Learning-Based Lagrangian Traffic Control Using Cooperative CAV Fleets for Bottleneck Decongestion

Siwen Yang, Guanxiao Li, Zhitong He, and Lingxi Li

***Abstract*—Highway bottlenecks caused by temporary lane closures can lead to severe congestion and capacity drop, especially under high traffic demand. This paper proposes a learning-based Lagrangian traffic control framework in which a small fraction of connected and automated vehicles (CAVs) in mixed traffic are organized into specially structured multi-lane fleets and utilized as mobile actuators to mitigate congestion. A hierarchical control architecture is developed, where an upper-level reinforcement learning (RL) agent determines fleet-level control parameters, including cruising speed and inter-vehicle spacing, while a lower-level model predictive control (MPC) module ensures real-time trajectory tracking and safety. The proposed method explicitly captures cross-lane CAV cooperation within fleets through the design of the RL action space, and incorporates CAV-induced traffic regulation effects into the reward formulation to optimize traffic dynamics and alleviate bottleneck congestion. Simulation results in a multi-lane highway scenario demonstrate that the proposed strategy significantly improves traffic performance across a wide range of demand conditions, reducing travel delay and congestion length while increasing average traffic speed. In addition, the method exhibits generalization across varying traffic demands and improved performance with increased availability of CAV fleets.**

***Index Terms*—Mixed traffic control, connected and automated vehicle (CAV), reinforcement learning (RL), bottleneck decongestion**

I. Introduction

HIGHWAY bottlenecks, such as work zones, lane drops, and ramp areas, are major sources of congestion and can significantly reduce travel efficiency while increasing operational risk, especially under heavy traffic demand [1, 2]. Studies have shown that once a bottleneck becomes active and congestion forms, the maximum traffic flow capacity can drop by up to 25% [3]. This not only increases travel time and delay, but also contributes to higher accident and incident rates under severe queuing conditions [4]. Such bottlenecks also pose a persistent challenge to the deployment and large-scale operation of connected and automated vehicles (CAVs) [5].

Existing traffic control methods, such as macro-level control using roadside traffic signals, are often limited by infrastructure availability in non-recurrent bottleneck scenarios and by the high cost of installation and maintenance [6]. In addition to roadside signal-based control, another class of solutions focuses on optimizing and controlling CAV trajectories to improve traffic performance at bottlenecks [7, 8]. However, such methods often become computationally demanding when a large number of CAVs need to be coordinated. Moreover, because mixed traffic will remain prevalent in the foreseeable future [9], the noncompliant and uncertain behaviors of human-driven vehicles (HDVs) further limit the practical applicability of trajectory-optimization-based control [10]. Overall, non-recurrent bottlenecks that can arise unpredictably over time and across locations pose a persistent challenge for effective mixed traffic control.

In recent years, a new paradigm of mixed traffic control has emerged, which leverages CAVs as mobile actuators for traffic control and is referred to as Lagrangian traffic control [11]. It leverages CAVs already existed in the traffic stream for control, making it naturally suited for flexible deployment across different locations and times, with lower deployment cost. To mitigate bottleneck congestion in mixed traffic, prior research has explored the use of individual CAVs as decentralized mobile controllers to locally regulate traffic dynamics [12–15], as well as organizing multiple CAVs into groups that function as mobile actuators for corridor-level flow control [16–21]. By adjusting the CAVs' driving behaviors, such as cruising speed and car-following distance, the surrounding mixed flow rate and speed can be actively influenced and regulated, thereby helping to mitigate downstream congestion. Existing works developed macroscopic traffic models based on cell transmission model (CTM) [16–19, 21] or the Lighthill-Whitham-Richards (LWR) model [12, 13, 20] to characterize the macroscopic dynamics induced by controlled CAVs and capture capacity drop at bottlenecks. Nevertheless, the mechanism-based model requires many parameters that are difficult to calibrate in real traffic environments, particularly for temporary bottlenecks that can arise at arbitrary locations. Moreover, these methods typically rely on model-based, optimization-driven traffic control. While offering good interpretability, the model complexity leads to high computational cost, limiting their applicability in real-time settings.

Accordingly, the central research question is how to design a data-driven and real-time Lagrangian traffic control algorithm for mitigating bottleneck congestion in mixed multi-lane traffic involving both CAVs and HDVs. Different from existing Lagrangian traffic control studies that mainly treat CAVs

Manuscript received: 28 April 2026; revised: 4 June 2026; accepted: 5 June 2026. *(Corresponding author: Lingxi Li.)*

Citation: S. Yang, G. Li, Z. He, and L. Li, Learning-based Lagrangian traffic control using cooperative CAV fleets for bottleneck decongestion, *Int. J. Intell. Control Syst.*, 2026, 31(2), 160–172.

Siwen Yang, Guanxiao Li, Zhitong He, and Lingxi Li are with Elmore Family School of Electrical and Computer Engineering, Purdue University, Indianapolis, IN 46202, USA (e-mail: yang3062@purdue.edu; li4992@purdue.edu; he733@purdue.edu; lingxili@purdue.edu).

Digital Object Identifier 10.62678/IJICS202606.10365

as independent mobile controllers, this work introduces a structured cross-lane CAV fleet strategy as an intermediate control abstraction between individual vehicle behavior and system-level traffic dynamics. Specifically, CAVs across adjacent lanes are organized into coordinated fleets with adjustable formation parameters. These parameters jointly determine the fleet structure and behavior, which can influence flow speed and traffic redistribution across lanes. As a result, the proposed fleet structure serves as a bridge between microscopic CAV motion and macroscopic traffic regulation, enabling corridor-level traffic control through coordinated vehicle-level behaviors. Motivated by recent reinforcement learning (RL) applications in traffic control, this paper introduces a hierarchical framework that combines an RL agent for upper-level fleet parameters decision-making with a model predictive control (MPC)-based lower-level controller for fleet vehicle's microscopic motion control. The overall objective is to improve traffic efficiency under dynamic and uncertain bottleneck conditions in an effective and computationally efficient manner. The main contributions of this paper are summarized as follows:

(1) We introduce a cooperative cross-lane control strategy in which a small fraction of CAVs are organized into coordinated multi-lane fleets to actively regulate lane-wise traffic distribution and bottleneck inflow. Compared with existing Lagrangian control methods using non-cooperative CAVs, this structured cooperation enables more effective system-level traffic regulation.

