at the start moment are the same as those of the following vehicle in the training dataset. After that, the location and velocity of the CAV at each moment are generated by the control strategy. When a car-following event ends, the simulation switches to the next event. If a collision occurs during the training process, the event is terminated early and switched to the next event. The car-following events in the training dataset were repeatedly input for training to ensure thorough exploration and learning. The training was stopped after 700 non-collision terminations to prevent overfitting.

During the testing process, car-following events in each testing dataset are used to test the trained strategy. The testing dataset is initialized the same as the training dataset, and after each event ends, the simulation switches to the next car-following event. However, unlike the training process, collisions during testing do not terminate the event. The performance of the strategy in the testing dataset will be assessed using evaluation metrics.

4.1.3. Parameters setting

The IDL framework adopts a standard TD3 architecture, consisting of one actor network and two critic networks, each implemented as a fully connected feedforward network. Both the actor and critic adopt a 2×128 MLP structure, comprising two hidden layers with 128 units each and ReLU activations. Table 1 lists the parameters and their values for the DRL agent in the IDL strategy used in this study. Table 2 lists the parameters for the longitudinal control and their values. The parameter sensitivity analysis, including network architecture, exploration noise, and weight settings, is discussed in detail in subsection 5.1.

Notice that TTC* is a widely used safety threshold to distinguish situations where drivers unintentionally enter dangerous conditions from those in which they remain in control. In existing literature, the value of TTC* typically ranges between 2 and 5 s (Li et al., 2014; Minderhoud and Bovy, 2001). According to the findings of Hogema and Janssen (1996), using a TTC* of 4 or 5 s can result in an excessive number of false alarms. Their study concluded that a threshold of 3 s yielded the fewest false alarms, while still effectively capturing the most critical situations. Based on this, a 3-second TTC* threshold has been widely recognized as an adequate compromise for identifying dangerous approach scenarios (Minderhoud and Bovy, 2001). Consequently, we adopt TTC* = 3 s in this study. In addition, Shi et al. (2018) applied TIT to evaluate pre-crash risk levels. In their study, six TTC* thresholds ranging from 1.5 to 4 s were examined. While higher TTC* values demonstrated greater sensitivity in detecting risk conditions, the authors noted that the impact of threshold variation was not critical in distinguishing high-risk scenarios.

Our study similarly found that slightly stricter or looser TTC* thresholds did not significantly alter the overall safety outcomes. This is attributed to the design of the safety reward function, which is directly tied to TTC* and guides the agent to minimize occurrences where TTC $<$ TTC*. We experimented with TTC* values between 2 s and 5 s within the IDL strategy and compared the resulting safety metrics against HDV baselines. While the safety metrics for the IDL strategy naturally increase as TTC* increases—since the same dataset is evaluated under a more stringent safety criterion—the relative improvement of the IDL strategy over the HDV baseline remains consistent. Moreover, other performance metrics, including string stability, efficiency, and comfort, exhibit minimal sensitivity to changes in TTC*, further confirming the robustness of the proposed strategy under various threshold settings. In summary, although different TTC* values may slightly influence the magnitude of safety metrics, the overall performance trends and conclusions remain stable, validating the reliability of the chosen 3-second threshold.

To evaluate the learning process, we plot the rolling reward curves for the IDL strategy, the pure TD3 strategy and the pure DDPG strategy throughout training, as shown in Fig. 3. The rolling reward is defined as the mean reward over a moving window: if the total number of training episodes exceeds 10, it represents the mean reward over the most recent 10 episodes; otherwise, it reflects the cumulative mean reward up to the current episode. This metric provides a smoothed representation of the learning progression and stability. The results indicate that both strategies successfully converge during the training process. However, the IDL strategy

Table 2
Parameters for the longitudinal control.

Parameter	Description	Value
$acc_{\max}$	The upper bound of acceleration	3 m/s^2
$acc_{\min}$	The lower bound of acceleration	−7.6 m/s^2
α_p	Scaling factor	25
α_v	Scaling factor	2.5
α_{acc}	Scaling factor	4.5
δ	Preceding vehicle information delay	0.2 s
$k_{x,i}$	Coefficient of the linear controller related to the relative distance	[−2, 2]
$k_{v,i}$	Coefficient of the linear controller related to relative velocity	[−2, 2]
$k_{a,i}$	Coefficient of the linear controller related to acceleration	[−2, 2]
p_e	Penalty value	−1
l_0	Standstill spacing between vehicles	2 m
TTC*	The threshold for safety metrics	3 s
T^*	The desired time headway for the linear controller	1.1 s
$v_{\min}$	The lower bound of velocity	0.001 m/s
$v_{\max}$	The upper bound of velocity	33.3 m/s
ω_1	Coefficient of the safety reward	1/3
ω_2	Coefficient of the comfort reward	1/3
ω_3	Coefficient of the efficiency reward	1/3

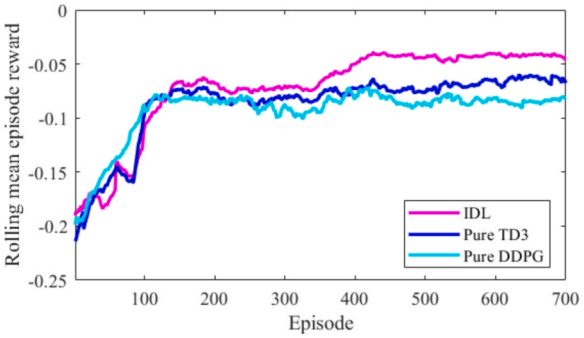


Fig. 3. Rolling reward over training episode (window size = 10).

converges more rapidly and consistently achieves higher reward values compared to the pure TD3 and pure DDPG strategies. Overall, the learning curves demonstrate that all DRL-based strategies exhibit effective learning behavior and achieve convergence.

4.1.4. Benchmark methods

The strategy proposed in this study develops string stability reward functions at the theoretical level. However, most of the existing DRL-based longitudinal control strategies do not incorporate the analyzed string stability condition into the reward function. These algorithms are significantly different from the proposed strategy in this study in terms of structure, output, and string stability evaluation metrics. This study will be compared with the following benchmark strategies:

(1) **Human-driven vehicle (HDV) data**: The HDV data is sourced from the NGSIM dataset, representing the original follower in each car-following pair. This serves as a baseline to compare the traffic flow benefits between CAVs and HDVs.
(2) **Model-based cooperative adaptive cruise control strategy (CACC strategy)**: This study employs the CACC controller proposed by Arem et al. (2006), with fixed controller parameters. In addition, a δ-second delay was introduced to the acceleration of the preceding vehicle to ensure consistency with the experimental setup adopted in this study. Since the strategy proposed in this study combines the DRL approach with a linear controller, selecting the CACC strategy enables us to compare the benefits of the integrated DRL-linear controller with the fixed-parameter linear controller.
(3) **DDPG-based CAV longitudinal control strategy (pure DDPG strategy)**: The DDPG-based CAV longitudinal control strategy proposed by Zhu et al. (2020) is used as one of the benchmarks in this study. In this research, the pure DDPG strategy has been shown to achieve a good balance in terms of safety, comfort, and efficiency, outperforming MPC-based ACC strategies. To ensure a fair comparison, we adopted the optimization framework of DDPG while replacing its original reward function with our proposed one. The purpose of selecting this strategy is to compare the proposed approach with another well-performing DRL method, thereby evaluating the effectiveness of the proposed strategy.
(4) **TD3-based CAV longitudinal control strategy (pure TD3 strategy)**: The pure TD3 strategy adopts a reward function identical to that of the IDL strategy, which is also consistent with the pure DDPG strategy. Notice that neither the pure TD3 nor the pure DDPG strategy includes a linear controller, and each generates an action with only one parameter — the acceleration of the CAV, while the IDL strategy produces three parameters. By comparing the performance of the pure TD3 strategy with that of the IDL strategy, it can be assessed whether integrating a linear model affects the performance of the pure TD3 algorithm.