(2) We develop a data-driven formulation of Lagrangian traffic control for bottleneck decongestion by modeling the cooperative behavior of structured CAV fleets as a sequential decision-making problem. This formulation enables the direct learning of the relationship between fleet-level control actions and macroscopic traffic outcomes, without relying on explicit mechanistic traffic models.

(3) We propose a hierarchical multi-scale control framework that integrates a learning-based upper-level policy for corridor-level decision-making with a decentralized MPC-based lower-level controller for vehicle-level motion tracking. This design effectively bridges macroscopic traffic objectives and microscopic vehicle dynamics, enabling real-time implementation.

(4) We design a multi-objective reward function that jointly captures throughput maximization, speed improvement, lane redistribution, and congestion suppression. This reward structure provides informative learning signals and facilitates the efficient training of control policies under complex mixed-traffic dynamics.

The remainder of this paper is organized as follows. Section II reviews representative studies on traffic decongestion. Section III presents the problem formulation. Section IV introduces the proposed hierarchical traffic regulation strategy. Simulation experiments and analysis are presented in Section V. Finally, conclusions are drawn in Section VI.

II. Related Work

Existing studies have shown that Lagrangian traffic control using both individual CAV control and multi-CAV coordination can improve traffic efficiency near bottleneck areas. Early model-based works typically formulated the problem through macroscopic traffic-flow models, such as the LWR model or the cell transmission model, and optimized CAV behaviors including speed, spacing, or platoon structure to regulate upstream inflow and suppress capacity drop [12, 13, 16–21]. However, existing strategies typically rely on mechanistic modeling to capture the microscopic-to-macroscopic traffic dynamics. Such models are highly complex (e.g., involving partial differential equations or mixed-integer optimization) and computationally demanding. As traffic scenarios become more realistic, especially in multi-lane mixed traffic with lane changes and uncertain human driving behaviors [22], modeling and predicting traffic evolution become increasingly challenging [23].

RL-based traffic control has received increasing attention in mixed traffic settings. In recent years, numerous studies have explored leveraging the control of CAV behaviors in mixed traffic to improve overall macroscopic traffic flow performance. Wei et al. [24] studied mixed-autonomy traffic control by using deep reinforcement learning (DRL) to regulate the longitudinal behaviors of CAVs. Their approach focuses on learning distributed control policies that improve traffic fluidity, particularly by increasing the average velocity and mitigating stop-and-go waves at the network level. Similarly, Wu et al. [25] further advanced the method by proposing a modular DRL framework that enabled the composition of diverse mixed-autonomy traffic scenarios. It provides a scalable and flexible platform for learning control policies and demonstrates strong performance across a range of benchmark environments, such as ring roads. More recently, Liu et al. [26] proposed a DRL-based control framework that explicitly treats autonomous vehicles (AVs) as mobile actuators to regulate mixed traffic at the system level, thereby moving beyond individual vehicle control toward a more unified system-level optimization perspective. However, the above studies primarily focus on traffic smoothing in simplified single-lane or homogeneous settings and do not explicitly address more challenging multi-lane scenarios with lane closures, where complex cross-lane interactions and flow redistribution become critical.

Vinitsky et al. [14] introduced DRL-based Lagrangian bottleneck control and showed that controlling a small fraction of AVs can improve bottleneck throughput in a two-stage lane-drop scenario. Their later work [15] extended this idea to a decentralized multi-agent setting, where AVs use local observations to regulate mixed-autonomy traffic without centralized coordination. However, these approaches do not explicitly consider cooperation among CAVs and often rely on restrictive assumptions, such as no lane-changing, which limits their capability to capture lane-wise traffic dynamics and reduces their effectiveness in lane-reduction bottleneck scenarios. Ha et al. [27] further investigated highway bottleneck congestion mitigation by leveraging vehicle connectivity and autonomy through RL-based speed harmonization.

Beyond bottleneck control based on speed harmonization, several recent studies have further explored learning-based CAV regulation and multi-vehicle decision-making strategies from different perspectives [28]. Liu et al. [29] formulated highway bottleneck decongestion as a multi-agent decision-making problem and proposed a rollout-based coordination method for AVs in mixed traffic. Veksler et al. [30] focused on cooperative cruising and used RL to adjust CAV time headway near bottlenecks, providing a practical way to influence local traffic flow through existing car-following control. Lin et al. [31] studied RL-based CAV control in a two-lane moving-bottleneck scenario, showing that learned CAV behaviors can mitigate congestion induced by slow-moving disturbances. However, existing RL-based Lagrangian traffic control studies mainly focus on regulating the behaviors of individual CAVs to influence surrounding traffic flow. Although such approaches can improve local traffic conditions, they do not fully exploit the cooperative regulation capability of CAVs distributed across multiple lanes, particularly for actively coordinating lane-wise flow redistribution in multi-lane bottleneck scenarios. Moreover, vehicle-level control lacks an explicit mechanism for connecting microscopic CAV behaviors with macroscopic traffic objectives. A mesoscopic structured control abstraction is introduced to provide an intermediate layer that bridges coordinated vehicle-level behaviors and corridor-level traffic regulation.

Overall, these limitations motivate the development of a learning-based Lagrangian traffic control framework with explicit structured cross-lane cooperation for multi-lane bottleneck scenarios involving dynamic lane closures. To address this challenge, this work proposes a structured cross-lane fleet representation with controllable formation parameters. By regulating fleet-level behaviors rather than directly optimizing individual vehicle actions, the proposed approach enables coordinated CAV movements to systematically reshape lane-wise traffic dynamics and facilitate corridor-level traffic regulation. Furthermore, the RL-based fleet parameter configuration allows the framework to learn effective control strategies directly through interaction with the environment, avoiding explicit multi-scale modeling of complex microscopic-to-macroscopic traffic dynamics while enabling adaptive and computationally efficient real-time decision making.

III. Problem Statement

The problem considers a multi-lane highway corridor in which a temporary lane closure caused by an incident creates a bottleneck. Under high traffic demand, the bottleneck is activated and leads to a capacity drop, which in turn triggers congestion that propagates upstream. To mitigate the bottleneck, several CAV fleets are generated within the upstream mixed traffic flow. From downstream to upstream, the fleets are indexed by $m = 1, 2, \ldots, N_{\mathrm{m}}$, where N_{m} denotes the total number of fleets. The goal is to leverage all the formed CAV fleets to regulate the multi-lane traffic and decongest the bottleneck. In this study, all CAV fleets share the same structure, consisting of one CAV per lane, as illustrated in Fig. 1. The adjustable formation parameters of each fleet m include the fleet's traveling speed $\hat{v}_m(t)$, and the longitudinal spacings $\hat{s}_{1,m}(t)$ and $\hat{s}_{2,m}(t)$ between vehicles in adjacent lanes (with this work focusing on the three-lane case). The effects of these parameters on traffic flow can be intuitively summarized as follows:

(1) By adjusting the speed $\hat{v}_m(t)$, the fleets can regulate the inflow rate into the bottleneck. By temporarily reducing the upstream inflow, the fleets help prevent vehicle queuing at the bottleneck and mitigate the capacity drop.

(2) Surrounding vehicles tend to perform lane changes to overtake the fleet moving at a lower speed than the surrounding traffic. Thus, $\hat{s}_{1,m}(t)$ and $\hat{s}_{2,m}(t)$ regulate the opportunities for lane changing and the number of vehicles that change lanes to pass the fleet. This helps shape the lane-wise distribution of traffic and encourages vehicles to shift earlier toward the lanes that remain open.

Therefore, $\hat{v}_m(t)$, $\hat{s}_{1,m}(t)$, and $\hat{s}_{2,m}(t)$, $\forall m = 1, 2, \ldots, N_{\mathrm{m}}$ are critical decision variables that must be optimized in real time based on the macroscopic traffic state and the operating conditions of the bottleneck. After the dynamic decision-making of these fleet parameters, the next step at the lower level is to design coordinated control strategies for the CAVs within each fleet to maintain the fleet formation and track the updated fleet parameters in real time.

IV. Hierarchical Traffic Regulation Strategy

To coordinate a subset of CAVs in mixed traffic to form fleets for traffic regulation and mitigate bottleneck congestion, this paper proposes a hierarchical regulation strategy. The overall framework of the proposed method is illustrated in Fig. 2. At the upper level, we construct a Markov decision process (MDP) model for the dynamic decision-making of fleet parameters, which defines the state space, action space, and reward function of the traffic regulation problem. A deep reinforcement learning approach is then employed to learn effective parameter-setting policies. The upper-level RL decision-making process operates with a period T. At the lower level, an MPC-based fleet control is developed to coordinate the motion of CAVs within a fleet. The controller operates at a smaller control interval Δt and functions in a decentralized manner at the fleet level. It generates acceleration commands for the CAVs to track the parameters assigned by the upper-level controller while maintaining fleet formation.

A. RL-Based Fleet Coordination

a. MDP Formulation

The interactions between multiple fleets and the traffic environment can be modeled as an MDP, which can be character-

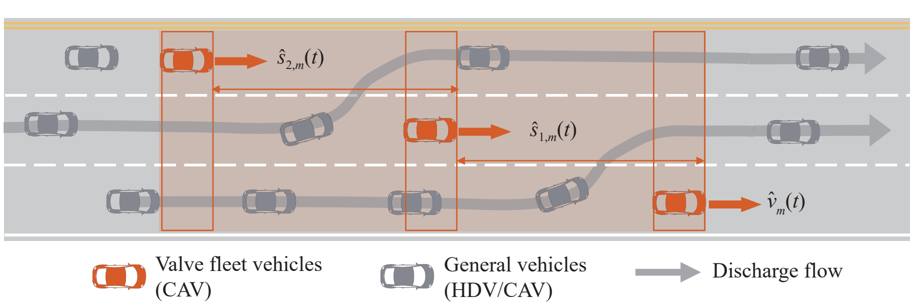


Fig. 1 The structure and decision variables of the fleets. Each fleet consists of one CAV in each lane, and the fleet-level decision variables include the target fleet speed $\hat{v}_m(t)$ and the inter-lane longitudinal spacings $\hat{s}_{1,m}(t)$ and $\hat{s}_{2,m}(t)$, which are used to regulate bottleneck inflow and lane-wise traffic redistribution.