4.2. Evaluation metrics

4.2.1. String evaluation metrics

In frequency domain analysis, the string stability criterion is determined by the relationship between inputs and outputs. Since the relationship between vehicle acceleration and motion states is clearly defined in both the IDL and CACC strategies, the string stability conditions of these two strategies can be explicitly evaluated based on Proposition 1. We define the string stability ratio, M_{fss}, to evaluate the frequency domain string stability, which is the ratio of the duration during which all vehicles satisfy the string stability conditions to the total simulation duration. A higher value of M_{fss} indicates that more vehicles satisfy the string stability conditions for a longer period, implying better overall string stability for the strategy. The expression is as follows:

$$M_{\text{fss}} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} m_{\text{fss},t}^{i}}{N \bullet T/\Delta t} \times 100\%, \tag{27}$$

where N is the total number of vehicles in the simulation, T is the simulation time for each vehicle. and $m_{\text{tss},t}^{i}$ represents the number of string stability moments, calculated by

$$m_{\text{fss},\,t}^{i} = \begin{cases} 1, & \text{if the string stability conditions are satisfied,} \\ 0, & \text{otherwise.} \end{cases} \tag{28}$$

However, the situation is different for the three baseline strategies: HDV, pure DDPG, and pure TD3. For the pure DDPG and pure TD3 strategies, where the output is acceleration and the input is the environment state, the complex relationship between input and output cannot be clearly expressed using a well-defined mathematical relation. Therefore, evaluating string stability for the pure DDPG and pure TD3 strategies is challenging. As for the NGSIM dataset, they are generated by drivers exhibiting random behaviors (Gunter et al., 2020), and thus, modeling and calibrating vehicle following behavior may lead to significant errors. Consequently, this study does not use M_{fss} to analyze the string stability of HDV, pure DDPG, and pure TD3 strategies. Instead, the study adopts the time domain statistical metric, the acceleration L_2 norm cumulative damping ratio, m_p^i, as proposed by Shi et al. (2021), to approximate the string stability. This metric evaluates the propagation of fluctuations in vehicle acceleration in the time domain. The expression is as follows:

$$m_p^i = \sqrt{\frac{\sum_{t=1}^{T} {acc_t^i}^2}{\sum_{t=1}^{T} {acc_t^{i-1}}^2}}. \tag{29}$$

We define M_{tss} to evaluate the average m_p^i across multiple vehicles, with the expression as follows:

$$M_{\text{tss}} = \frac{\sum_{i=2}^{N} m_p^i}{N-1}. \tag{30}$$

Among the four metrics mentioned above, m_p^i and $m_{\text{fss},\,t}^i$ are microscopic metrics, representing the fluctuation transmission of a vehicle over a period of time, and whether the parameters of a vehicle satisfy the string stability conditions at a specific moment, respectively. M_{tss} and M_{fss} are macroscopic metrics, representing the overall fluctuation transmission across all vehicles during the simulation time, and the overall performance of all vehicles meeting the string stability conditions throughout the entire simulation time, respectively.

4.2.2. Safety evaluation metrics

The safety evaluation metrics are represented by the values of TTC and time integrated time-to-collision (TIT), with higher values indicating greater risk. TTC is a microscopic safety metric used to observe the safety of the vehicle at each moment. TIT (Minderhoud and Bovy, 2001) is a macroscopic safety metric that is used to quantify the severity of different TTC values below the threshold TTC^*. The TTC is expressed as Eq. (19), and the formula of TIT is as follows:

$$TIT = \sum_{t=1}^{T} \left(\text{TTC}^* - TTC_t^i\right) \bullet \tau \bullet \vartheta_t, \tag{31}$$

$$\vartheta_t = \begin{cases} 1, & \text{if} 0 \le TTC_t^i \le \text{TTC}^*, \\ 0, & \text{otherwise}, \end{cases} \tag{32}$$

where the coefficient ϑ_t determines whether the TTC is below the threshold TTC*, and τ is the time interval.

4.2.3. Comfort evaluation metrics

The comfort evaluation metrics are represented by the square of acceleration change rate (jerk), *ci*, and the mean squared jerk, M_{CI}, with lower values indicating greater comfort. The *ci* is a microscopic comfort metric used to observe the instantaneous comfort of each vehicle at each moment, while M_{CI} is a macroscopic comfort metric reflecting the overall average comfort performance of all vehicles during the simulation period. They are expressed as follows:

$$ci_t^i = \left(\frac{acc_t^i - acc_{t-\Delta t}^i}{\Delta t}\right)^2, \tag{33}$$

$$M_{\text{CI}} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} ci_t^i}{N \bullet T}. \tag{34}$$

4.2.4. Efficiency evaluation metric

The efficiency evaluation metrics are represented by the time headway, h_t^i, calculated in Eq. (22), and the average time headway, M_{effi}. Lower values of h_t^i and M_{effi} indicate greater efficiency. The h_t^i is a microscopic metric that represents the performance of one vehicle at each moment, while M_{effi} is a macroscopic metric that represents the performance of all the vehicles during the simulation time on their average level, and is expressed as:

$$M_{\text{effi}} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} h_t^i}{N \bullet T/\Delta t}. \tag{35}$$

5. Results and discussions

This section presents a sensitivity analysis of network architecture, exploration noise, and weight settings. In addition, the CAV vehicle following behavior based on the IDL strategy and four benchmark methods is compared. The string stability, efficiency, safety, and comfort of the five strategies are analyzed from both micro and macro perspectives, respectively. The macro-analysis assessed the

overall performance of the five strategies in terms of each benefit throughout the entire simulation period, while the micro-analysis evaluated the performance of single-vehicle car-following behavior over short time intervals for each strategy. Finally, an analysis of the overall improvement in the IDL strategy relative to the four benchmark strategies is conducted.

Sensitivity analysis of network architecture, exploration noise, and weight settings.

(1) Network architecture selection

To ensure fairness and provide a systematic justification, we conducted additional experiments across four representative neural network architectures of increasing complexity: small (2 × 64 MLP, lightweight baseline), medium (2 × 128 MLP, moderate baseline; adopted in this study), large (2 × 256 MLP, RLlib default), and extra large (3 × 256 MLP, Acme default). We evaluated the performance of these architectures in terms of the reward convergence curves. Based on this analysis, the network parameters achieving the best overall performance were selected. All experiments were performed under a noise level of $\sigma = 0.2$, and the corresponding reward convergence curves are presented in Fig. 4.

The results indicate that network size has a clear impact on training stability and overall performance. The small network converges stably but achieves relatively lower final rewards, suggesting limited representation capacity. In contrast, larger networks exhibit higher peak rewards but also stronger oscillations during training, implying increased sensitivity to noise and potential overfitting. A moderately sized network achieves a favorable balance between stability and expressiveness — it maintains smooth convergence, reaches near-optimal rewards, and avoids excessive fluctuations observed in larger architectures. Therefore, the 2 × 128 MLP was selected as the default configuration, as it provides stable learning dynamics and reliable performance without unnecessary model complexity.

(2) Noise exploration analysis.

To investigate the influence of exploration noise, we performed a comprehensive sensitivity study on the standard deviation of the exploration noise (σ) for both IDL and TD3 strategies. The analysis covered $\sigma \in \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$. Each model was trained to convergence under every noise level, and the performances were compared in terms of string stability, safety, comfort, and efficiency. The corresponding results are shown in Fig. 5.

Since all evaluation metrics are minimized, the optimal performance corresponds to achieving relatively low values across all four indicators. The results demonstrate that the IDL strategy attains its best overall performance at $\sigma = 0.15$, whereas the TD3 strategy performs best at $\sigma = 0.25$. These values were therefore adopted for subsequent experiments to ensure fair and comparable exploration capacities between the two methods. By explicitly tuning σ for each agent, we eliminate potential biases arising from differences in action-space scaling and establish a more rigorous baseline comparison.

(3) Weight tuning and Pareto front exploration

To further explore the Pareto front, we varied the reward weights among safety, comfort, and efficiency objectives. Specifically, we compared the balanced weighting scheme against configurations with stronger or weaker biases toward a particular objective, and examined their impacts on stability, safety, comfort, and efficiency performance. It is worth noting that string stability is not explicitly included in the reward function—it is inherently embedded within the IDL framework. Nevertheless, because string stability is an essential indicator for evaluating longitudinal control, we reported it as an outcome metric to provide a complete view of the Pareto front, as summarized in Table 3.

The results show that adjusting the reward weights shifts the outcomes toward the emphasized objective, though the magnitude of these shifts is moderate, suggesting that the IDL framework remains robust across different trade-off preferences. These findings confirm the Pareto consistency of the results: emphasizing one objective improves its corresponding metric while marginally reducing others. Given the moderate shifts and the need for balanced real-world performance, a neutral trade-off scheme [1/3,1/3,1/3] was adopted in subsequent experiments. Importantly, under this revised configuration, IDL continues to exhibit clear advantages in maintaining string stability while balancing safety, comfort, and efficiency effectively.

5.1. String stability analysis

Model-free strategies such as HDV, pure DDPG strategy, and pure TD3 strategy cannot be evaluated analytically and can only be

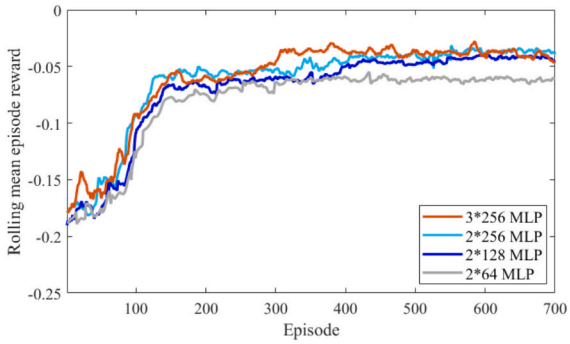


Fig. 4. Comparison of reward convergence across MLP architectures.

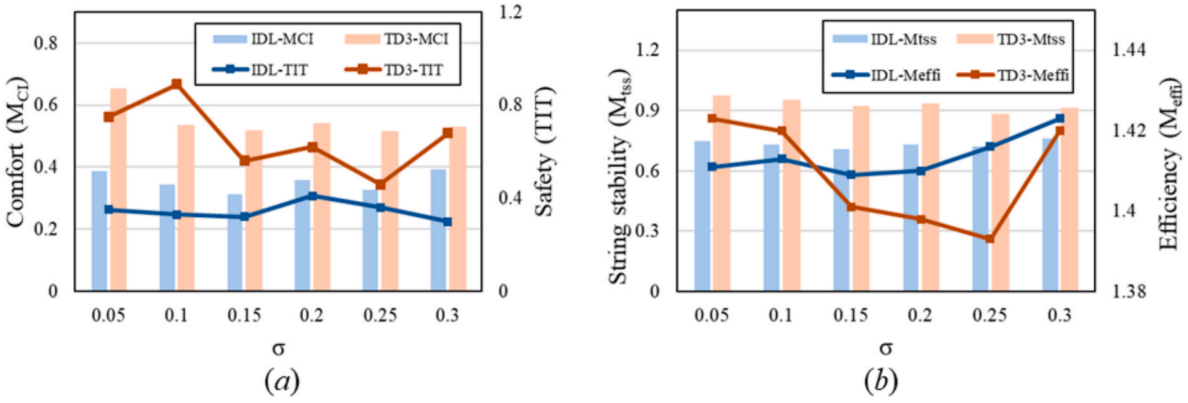


Fig. 5. Sensitivity analysis of exploration noise (σ) on the performance of IDL and TD3 strategies.

Table 3
Performance metrics under different reward weight settings for Pareto front exploration.

Experiment	$[\omega_1, \omega_2, \omega_3]$	TIT	M_{CI}	M_{effi}	M_{tss}
Tradeoff	[1/3, 1/3, 1/3]	0.32	0.31	1.41	0.71
Mild safety bias	[0.4, 0.3, 0.3]	0.29	0.35	1.42	0.72
Strong safety bias	[0.6, 0.2, 0.2]	0.19	0.41	1.45	0.74
Mild comfort bias	[0.3, 0.4, 0.3]	0.37	0.32	1.43	0.71
Strong comfort bias	[0.2, 0.6, 0.2]	0.46	0.30	1.42	0.70
Mild efficiency bias	[0.3, 0.3, 0.4]	0.31	0.35	1.40	0.72
Strong efficiency bias	[0.2, 0.2, 0.6]	0.34	0.38	1.39	0.75

approximated by using M_{tss}. Both the CACC and IDL strategies, however, can be assessed not only through M_{tss} but also by the more precise metric, M_{fss}. The values of M_{tss} for the five strategies are as follows: 0.71 for the IDL strategy, 0.94 for the pure DDPG strategy, 0.88 for the pure TD3 strategy, 1.01 for the CACC strategy, and 1.04 for the HDV. It is evident that only the M_{tss} for the IDL strategy, pure TD3 strategy, and pure DDPG strategy are less than 1, indicating that these strategies are string stable, and other strategies are string unstable. And the M_{tss} for the pure DDPG strategy is very close to 1, suggesting a weak ability to dissipate disturbances and placing it on the verge of string instability. In contrast, the M_{tss} for the IDL strategy is significantly below 1, indicating a stronger capacity to dissipate disturbances. Additionally, the M_{fss} for the IDL strategy is 100%, further confirming that the IDL strategy is string stable. The IDL strategy is the only one that meets string stability conditions based on both frequency domain theoretical derivation and time domain statistical analysis.

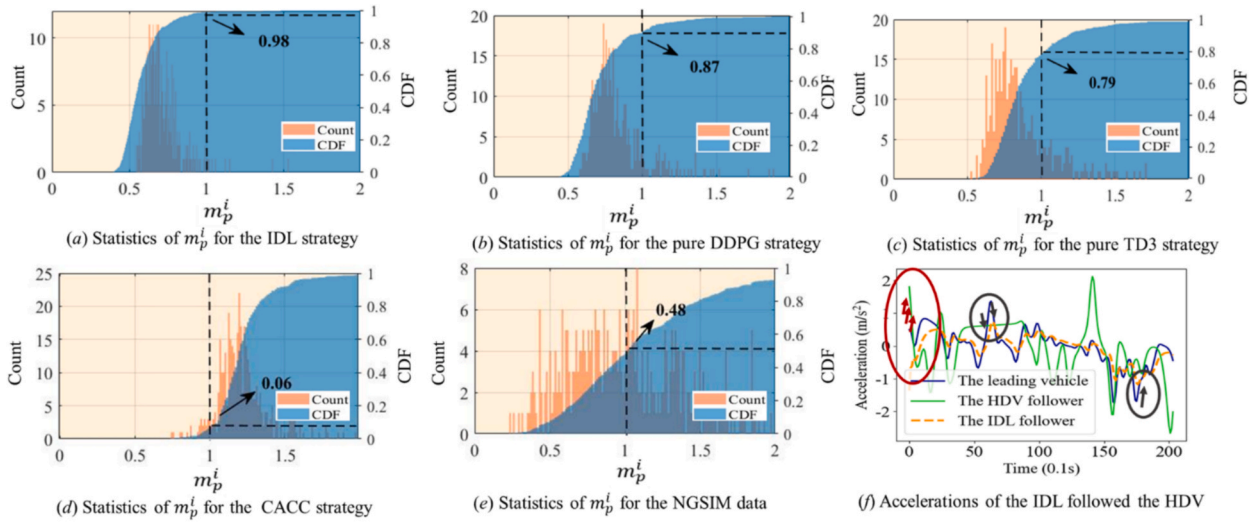


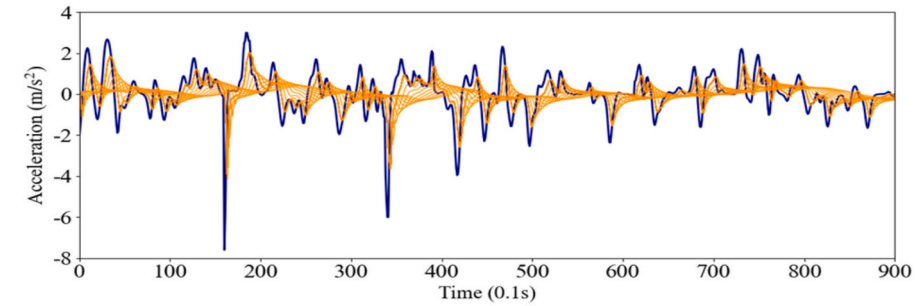
Fig. 6. The transmission of fluctuations in each episode across five strategies.

Fig. 6 illustrates the distribution of m_p^i values across five strategies within the testing dataset. The orange vertical bars represent the frequency of different m_p^i values, while the blue-shaded area indicates the cumulative distribution function (CDF) of m_p^i. When $m_p^i < 1$, the cumulative squared acceleration of the following vehicle is lower than that of the leading vehicle over the testing duration, suggesting a reduction in overall fluctuation during transmission. Conversely, $m_p^i > 1$ indicates an increase in fluctuation.