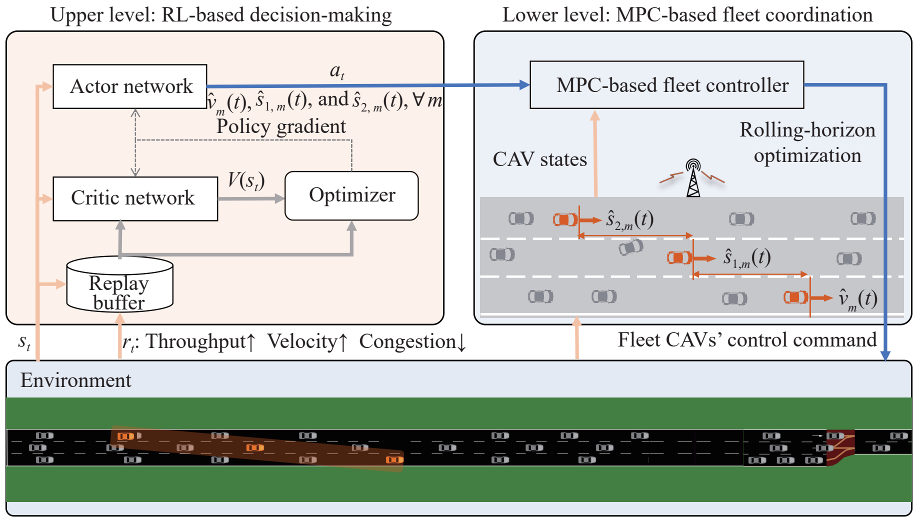


Fig. 2 Schematic diagram of the hierarchical traffic regulation strategy framework.

ized by a six-tuple $\mathcal{M} = (\mathcal{S}, \mu_0, \mathcal{A}, P, R, \gamma)$, where $\mathcal{S}$ (state space) denotes the set of traffic states observed from the environment on the bottleneck segment; μ_0 represents the initial state distribution at the beginning of each episode, as provided by the simulation environment; $\mathcal{A}$ (action space) specifies the possible control operations that can be applied to the fleets on the road; P is the state transition function, which captures the underlying traffic dynamics from microscopic vehicle behaviors to mesoscopic fleet effects and ultimately macroscopic traffic outcomes, although no explicit model is constructed in this work; R denotes the reward function from interacting with the traffic environment; and γ is the discount factor. Next, we focus on the design of the state space, action space, and reward function.

(1) State space: Three categories of states are included in the state space. The first is the macroscopic traffic state of the roadway. Starting from the bottleneck and moving upstream, the road is divided into cells, each of length L_c, indexed by c, with a total of N_c cells. Each cell is further divided into $N_\mathrm{l} = 3$ lanes, indexed by $l = 1,2,3$. In each cell, the lane-wise average speed $v_{c,l}(t)$ and vehicle density $\rho_{c,l}(t)$ are included as state components. The second is the congestion indicator. Road segments with a density above 50 veh/km and an average speed below 30 km/h are considered congested. The congestion length (CL) can be directly observed from the environment and is denoted by $l_\mathrm{cong}(t)$. The third category is the position of all fleets. The locations of all fleets, denoted by $p_m(t)$, are also included as state components. Therefore, the state at time t is defined as

$$s_t = \left\{ \{v_{c,l}(t), \rho_{c,l}(t)\}_{l=1,\,c=1}^{N_\mathrm{l}\ \ N_\mathrm{c}}, l_\mathrm{cong}(t), \{p_m(t)\}_{m=1}^{N_\mathrm{m}} \right\} \tag{1}$$

which has a total dimension of $2N_\mathrm{c} \times N_\mathrm{l} + 1 + N_\mathrm{m}$. The state space $\mathcal{S}$ is the Cartesian product of the continuous domains of all state components, forming a high-dimensional continuous space.

(2) Action space: Consistent with the statement in Section III, the action at time t is defined as

$$a_t = \left\{ \{\hat{v}_m(t), \hat{s}_{1,m}(t), \hat{s}_{2,m}(t)\}_{m=1}^{N_\mathrm{m}} \right\} \tag{2}$$

where the traveling speed and structural spacings of each fleet are included as decision variables, resulting in a total dimension of $3N_\mathrm{m}$. In addition, the decision variables are constrained within predefined ranges

$$\begin{aligned} &\hat{v}_m(t) \in [v_\mathrm{min}, v_\mathrm{max}], \\ &\hat{s}_{1,m}(t),\ \hat{s}_{2,m}(t) \in [0, s_\mathrm{max}], \quad \forall m \end{aligned} \tag{3}$$

where v_min and v_max represent the lower and upper bounds on the fleet cruising speed. s_max is the maximum allowable longitudinal spacing.

(3) Reward function: The MDP reward function consists of four components

$$r_t = \zeta_1 r_\mathrm{thp} + \zeta_2 r_\mathrm{vel} + \zeta_3 r_\mathrm{dist} + \zeta_4 r_\mathrm{cong} \tag{4}$$

where ζ_1, ζ_2, ζ_3, and ζ_4 are weight coefficients, r_thp is the throughput reward, r_vel is the average traffic speed reward, r_dist penalizes vehicles remaining in the closed lane, and r_cong penalizes the congestion length.

The first component rewards the overall throughput of the roadway (i.e., the number of vehicles exiting the bottleneck). First, a reference throughput capacity is defined as

$$n_\mathrm{cap} = \min\left(\frac{q_\mathrm{in} N_\mathrm{l}}{3600} T,\ \frac{q_\mathrm{bn}}{3600} T \right) \tag{5}$$

where q_in is the per-lane inflow rate, and q_bn is the ideal discharge capacity of the bottleneck, corresponding to its highest outflow under uncongested conditions. The reference throughput represents the maximum achievable outflow constrained by the upstream traffic demand and the bottleneck capacity. Then, the throughput reward is given by

$$r_\mathrm{thp} = \begin{cases} 0, & \text{if } n_\mathrm{out} \leqslant 0.6 n_\mathrm{cap}; \\ \dfrac{n_\mathrm{out} - 0.6 n_\mathrm{cap}}{0.4 n_\mathrm{cap}}, & \text{if } 0.6 n_\mathrm{cap} < n_\mathrm{out} \leqslant n_\mathrm{cap}; \\ 1, & \text{if } n_\mathrm{out} > n_\mathrm{cap} \end{cases} \tag{6}$$

where $n_\mathrm{out} = \sum_{c=1}^{N_\mathrm{c}} \sum_{l=1}^{N_\mathrm{l}} L_\mathrm{c}(\rho_{c,l}(t-1) - \rho_{c,l}(t))$ denotes the number of vehicles that exit the bottleneck during the interval from time step $t-1$ to t. By rewarding this quantity in a piecewise manner, we encourage increased throughput at the bottleneck.