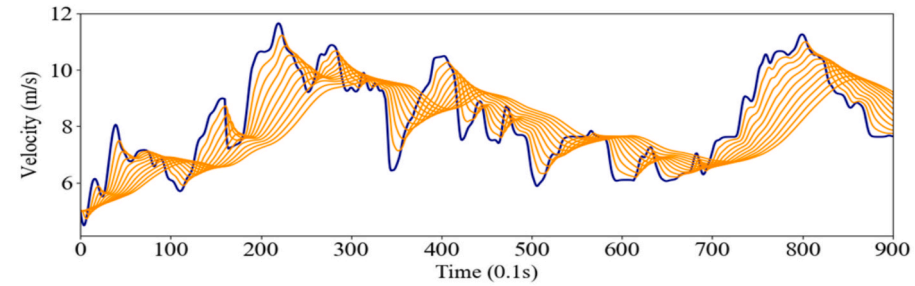
As shown in Fig. 6, the IDL strategy achieves the highest proportion of $m_p^i < 1$, accounting for 98% of the dataset. This is followed by the pure DDPG and pure TD3 strategies, with proportions of 87% and 79%, respectively. The HDV strategy shows 48%, while the CACC strategy, which does not satisfy string stability conditions, shows only 7%. The proportion of $m_p^i < 1$ serves as an indicator of string stability to some extent, although some bias may be present, as the experiments were conducted under unbalanced traffic flow conditions. Consequently, the CACC strategy, despite not meeting string stability conditions, still shows 6% of cases with $m_p^i < 1$, while the IDL strategy, despite meeting these conditions, has instances where $m_p^i > 1$.

Instances of $m_p^i > 1$ are mainly due to initial state imbalances. In Fig. 6 (f), the red circled area highlights cases where initial imbalances caused the following vehicle's acceleration fluctuations to exceed those of the leading vehicle, often due to a large initial gap that prompts the following vehicle (under the IDL strategy) to accelerate quickly, temporarily increasing fluctuations. However, in later states, the following vehicle's fluctuations decrease to be lower than that of the leading vehicle, as shown in the black circled area. As a result, the proportion of $m_p^i < 1$ serves as a partial indicator of string stability, as it reflects fluctuation dissipation to some degree but does not fully capture all aspects of stability. This limitation underscores the value of theoretical string stability conditions, which provide a more precise assessment of stability compliance.

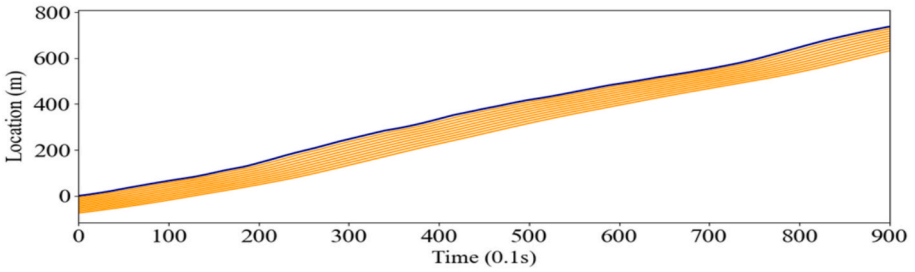
Nonetheless, the proportion of $m_p^i < 1$ offers a clear visualization of fluctuation dissipation in the time domain. The results show that the IDL-based CAV effectively mitigates transmitted perturbations compared to other algorithms, as evidenced by the m_p^i metric in



(a) Accelerations of the CAV platoon based on the IDL strategy



(b) Velocities of the CAV platoon based on the IDL strategy



(c) Locations of the CAV platoon based on the IDL strategy

Fig. 7. The performance of the CAV platoon based on the IDL strategy.

the time domain.

To demonstrate string stability, the motion performance of a CAV platoon based on the IDL strategy following the HDVs is depicted in Fig. 7. A platoon consisting of 10 CAVs follows the leading vehicle. The leading vehicle trajectory is synthesized from multiple representative driving profiles, covering a wide range of acceleration and deceleration behaviors. The new scenarios incorporate abrupt braking events of varying severity as well as sudden accelerations, with the maximum acceleration set to 3 m/s^2 and the maximum deceleration to -7.6 m/s^2. This design enables the evaluation of platoon performance under a broader spectrum of traffic disturbances.

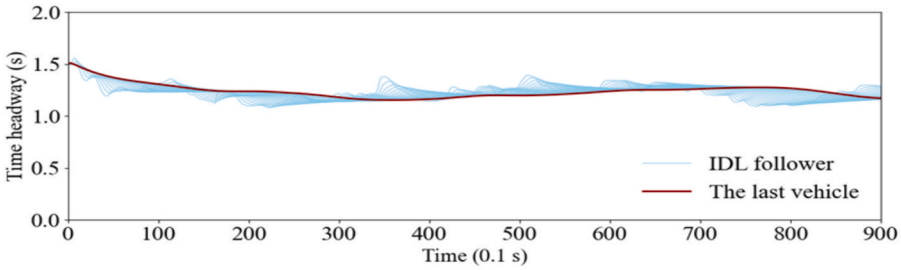
As shown in Fig. 7, it can be observed that the magnitude of accelerations and decelerations converges progressively along the vehicle sequence, and the platoon remains stable even under sudden braking and acceleration. We further report the m_p^i of each vehicle relative to its predecessor as: (0.45, 0.53, 0.61, 0.68, 0.73, 0.77, 0.81, 0.83, 0.85, 0.86). It is worth noting that the platoon is string stable whenever $m_p^i < 1$. The m_p^i increase gradually along the platoon, which means that the improvement effect diminishes downstream. However, the system still maintains disturbance attenuation. This conclusion still holds when the platoon is extended to 100 vehicles. In such cases, the m_p^i increases more gradually: it reaches 0.90 at the 15^{th} vehicle, 0.97 at the 39^{th} vehicle, and then slowly approaches 0.98, fluctuating between 0.978 and 0.989 for the rest of the vehicles. All values remain below 1.

This phenomenon can be explained by the fact that as the platoon size grows, the disturbances have already been largely dissipated by the preceding vehicles. Therefore, the marginal improvement achieved by the trailing vehicles becomes smaller. This further validates the stability of the proposed strategy, as it avoids overcorrection in response to sufficiently small disturbances and therefore prevents disturbance amplification in large-scale platoons.

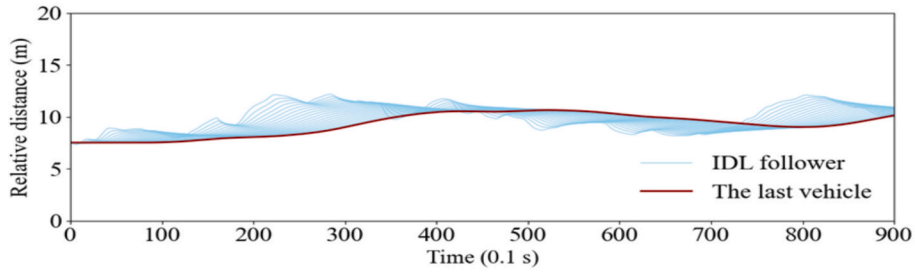
The time headway of each vehicle relative to its predecessor and the inter-vehicle distance between consecutive vehicles over time are depicted in Fig. 8. As shown, the time headway of all vehicles consistently remains above 1.0 s, and the safety indicator (TIT) remains zero, confirming that the platoon operates safely under the tested scenario. Although smaller spacings are occasionally observed, these occur when vehicle speeds are relatively low and the TTC remains within the safe threshold. In such cases, the reduced spacing does not compromise safety. For instance, before 30 s, the spacing between the last vehicle and its predecessor is smaller than that of the others because its speed is lower; however, its time headway is correspondingly larger, indicating that it still requires more time to reach the preceding vehicle. Conversely, between 50 – 70 s, the opposite pattern is observed. Nevertheless, in both cases, all indicators remain within the safe range, demonstrating stable and safe platoon operation.

5.2. Safety analysis

TIT values under hazardous conditions for each strategy are as follows: 0.32 (IDL strategy), 0.60 (pure DDPG strategy), 0.46 (pure TD3 strategy), 0.63 (CACC strategy), and 9.13 (HDV). Lower TIT values correspond to greater safety. These results indicate that the IDL strategy achieves the highest level of safety, significantly improving safety compared to both linear models and DRL strategies, with a substantial enhancement over HDV. Lower TIT value corresponds to greater safety. These results indicate that the IDL strategy achieves



(a) Time headway of the CAV platoon based on the IDL strategy



(b) Relative distance of the CAV platoon based on the IDL strategy

Fig. 8. The time headway and inter-vehicle distance of the CAV platoon based on the IDL strategy.

the highest level of safety, significantly improving safety compared to both linear models and DRL strategies, with a substantial enhancement over HDV. According to the research from Minderhoud and Bovy (2001), when simulating a 9000-second traffic flow involving approximately 10,000 vehicles, the probability of a randomly selected vehicle encountering unsafe conditions was approximately 2.2×10^{-7} for the scenario that TET (time exposed time-to-collision, which measures the duration of dangerous moments when TTC is below the threshold TTC*) equals 19.5 s, indicating an extremely safe traffic environment. In our study, the total simulation time was approximately 9300 s, involving 386 vehicles. Consequently, the average probability of a vehicle experiencing unsafe conditions was 2.8×10^{-8} for the scenario that TET equals 0.6 s (IDL strategy), and 2.8×10^{-7} for the scenario that TET equals 1.2 s (CACC strategy). Both values indicate a high level of safety. Furthermore, both lower TIT and lower TET values are indicative of safer traffic conditions, as supported by Hirst and Graham (2020). Our results demonstrate that the IDL strategy reduces exposure to unsafe conditions by an order of magnitude compared to baseline strategies, highlighting its substantial safety advantage.

Fig. 9 shows the TTC distribution for vehicles under different strategies at each time step, with TTC values ranging from 0 to 100 s. It can be observed that the HDV distribution has a sharper peak, with the highest TTC around 10 s, whereas other strategies display a greater number of higher TTC values. Although safety metrics are only calculated when TTC is below a specified threshold, larger TTC values are generally more favorable for driving safety. Consequently, all other strategies exhibit better safety performance compared to HDV. Fig. 9(f) compares the TIT values across different strategies, clearly demonstrating that the IDL strategy exhibits superior safety performance.

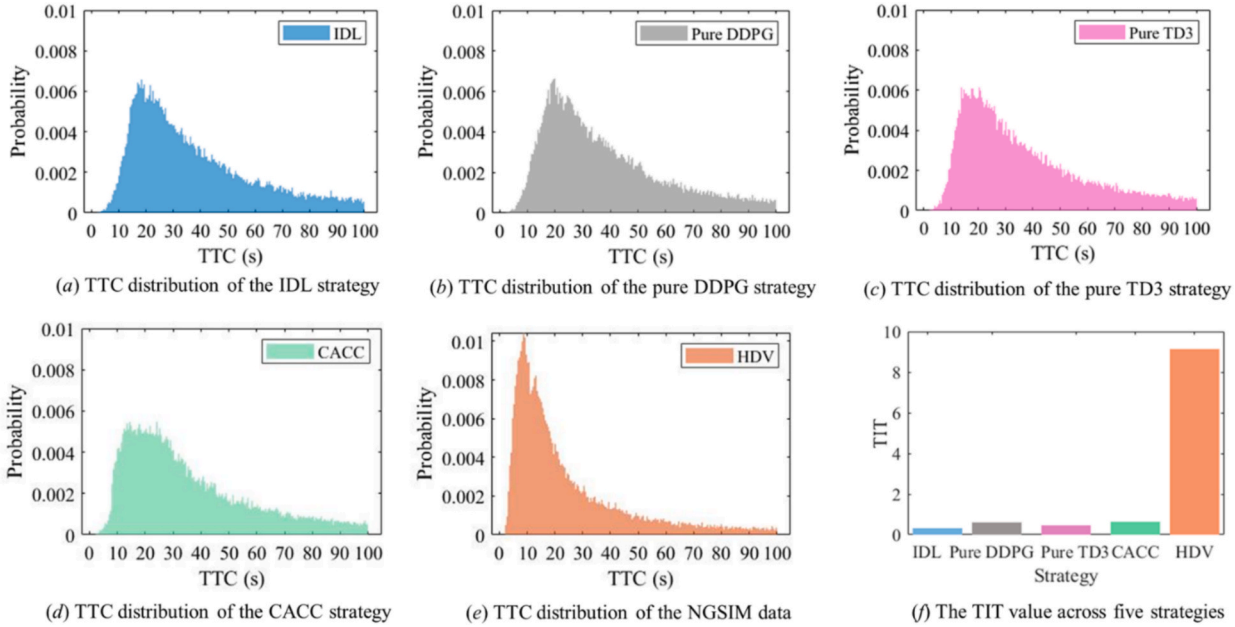
5.3. Comfort analysis

The comfort index M_{CI} values for the five strategies are as follows: 0.31 m/s^6 for the IDL strategy, 0.54 m/s^6 for the pure DDPG strategy, 0.51 m/s^6 for the pure TD3 strategy, 0.84 m/s^6 for the CACC strategy, and 0.68 m/s^6 for the HDV. Overall, all CAV strategies, except for the CACC strategy, improved comfort compared to the HDV, with the IDL strategy achieving the greatest improvement.

To further explore the comfort performance of these five strategies, Fig. 10 illustrates the total number of *ci* values for the following vehicle during the simulation period for each strategy, with the horizontal axis representing the *ci* sequence and the vertical axis representing the *ci* values. It can be observed that, compared to other strategies, the IDL strategy exhibits some larger *ci* values while achieving the mean squared jerk M_{CI}, indicating that the IDL, the pure DDPG, and the pure TD3 strategies minimize M_{CI} by allowing occasional larger changes in acceleration, thereby reducing continuous minor acceleration changes, while also maintaining high efficiency and safety.

In Fig. 11, *ci* values across different ranges are further analyzed, allowing for a clearer comparison of each strategy's performance across various *ci* value ranges. Fig. 11 categorizes *ci* values into four ranges and counts the number of *ci* values in each range for different strategies. The four ranges are: [0, 0.1], [0.1, 1], [1, 10], and [10, 60]. As observed in the figure, the three deep reinforcement learning-based strategies—IDL, pure DDPG, and pure TD3—exhibit similar patterns. These strategies show a higher count of *ci* values within [0, 0.1] (especially within [0, 0.01]) and [10, 60], while exhibiting fewer *ci* values in other ranges. In contrast, the CACC and HDV strategies show a higher count of *ci* values within [0.1, 1] and [1, 10].

This phenomenon suggests that CAVs based on DRL strategies perform more refined control, enabling finer acceleration adjust-



(a) TTC distribution of the IDL strategy (b) TTC distribution of the pure DDPG strategy (c) TTC distribution of the pure TD3 strategy

(d) TTC distribution of the CACC strategy (e) TTC distribution of the NGSIM data (f) The TIT value across five strategies

Fig. 9. Distribution of TTC for five strategies.

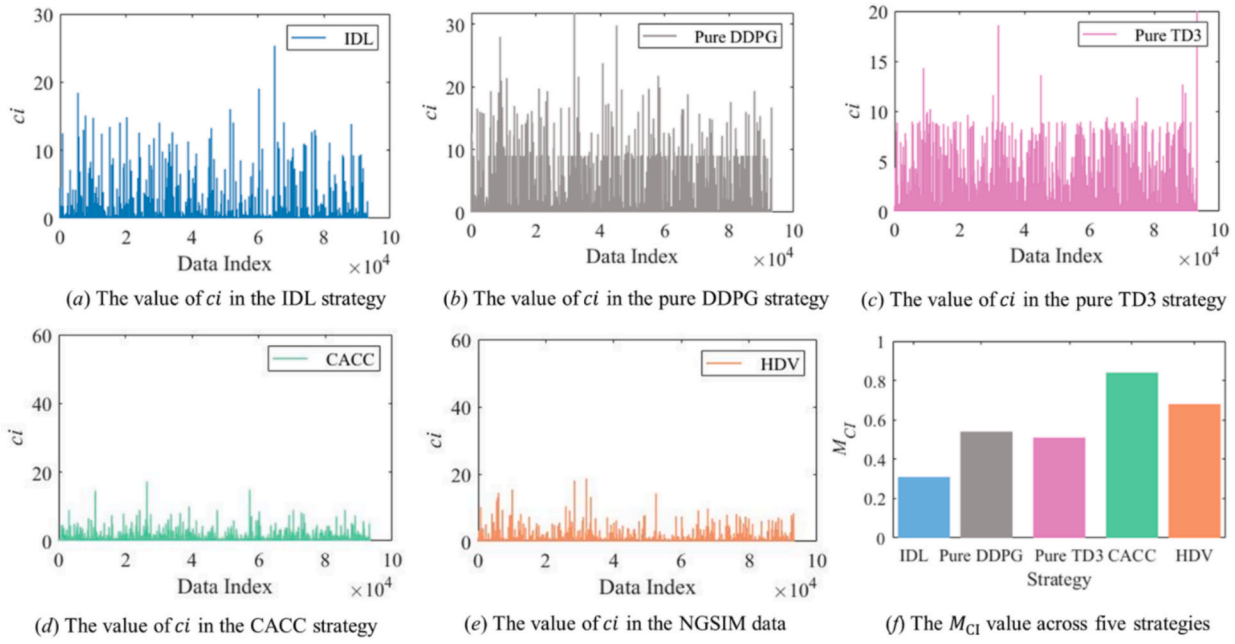


Fig. 10. The value of *ci* during the simulation period for five strategies.

ments within smaller ranges. These three strategies tend to avoid generating the medium *ci* values, which significantly contribute to the M_{CI}. Instead, they tend to generate some higher *ci* values to achieve a reduction in M_{CI} and enhance comfort. This approach also contributes to safety and efficiency.