The second component rewards the average traffic speed (ATS) over the entire roadway segment

$$r_{\text{vel}} = \frac{\sum_{c=1}^{N_{\text{c}}} \sum_{l=1}^{N_{\text{l}}} \rho_{c,l}(t) v_{c,l}(t)}{\sum_{c=1}^{N_{\text{c}}} \sum_{l=1}^{N_{\text{l}}} \rho_{c,l}(t) v_{\max}} \tag{7}$$

which is the average speed of all vehicles on the roadway at time t, normalized by $1/v_{\max}$.

The third component penalizes the proportion of vehicles remaining in the closed lane in the downstream portion of the road segment before reaching the bottleneck

$$r_{\text{dist}} = -\frac{1}{(N_{\text{c}} - 15)} \sum_{c=16}^{N_{\text{c}}} \frac{N_{\text{l}} \rho_{c,1}(t)}{\sum_{l=1}^{N_{\text{l}}} \rho_{c,l}(t)} \tag{8}$$

where $\rho_{c,1}(t)$ denotes the vehicle density in the closed lane (lane 1) of cell c. This normalized penalty discourages a high fraction of vehicles from remaining in the closed lane near the bottleneck, thereby encouraging earlier redistribution toward the open lanes.

The fourth component penalizes the congestion length on the road.

$$r_{\text{cong}} = -\frac{1}{\eta_{\text{cong}}} l_{\text{cong}}(t) \tag{9}$$

where η_{cong} is a normalization factor. Combining all four components yields the final reward.

b. Proximal Policy Optimization (PPO) Algorithm

PPO is adopted to solve the formulated MDP because it provides stable and sample-efficient policy updates in continuous state-action spaces, which is well suited for real-time traffic control. The proposed problem involves high-dimensional continuous fleet-level actions and safety-sensitive traffic regulation decisions, where the action dimension increases with the number of coordinated CAV fleets. Compared with deterministic policy-gradient methods such as deep deterministic policy gradient (DDPG) and twin delayed deep deterministic policy gradient (TD3), PPO generally provides more stable policy updates through its clipped objective and exhibits lower sensitivity to hyperparameter settings. Moreover, compared with entropy-based methods such as soft actor-critic (SAC), PPO often achieves a favorable balance between convergence stability and training efficiency in complex traffic control tasks. In this paper, we focus on several aspects in our customized PPO algorithm, which are critical for effective learning in the proposed traffic regulation problem.

The state variables exhibit significantly different scales (e.g., speeds versus positions), which may bias learning toward large-magnitude features. We apply online state normalization using Welford's algorithm to maintain running estimates of the mean and variance

$$\tilde{s}_t = \frac{s_t - \mu_t}{\sigma_t} \tag{10}$$

where μ_t and σ_t are estimates of the states' mean and standard deviation updated by $\mu_t = \mu_{t-1} + (s_t - \mu_{t-1})/t$ and $\sigma_t^2 = M_t/(t-1)$ with accumulated variance $M_t = M_{t-1} + (s_t - \mu_{t-1})(s_t - \mu_t)$.

To ensure that the policy actions satisfy the constraints in Formula (3), a squashing function based on the hyperbolic tangent is used to bound the actor output, followed by rescaling to the feasible action interval. Specifically, a raw Gaussian action a_t^{raw} is transformed as

$$a_t = a_{\min} + \frac{\tanh(a_t^{\text{raw}}) + 1}{2} (a_{\max} - a_{\min}) \tag{11}$$

which guarantees that the executed action lies within the prescribed bounds ($a_{\max}$ and $a_{\min}$ are derived from $s_{\max}$, $v_{\max}$, and $v_{\min}$). To correctly evaluate the policy likelihood under this transformation, the log-probability is adjusted using the Jacobian of the tanh-based mapping

$$\log \pi_\theta(a_t|s_t) = \log \mathcal{N}(a_t^{\text{raw}}|\mu_\theta(s_t), \sigma_\theta) - \sum_i \log\left(1 - \tanh^2(a_{t,i}^{\text{raw}})\right) \tag{12}$$

Finally, we adopt the clipped version of PPO to ensure stable policy updates during training. The policy is optimized using the clipped PPO objective

$$\max_\theta E_{s \sim \bar{\rho}_{\mu_0}^{\pi_{\theta_t}}} \left[E_{a \sim \pi_{\theta_t}} \left[\min\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_t}(a|s)} A^{\pi_{\theta_t}}(s,a), \operatorname{clip}\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_t}(a|s)}, 1-\epsilon, 1+\epsilon \right) A^{\pi_{\theta_t}}(s,a) \right) \right] \right] \tag{13}$$

where $\bar{\rho}_{\mu_0}^{\pi_{\theta_t}}$ denotes the occupancy measure induced by the initial state distribution μ_0 and the current policy π_{θ_t}. The designed PPO algorithm is summarized in Algorithm 1.

B. MPC-Based Fleet Control

To coordinate the motion of CAVs among vehicles in each fleet, we adopt an MPC framework. The schematic diagram of the lower-level fleet coordination is shown in Fig. 3. Each fleet consists of one CAV per lane, and the controller regulates their longitudinal motion to track the desired fleet-level parameters provided by the upper-level RL policy. Compared to the longer decision interval T of the upper level, the lower-level fleet control operates at a shorter control interval Δt. In the following, we use k to index the time steps in the lower-level control.

Let $\tilde{v}_l(k)$ and $\tilde{a}_l(k)$ denote the speed and acceleration of the CAV in lane l at prediction step k, respectively. Let $\tilde{s}_l(k)$ denote the longitudinal spacing between the CAV in lane l and the CAV in lane $l-1$ at step k. At each time step, the speed $\tilde{v}_l^{\text{lead}}$ and distance $\tilde{s}_l^{\text{lead}}$ of the leader ahead of each fleet CAV in the same lane are measured to account for collision avoidance. The control input is the acceleration increment $\Delta\tilde{a}_l(k)$ of all fleet CAVs.

At each decision step, the MPC solves a finite-horizon optimal control problem over a prediction horizon N_{p} in a receding horizon manner. Given the current system state, it predicts the future evolution of vehicle speeds and spacings over the horizon based on candidate control inputs. An optimal sequence of control actions is then obtained by minimizing a

Algorithm 1 PPO training for CAV fleet-based traffic regulation strategy

Input: Environment $\mathcal{E}$, policy/value network π_θ and V_ϕ, and action bounds in Formula (3).
Initialize running statistics for state normalization: mean μ_t and variance accumulator M_t.
for training iteration $k = 1, 2, \ldots,$ **do**
  Initialize an empty rollout buffer $\mathcal{B}$.
  for rollout step $t = 1, 2, \ldots, N_{\mathrm{roll}}$ **do**
    Observe raw state s_t from environment, update running statistics (μ_t, M_t, σ_t^2), and compute normalized state via Eq. (10).
    Compute Gaussian policy and sample raw Gaussian action a_t^{raw}.
    Apply squashing and rescaling via Eq. (11) to ensure that a_t satisfies Formula (3).
    Compute corrected log-probability with tanh-based Jacobian by Eq. (12).
    Execute action a_t in $\mathcal{E}$ and observe reward r_t, next state s_{t+1} and terminal flag.
    Store $(s_t, a_t, r_t, \log \pi(a_t|s_t), V_\phi(s_t))$ into $\mathcal{B}$.
  end for
  Compute generalized advantage estimation (GAE) advantages $\hat{A}_t$ and returns $\hat{R}_t$ using bootstrap value. Normalize advantages.
  for epoch $e = 1, 2, \ldots, E$ **do**
    Shuffle indices and split $\mathcal{B}$ into mini-batches.
    for each mini-batch **do**
      Recompute normalized states using current running statistics by Eq. (10).
      Recompute log-probability $\log \pi_\theta(a_t|s_t)$ via the arctanh transformation.
      Compute PPO surrogate losses by Formula (13).
      Compute value loss and entropy bonus.
      Update θ and ϕ via gradient descent and clip gradient norm.
    end for
  end for
end for

cost function that balances speed tracking, spacing regulation, and control effort, subject to dynamic and safety constraints. Only the first control input is applied, and the process is repeated at the next time step with updated state information. The receding-horizon optimization model is given by

$$\min_{\{\Delta\tilde{a}_l(k)\}} \sum_{k=1}^{N_{\mathrm{p}}} \left(\sum_{l=1}^{N_{\mathrm{l}}} w_{\mathrm{v}} (\tilde{v}_l(k) - \hat{v}_m(t))^2 + \sum_{l=2}^{N_{\mathrm{l}}} w_{\mathrm{s}} (\tilde{s}_l(k) - \hat{s}_{l-1,m}(t))^2 + \sum_{l=1}^{N_{\mathrm{l}}} w_{\mathrm{a}} \tilde{a}_l(k)^2 \right) \tag{14a}$$

$$\text{s.t., } \tilde{s}_l(k+1) = \tilde{s}_l(k) + \Delta t(\tilde{v}_{l-1}(k) - \tilde{v}_l(k)) + \frac{1}{2}\Delta t^2 (\tilde{a}_{l-1}(k) - \tilde{a}_l(k)), \quad \forall l \geqslant 2, k \tag{14b}$$

$$\tilde{v}_l(k+1) = \tilde{v}_l(k) + \tilde{a}_l(k)\Delta t, \quad \forall l, k \tag{14c}$$

$$\tilde{a}_l(k) = \tilde{a}_l(k-1) + \Delta\tilde{a}_l(k), \quad \forall l, k \tag{14d}$$

$$-\Delta a_{\max} \leqslant \Delta\tilde{a}_l(k) \leqslant \Delta a_{\max}, \quad \forall l, k \tag{14e}$$

$$a_{\min} \leqslant \tilde{a}_l(k) \leqslant a_{\max}, \quad \forall l, k \tag{14f}$$

$$v_{\min} \leqslant \tilde{v}_l(k) \leqslant v_{\max}, \quad \forall l, k \tag{14g}$$

$$\sum_{j=0}^{k-1} \left(\tilde{v}_l(j)\Delta t + \frac{1}{2}\tilde{a}_l(j)\Delta t^2 \right) + \tau \tilde{v}_l(k) + s_{\mathrm{safe}} \leqslant \tilde{s}_l^{\mathrm{lead}} + \tilde{v}_l^{\mathrm{lead}} k\Delta t, \quad \forall l,\ k = 1, 2, \ldots, N_{\mathrm{p}} \tag{14h}$$