5.4. *Efficiency analysis*

The average time headway values for CAVs following the IDL, pure DDPG, pure TD3, and CACC strategies are 1.41 s, 1.39 s, 1.39 s, and 1.44 s, respectively, while that for HDV is 1.65 s. Our efficiency reward function is designed to maximize the time headway at 1.26 s. As a result, the IDL strategy, pure DDPG strategy, and pure TD3 strategy are closest to the desired time headway. Since time headway serves as an efficiency metric, with higher efficiency generally associated with shorter time headways, these findings indicate that the IDL strategy improves the randomness and efficiency of the original traffic flow and achieves a relatively high-efficiency level compared to other strategies.

The microscopic perspective metric h_t^i across the five strategies further support this conclusion. To further explore the efficiency performance of different strategies in other episodes, Fig. 12 presents the CDF curves of time headways (ranging from 0 to 3 s) for all vehicles during the simulation across five strategies.

It can be observed that the three DRL-based strategies—IDL, pure DDPG, and pure TD3—demonstrate comparable and relatively high efficiency, followed by CACC and then HDV. Compared to other strategies, HDV exhibits more time headways shorter than 1 s and greater than 3 s. This observation suggests a high variability in HDV behavior, as HDVs are driven by different types of drivers, leading to a wider spread of time headways in the dataset. Aggressive drivers tend to maintain shorter time headways, while conservative drivers prefer larger ones. For CAVs following DRL strategies or parameter-fixed model-based controllers, the behavioral patterns are more consistent, with most time headways concentrated between 1 and 1.5 s.

5.5. *Overall performance*

Table 4 presents the statistical values for five strategies in terms of safety, comfort, efficiency, and stability on the testing dataset. Additionally, the percentage increases or decreases of each metric for the IDL strategy relative to other comparison strategies are presented. Comparing the IDL strategy to both HDVs and the CACC strategy reveals significant improvements across all metrics, particularly in safety and string stability. Safety metrics improve by over 96% compared to HDVs and 49% compared to the CACC strategy. The IDL strategy is string stable, while the CACC strategy and HDV are string unstable. String stability M_{tss} improves by 31.73% compared to HDVs and 29.70% compared to the CACC strategy. When comparing the IDL strategy to the pure TD3 and pure DDPG strategies, the IDL strategy shows improvements across all metrics, except for a slight decrease in efficiency. However, the efficiency levels of the three strategies are very similar. The pure TD3 strategy already achieves sufficiently low TIT values, indicating good safety performance. However, the IDL strategy can further improve it. The safety improvement is particularly notable, with increases of over 30% compared to the pure TD3 strategy and over 46% compared to the pure DDPG strategy. The enhancement in string stability is also significant, with improvements of 19.32% and 24.47% over the pure TD3 and pure DDPG strategies, respectively.

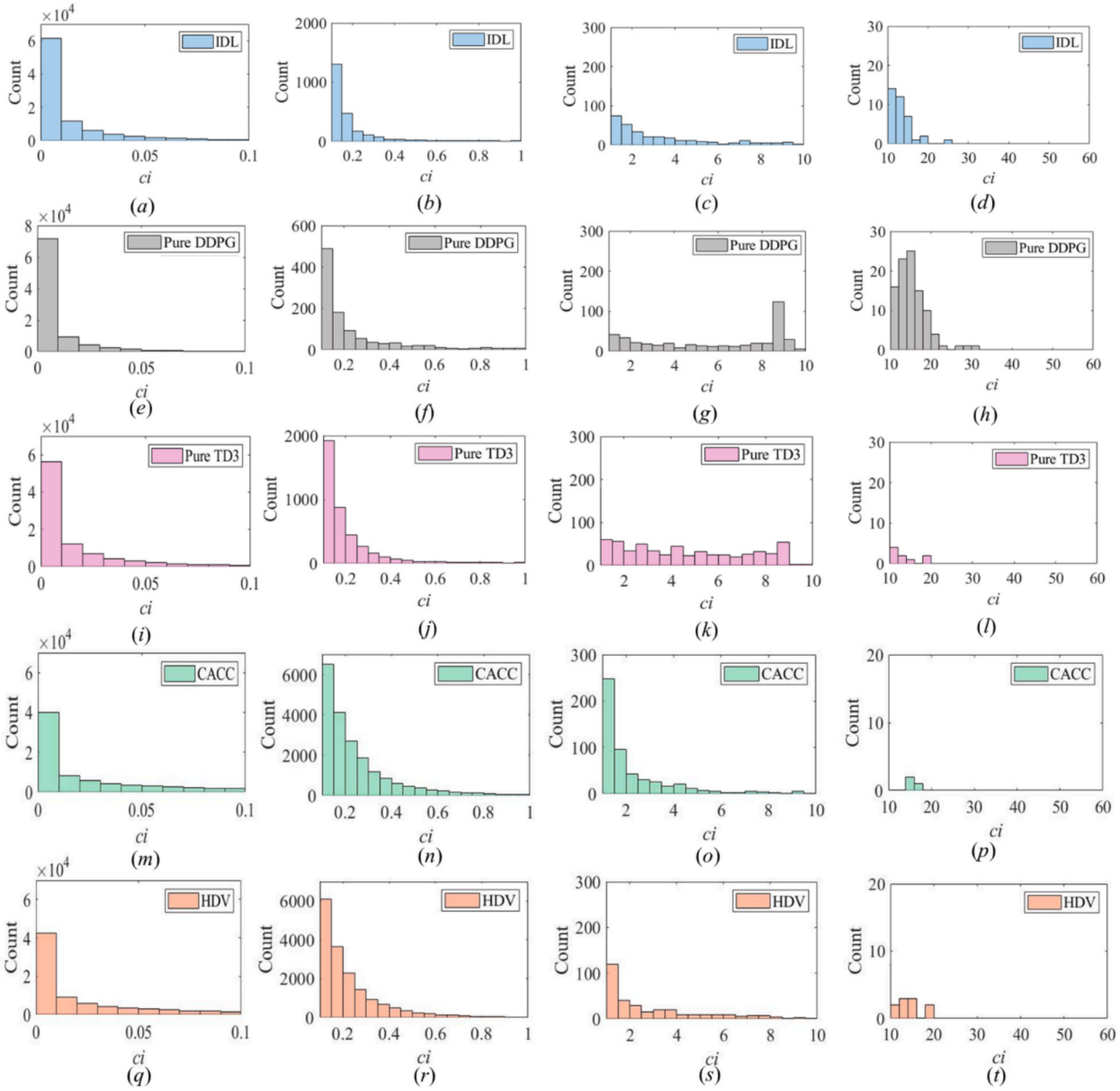
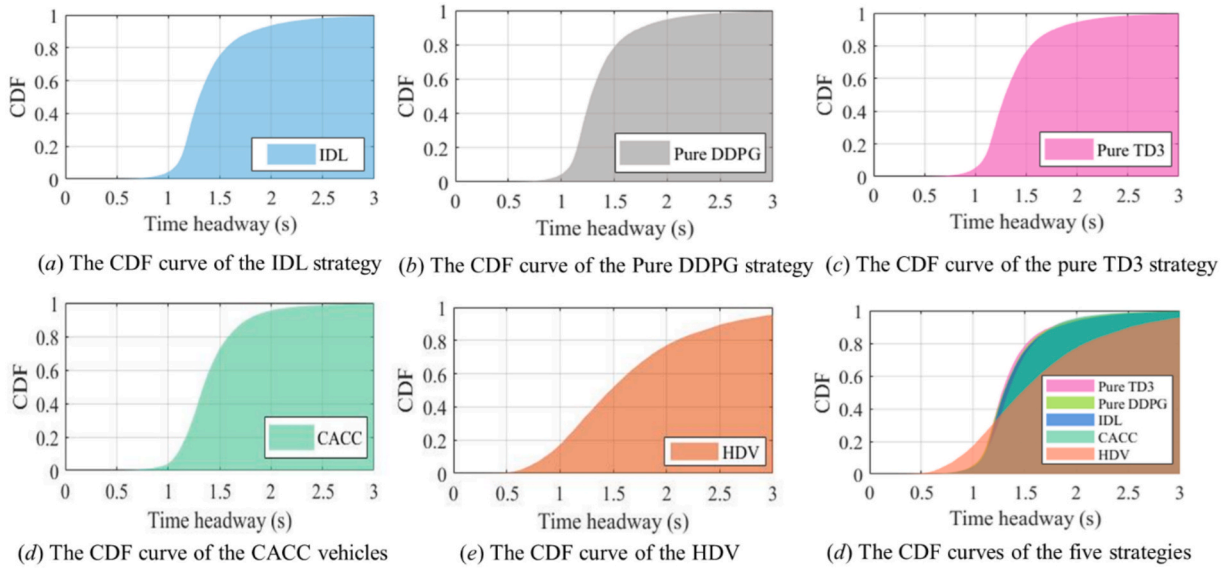


Fig. 11. The quantity of *ci* in the four ranges for vehicles with different strategies.

Notably, the IDL strategy combines the pure TD3 approach with a linear model. The comparison results between these two strategies suggest that integrating a linear model into the DRL framework for CAV longitudinal control can improve string stability, safety, and comfort. In terms of string stability assessment, this approach not only provides the time domain acceleration L_2 norm cumulative damping ratio, M_{tss}, but also enables the theoretical evaluation of string stability via M_{fss}, offering a more precise characterization of string stability.

Overall, compared to other strategies, the IDL strategy shows improvements across all metrics, with only a slight decrease in efficiency relative to the pure DDPG and pure TD3 strategies. The most significant improvements are seen in safety and comfort, with safety performance increasing by 30% to 96%, and comfort performance increasing by 39% to 62%. In terms of string stability, only the IDL strategy meets the string stability conditions, with its M_{tss} value showing a 19% to 31% improvement over other strategies. Additionally, the IDL strategy effectively balances efficiency, with 14% improvement compared to the HDV, and just a slight decrease of 1% compared to the pure DDPG and pure TD3 strategies. This indicates that the IDL strategy, compared to pure DRL or model-based control strategies, achieves enhanced stability while more effectively balancing safety, comfort, and efficiency, demonstrating a particular advantage in string stability.

To evaluate the individual contribution of each component in the reward function, we conducted ablation studies by selectively



(*a*) The CDF curve of the IDL strategy (*b*) The CDF curve of the Pure DDPG strategy (*c*) The CDF curve of the pure TD3 strategy

(*d*) The CDF curve of the CACC vehicles (*e*) The CDF curve of the HDV (*d*) The CDF curves of the five strategies

Fig. 12. The CDF curves of the time headway from 0 to 3 s for all vehicles across five strategies.

Table 4
Testing results and analysis.

Experiment	TIT	M_{CI}	M_{effi}	M_{tss}	M_{fss}
IDL	0.32	0.31	1.41	0.71 (stable)	100%
Compared to HDV	96.50% ↑	54.15% ↑	14.55% ↑	31.73% ↑	–
Compared to CACC	49.21% ↑	62.94% ↑	2.08% ↑	29.70% ↑	100% ↑
Compared to pure TD3	30.43% ↑	39.54% ↑	–1.44% ↓	19.32% ↑	–
Compared to pure DDPG	46.67% ↑	42.25% ↑	–1.44% ↓	24.47% ↑	–
HDV	9.13	0.68	1.65	1.04 (unstable)	–
CACC	0.63	0.84	1.44	1.01 (unstable)	0.00%
Pure TD3	0.46	0.51	1.39	0.88 (stable)	–
Pure DDPG	0.60	0.54	1.39	0.94 (stable)	–

* "-" indicates that the data is not available.

removing one component at a time and observing the resulting impact on system performance. The reward function is composed of three benefit terms, so a total of three ablation experiments were performed: the full reward (baseline), removing each term (– safety, – efficiency, – comfort). The comparative results of these ablation scenarios are summarized in Table 5, which highlights the performance degradation or changes caused by the removal of each component. In addition, we have added symbolic annotations (↓) to the metric headers to indicate the desired directionality, which represents "lower is better". All the metrics are lower is better.

The results show that removing the safety, efficiency, or comfort reward components leads to a decline in the corresponding performance metrics. These ablation studies collectively verify the necessity and effectiveness of each reward component in contributing to the overall control performance.

6. Conclusions

This study proposes a novel integrated DRL-linear control (IDL) strategy for CAV longitudinal control, designed to ensure string stability and enhance safety, comfort, and efficiency by incorporating a linear controller within a deep reinforcement learning (DRL) framework. In this structure, the DRL agent functions as a learning algorithm that interprets the state of the environment to optimize the linear controller's parameters, while the linear controller calculates the desired acceleration for the CAV based on outputs from the DRL agent. The linear controller evaluates string stability, while the metrics for safety, comfort, and efficiency are derived from environmental feedback, collectively guiding the training process within the framework. The twin delayed deep deterministic policy gradient (TD3) algorithm is used as the DRL method for numerical simulations, with real-world human-driving data from the NGSIM dataset used to model leading vehicle behavior.

Experimental results indicate that the IDL strategy achieves improvements across all key metrics, including string stability, safety, and comfort, with only a minor reduction in efficiency compared to the pure DDPG and pure TD3 strategies. But the efficiency levels of the three strategies are very similar, and all are close to the preset desired time headway. The most notable gains are observed in safety and comfort, with safety performance increasing by 30% to 96%, and comfort performance increasing by 39% to 62%. The IDL strategy

Table 5
Ablation studies for the reward function.

Experiment	TIT↓	M_{CI}↓	M_{effi}↓	M_{tss}↓	Collision↓
Full reward	0.32	0.31	1.41	0.71	0
– Safety	∞	0.20	1.32	0.87	2
– Efficiency	0.23	0.19	1.54	0.67	0
– Comfort	0.18	1.34	1.36	0.92	0

is the only one that satisfies the string stability conditions, both from the frequency domain theoretical derivation and time domain statistical analysis. Its time domain acceleration L_2 norm cumulative damping ratio is improved by 19% to 31% compared to the other strategies. Efficiency is effectively maintained with a minimal deviation from the preset desired time headway. These results indicate that the IDL strategy, in comparison to standalone DRL or model-based control methods, achieves enhanced string stability while balancing safety, comfort, and efficiency, with particular strengths in string stability. This is further validated in a platooning scenario, where fluctuations in velocity and acceleration gradually dissipate. In conclusion, the IDL strategy leverages the linear model's advantage in string stability analysis while retaining the DRL approach's capability to optimize multiple objectives, thereby achieving a well-balanced and string stable control solution for CAVs.

Future studies can be carried out based on the results of this study. For instance, more complex information flow topologies can be applied in the control framework of CAVs to further improve the benefits in various aspects. Moreover, vehicle dynamics can be incorporated into the control framework, and more complex vehicle control models can be built to enhance robustness. On the other hand, other AI models can be incorporated into the framework to analyze the potential of solving the problem with other AI models.

CRediT authorship contribution statement

Ziwei Yi: Writing – original draft, Software, Methodology, Conceptualization. **Min Xu:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Shuaian Wang:** Writing – review & editing, Methodology, Funding acquisition, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China [Grant Nos. 72371221, 72361137006], the Research Grants Council of the Hong Kong Special Administrative Region, China [Project number HKSAR RGC TRS T32-707/22-N].

Data availability

Data will be made available on request.

References

Arem, B.V., Van Driel, C.J.G., Visser, R., 2006. The impact of cooperative adaptive cruise control on traffic-flow characteristics. IEEE Trans. Intell. Transp. Syst. 7 (4), 429–436. https://doi.org/10.1109/TITS.2006.884615.

Blum, C., Roli, A., 2003. Metaheuristics in combinatorial optimization: Overview and conceptual comparison. ACM Computing Surveys (CSUR) 35 (3), 268–308. https://doi.org/10.1145/937503.937505.

Brackstone, M., McDonald, M., 1999. Car-following: A historical review. Transport. Res. F: Traffic Psychol. Behav. 2 (4), 181–196. https://doi.org/10.1016/S1369-8478(00)00005-X.

Buechel, M., Knoll, A., 2018. Deep reinforcement learning for predictive longitudinal control of automated vehicles. In: *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 2391–2397. https://doi.org/10.1109/ITSC.2018.8569977.

Chen, D., Ahn, S., Chitturi, M., Noyce, D., 2018. Truck platooning on uphill grades under cooperative adaptive cruise control (CACC). Transp. Res. Part C Emerging Technol. 94, 50–66. https://doi.org/10.1016/j.trc.2017.08.025.

Chu, T., Kalabić, U., 2019. Model-based deep reinforcement learning for CACC in mixed-autonomy vehicle platoon. In: *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 4079–4084. https://doi.org/10.1109/CDC40024.2019.9030110.