In Formula (14a), objective function is formulated to penalize the deviation from the desired fleet speed $\hat{v}_m(t)$ and target inter-vehicle spacings $\hat{s}_{1,m}(t)$ and $\hat{s}_{2,m}(t)$ given by the upper level, while also regularizing control effort. w_{v}, w_{s}, and w_{a} are weighting coefficients that balance the tracking of desired speed, inter-vehicle spacing, and control effort, respectively. Constraint Eq. (14b) captures the evolution of inter-vehicle spacing based on relative motion. Constraint Eq. (14c) describes the vehicle speed dynamics. Constraint Eq. (14d) defines the acceleration update through incremental control inputs. Constraints Formulas (14e)–(14g) impose bounds on acceleration increments, acceleration, and speed, respectively. Constraint Formula (14h) enforces collision avoidance with the preceding vehicle by assuming a constant-speed motion of the leading vehicle in the prediction horizon N_{p}. $\tilde{s}_l^{\mathrm{lead}}$ and $\tilde{v}_l^{\mathrm{lead}}$ denote the measured initial spacing and speed of the leading vehicle at the current time step. s_{safe} denotes a minimum spacing and τ is a time headway parameter. Since the lower-level controller adopts a receding-horizon MPC framework, the optimization problem is repeatedly solved at each control interval using the latest traffic observations. Specifically, the distance and speed of the nearest preceding HDV for each fleet CAV are updated every control step and used to reconstruct the collision-avoidance constraint in Formula (14h). Therefore, when surrounding HDVs perform lane-changing maneuvers (e.g., cut-in behaviors), the controller can continuously adapt its formulation and generate updated control actions according to the latest traffic state.

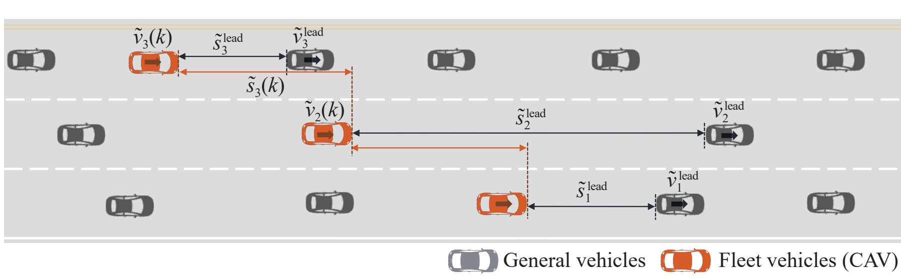


Fig. 3 Schematic diagram of the lower-level fleet coordination and control. The controller regulates the longitudinal motions of fleet CAVs using measured leader distances and speeds while tracking the desired fleet speed and inter-vehicle spacings generated by the upper-level RL policy.

C. Implementation Procedure

During the online execution phase, every T s, we collect data from the simulation environment, obtain the current state s_t, and apply state normalization. The normalized state is then fed into the trained networks from Algorithm 1. It outputs a continuous action distribution, from which an action a_t is sampled. The sampled action is subsequently decoded to the corresponding parameters $\hat{v}_m(t)$, $\hat{s}_{1,m}(t)$, and $\hat{s}_{2,m}(t)$, $\forall m = 1,2,\dots,N_\mathrm{m}$, which are dispatched to each fleet. The lower-level fleet controller then executes these parameters at a finer control interval Δt using an MPC-based controller in Formula (14) with collision-avoidance constraints to generate acceleration commands for the fleet CAVs. These commands are applied to the vehicles to regulate their longitudinal motion and inter-vehicle spacing, thereby tracking the desired fleet-level targets. The resulting traffic states are then fed back into the system, forming a closed-loop control process that continuously updates the control actions in response to evolving traffic conditions. In addition, the fleet control is mainly applied upstream of the bottleneck area for traffic-flow regulation, while the fleet structure is released before entering the lane-reduction region, allowing CAVs to return to normal driving behaviors during the merging process.

V. Experiments and Analysis

A. Simulation Scenario and Setup

The experiments in this work are conducted in the microscopic traffic simulator Simulation of urban mobility (SUMO) [32], and a schematic of the designed simulation environment is shown in Fig. 4. A three-lane, 16 km highway segment narrows to two lanes at its downstream end, where congestion emerges under high traffic demands. The proposed control framework focuses on regulating traffic in the upstream transition region before vehicles enter the bottleneck. Therefore, the control objective is to improve traffic redistribution and merging efficiency at the bottleneck entrance, while the downstream bottleneck section after lane reduction is not explicitly controlled. For vehicles other than the cooperative CAVs controlled within the fleets, their behaviors are determined by the built-in microscopic vehicle models in SUMO. Specifically, the car-following behavior is modeled using the Krauss model, with parameters accel = 3.5, decel = 5.5, desiredMaxSpeed = 27.78, length = 5.0, minGap = 1.0, $\tau = 1.0$, and $\sigma = 0.1$. The lane-changing behavior is modeled using the LC2013 model, with parameters lcSpeedGain = 5.0, lcSpeedGain-Urgency = 50, lcStrategic = 0.1, lcLookaheadLeft = 0.2, and lcOvertake-Right = 0.1. We train and evaluate the proposed RL-based decongestion strategy in this SUMO-based environment. The MPC optimization problem in Formula (14) is solved using the Gurobi solver [33].

The experimental parameters used in the proposed RL-based strategy are shown in Table 1.

B. Training Process

The parameters associated with the MDP setting and PPO algorithm in the proposed algorithm are set as follows: discount factor $\gamma = 0.99$, PPO clipping parameter $\epsilon = 0.2$, and GAE decay parameter $\lambda = 0.95$. The training setup uses a rollout length of 1024, a batch size of 512, 8 training epochs per update, and a learning rate of $\alpha = 3\times10^{-4}$. The actor is parameterized by a two-layer fully connected neural network with 256 hidden units per layer and rectified linear unit (ReLU) activations, followed by a linear layer that outputs the mean of a Gaussian action distribution. The critic adopts a similar two-layer fully connected structure with 256 hidden units and ReLU activations, and outputs a scalar state-value estimate. All training and experiments were conducted on a laptop with an Intel i7-14650HX CPU, an NVIDIA GeForce RTX 5060 GPU, and 16 GB of RAM.