Gunter, G., Janssen, C., Barbour, W., Stern, R.E., Work, D.B., 2020. Model-based string stability of adaptive cruise control systems using field data. IEEE Trans. Intell. Veh. 5 (1), 90–99. https://doi.org/10.1109/TIV.2019.2955368.

Hart, F., Okhrin, O., Treiber, M., 2024. Towards robust car-following based on deep reinforcement learning. Transp. Res. Part C Emerging Technol. 159, 104486. https://doi.org/10.1016/j.trc.2024.104486.

Hayward, J.C., 1972. Near miss determination through use of a scale of danger. Available Highw. Res. Rec. 384, 24–34. http://onlinepubs.trb.org/Onlinepubs/hrr/1972/384/384-004.pdf.

Hogema, J. H., & Janssen, W. H. (1996). Effects of intelligent cruise control on driving behavior: a simulator study. In *Intelligent Transportation: Realizing the Future. Abstracts of the Third World Congress on Intelligent Transport Systems ITS America.* Available: https://trid.trb.org/View/574311.

Hirst, S., Graham, R., 2020. The format and presentation of collision warnings. Ergonomics and Safety of Intelligent Driver Interfaces, CRC Press 203–219. https://doi.org/10.1201/9781003064107.

Huang, Z., Xu, X., He, H., Tan, J., Sun, Z., 2019. Parameterized batch reinforcement learning for longitudinal control of autonomous land vehicles. IEEE Transactions on Systems, Man, and Cybernetics: Systems 49 (4), 730–741. https://doi.org/10.1109/TSMC.2017.2712561.

Jiang, L., Xie, Y., Evans, N.G., Wen, X., Li, T., Chen, D., 2022. Reinforcement learning based cooperative longitudinal control for reducing traffic oscillations and improving platoon stability. Transp. Res. Part C Emerging Technol. 141, 103744. https://doi.org/10.1016/j.trc.2022.103744.

Li, Y., Wang, H., Wang, W., Xing, L., Liu, S., Wei, X., 2017. Evaluation of the impacts of cooperative adaptive cruise control on reducing rear-end collision risks on freeways. Accid. Anal. Prev. 98, 87–95. https://doi.org/10.1016/j.aap.2016.09.015.

Li, Z., Chu, T., Kolmanovsky, I.V., Yin, X., 2018. Training drift counteraction optimal control policies using reinforcement learning: An adaptive cruise control example. IEEE Trans. Intell. Transp. Syst. 19 (9), 2903–2912. https://doi.org/10.1109/TITS.2017.2767083.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., & Wierstra, D. (2015). Continuous control with deep reinforcement learning. *Computer Science*, arXiv preprint, arXiv:1509.02971. Doi: 10.48550/arXiv.1509.02971.

Li, Z., Li, Y., Liu, P., Wang, W., Xu, C., 2014. Development of a variable speed limit strategy to reduce secondary collision risks during inclement weathers. Accid. Anal. Prev. 72, 134–145. https://doi.org/10.1016/j.aap.2014.06.018.

Lin, Y., McPhee, J., Azad, N.L., 2021. Comparison of deep reinforcement learning and model predictive control for adaptive cruise control. IEEE Trans. Intell. Veh. 6 (2), 221–231. https://doi.org/10.1109/TIV.2020.3012947.

Lu, W., Yi, Z., Gu, Y., Rui, Y., Ran, B., 2023. TD3LVSL: A lane-level variable speed limit approach based on twin delayed deep deterministic policy gradient in a connected automated vehicle environment. Transp. Res. Part C Emerging Technol. 153, 104221. https://doi.org/10.1016/j.trc.2023.104221.

Ma, K., Wang, H., Zuo, Z., Hou, Y., Li, X., Jiang, R., 2022. String stability of automated vehicles based on experimental analysis of feedback delay and parasitic lag. Transp. Res. Part C Emerging Technol. 145, 103927. https://doi.org/10.1016/j.trc.2022.103927.

Ma, X., Shi, S., Lin, N., Li, Y., 2021. In: Research on platoon agent controller based on twin delayed deep deterministic policy gradient algorithm, pp. 1–5. https://doi.org/10.1109/CVCI54083.2021.9661225.

Minderhoud, M.M., Bovy, P.H.L., 2001. Extended time-to-collision measures for road traffic safety assessment. Accid. Anal. Prev. 33 (1), 89–97. https://doi.org/10.1016/S0001-4575(00)00019-1.

Montanino, M., Monteil, J., Punzo, V., 2021. From homogeneous to heterogeneous traffic flows: Lp string stability under uncertain model parameters. Transp. Res. B Methodol. 146, 136–154. https://doi.org/10.1016/j.trb.2021.01.009.

Ploeg, J., Van De Wouw, N., Nijmeijer, H., 2014. Lp string stability of cascaded systems: Application to vehicle platooning. IEEE Trans. Control Syst. Technol. 22 (2), 786–793. https://doi.org/10.1109/TCST.2013.2258346.

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. Transp. Res. Part CEmerging Technol. 148, 104019. https://doi.org/10.1016/j.trc.2023.104019.

Shi, X., Wong, Y. D., Li, M. Z. F., & Chai, C. (2018). Key risk indicators for accident assessment conditioned on pre-crash vehicle trajectory. *Accident Analysis & Prevention*, *117*, 346-356. https://doi.org/Doi: 10.1016/j.aap.2018.05.007.

Shi, H., Zhou, Y., Wu, K., Wang, X., Lin, Y., Ran, B., 2021. Connected automated vehicle cooperative control with a deep reinforcement learning approach in a mixed traffic environment. Transp. Res. Part C Emerging Technol. 133, 103421. https://doi.org/10.1016/j.trc.2021.103421.

Wang, M., Daamen, W., Hoogendoorn, S.P., Arem, B.V., 2016. Cooperative car-following control : Distributed algorithm and impact on moving jam features. IEEE Trans. Intell. Transp. Syst. 17 (5), 1459–1471. https://doi.org/10.1109/TITS.2015.2505674.

Wang, Y., & Jin, P. J. (2023). Model predictive control policy design, solutions, and stability analysis for longitudinal vehicle control considering shockwave damping. *Transportation Research Part C: Emerging Technologies*, *148*, 104038. https://doi.org/Doi: 10.1016/j.trc.2023.104038.

Yue, X., Shi, H., Zhou, Y., Li, Z., 2024. Hybrid car following control for CAVs: Integrating linear feedback and deep reinforcement learning to stabilize mixed traffic. Transp. Res. Part C Emerging Technol. 167 (August), 104773. https://doi.org/10.1016/j.trc.2024.104773.

Zhang, X., Sun, J., Zheng, Z., Sun, J., 2024. On the string stability of neural network-based car-following models: A generic analysis framework. Transp. Res. Part C Emerging Technol. 160, 104525. https://doi.org/10.1016/j.trc.2024.104525.

Zhang, Y., Tian, B., Xu, Z., Gong, S., Gao, Y., Cui, Z., & Chen, X. (2022). A local traffic characteristic based dynamic gains tuning algorithm for cooperative adaptive cruise control considering wireless communication delay. *Transportation Research Part C: Emerging Technologies*, *142*, 103766. https://doi.org/Doi: 10.1016/j.trc.2022.103766.

Zhou, J., Yan, L., Yang, K., 2023. Safe reinforcement learning for mixed-autonomy platoon control. In: *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 5744–5749. https://doi.org/10.1109/ITSC57777.2023.10422463.

Zhou, Y., Ahn, S., Chitturi, M., Noyce, D.A., 2017. Rolling horizon stochastic optimal control strategy for ACC and CACC under uncertainty. Transp. Res. Part C Emerging Technol. 83, 61–76. https://doi.org/10.1016/j.trc.2017.07.011.

Zhou, Y., Ahn, S., Wang, M., Hoogendoorn, S., 2020. Stabilizing mixed vehicular platoons with connected automated vehicles: An H-infinity approach. Transp. Res. B Methodol. 132, 152–170. https://doi.org/10.1016/j.trb.2019.06.005.

Zhou, Y., Wang, M., Ahn, S., 2019. Distributed model predictive control approach for cooperative car-following with guaranteed local and string stability. Transp. Res. B Methodol. 128, 69–86. https://doi.org/10.1016/j.trb.2019.07.001.

Zhu, M., Wang, Y., Pu, Z., Hu, J., Wang, X., Ke, R., 2020. Safe, efficient, and comfortable velocity control based on reinforcement learning for autonomous driving. Transp. Res. Part C Emerging Technol. 117 (May), 102662. https://doi.org/10.1016/j.trc.2020.102662.