The PPO training process is visualized in Fig. 5. During training, at the beginning of each episode, a traffic demand level is uniformly sampled from {1800, 2000, 2200, 2400, 2600} veh/(h·lane), and SUMO generates the corresponding traffic flow for that episode accordingly. The agent interacts with the environment for a total of 5000 episodes. Fig. 5 shows the variation of the average reward per step within each episode over training. Specifically, it illustrates the evolution of the mean rollout reward over training episodes, as well as the range of average rewards corresponding to episodes under

Table 1 Experimental parameter settings of the proposed RL-based strategy.

Parameter	Description	Value
N_l	Number of lanes	3
N_c	Number of road cells	20
N_m	Number of fleets (mobile actuators)	8
L_c	Length of road cells	0.8 km
T	Decision interval of the MDP	15 s
q_bn	Bottleneck capacity	5300 veh/h
v_min	Minimum allowable fleet speed	18 km/h
v_max	Maximum allowable fleet speed	100 km/h
s_max	Maximum fleet longitudinal spacing	70 m
η_cong	Normalization factor for congestion term	700
ζ_1, ζ_2, ζ_3, and ζ_4	RL reward weights	1, 2, 1, and 1.5
Δt	Control interval of the MPC	1 s
N_p	Prediction horizon of the MPC	8
a_min	Minimum allowable acceleration	-5.5 m/s^2
a_max	Maximum allowable acceleration	3.5 m/s^2
Δa_max	Maximum acceleration increment	4 m/s^2
s_safe	Minimum safety spacing for fleet CAVs	2 m
τ	Time headway parameter for fleet CAVs	0.8 s
w_v, w_s, and w_a	MPC weighting coefficients	1, 1, and 0.05

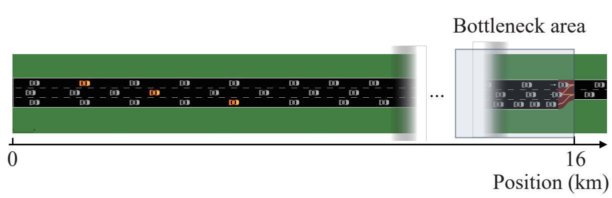


Fig. 4 The experimental scenario of a bottleneck segment designed in SUMO.

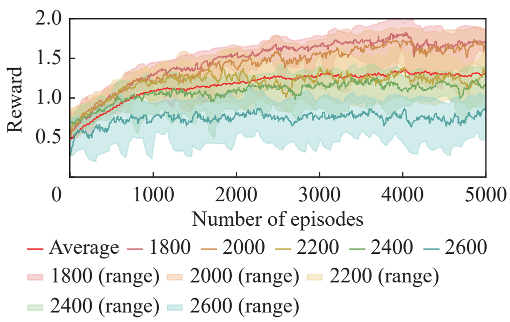


Fig. 5 PPO training performance under varying traffic demand levels. The solid curves represent the sliding-window average rollout reward, while the shaded regions indicate the range of episode rewards. A single policy is trained using traffic demands uniformly sampled from {1800, 2000, 2200, 2400, 2600} veh/(h·lane) over 5000 episodes.

different traffic demand levels. It is worth noting that a single agent was trained across a range of demand conditions, rather than training separate agents for each demand level; the cases corresponding to different traffic demand levels in the figure are derived by extracting, aggregating, and visualizing episodes associated with different sampled demand levels from the 5000 training episodes.

Overall, the PPO training process exhibits clear and relatively stable improvement. The mean rollout reward increases throughout training, indicating continuous policy enhancement. In the early stage of training (within the first 1000 episodes), episodes under all traffic demand levels show an increase in reward, and the overall mean reward also rises. Moreover, it is evident that episodes with lower traffic demand experience faster reward improvement, which is consistent with the fact that traffic regulation and congestion mitigation are easier under lighter demand conditions, allowing the agent to quickly learn effective strategies. In contrast, higher traffic demand levels (2400–2600 veh/(h·lane)) present greater learning difficulty, with reward curves exhibiting more pronounced oscillations. Nevertheless, with continued training, the rewards for high-demand episodes also show a fluctuating upward trend, indicating that the agent eventually learns effective strategies even in these more challenging scenarios. Although the policy requires a relatively large number of interaction steps to converge, this slower convergence is expected given the high-dimensional continuous state-action space and the complexity of mixed traffic dynamics.

C. Regulation Effectiveness Comparison and Analysis

We compare the following four cases.

No control (NC). In this baseline, all vehicles pass through the segment without any external control or intervention.

Non-cooperative CAV regulation strategy (NCCS). This baseline adopts the decentralized RL-based control strategy in Ref. [15], where each CAV independently regulates its longitudinal motion using local observations without explicit cross-lane coordination. The original two-stage 4-2-1 bottleneck (i.e., a two-stage bottleneck where four lanes reduce to two and then reduce to one) is adapted to our multi-lane freeway scenario while preserving its key merging dynamics, and the model is retrained based on the open-source implementation. A shared-policy multi-agent TD3 framework is employed, where all agents share the same actor-critic networks and act as independent learners. The control input is the longitudinal acceleration, and the state representation follows the "minimal + aggregate" setting. Most training parameters are kept consistent with Ref. [15], with scenario-specific adjustments including an AV penetration rate of 20%, a control zone of [14,000, 16,000] m. Consistent with the original setting in Ref. [15], all vehicles are prohibited from changing lanes upstream of the bottleneck.

Fixed-parameter strategy (FPS). Fleets with the proposed structure are used for bottleneck regulation, but no RL-based decision-making or parameter updating is applied. All fleets are assigned fixed yet reasonable parameters, with $\hat{v}_m(t) = 72$ km/h and $\hat{s}_{1,m}(t) = \hat{s}_{2,m}(t) = 55$ m.

The proposed RL-based fleet regulation strategy (RLFRS). $N_{\mathrm{m}} = 8$ fleets are deployed. The fleet formation parameters are dynamic and generated online using the trained policy. The RL agent is trained with the following reward weights: $\zeta_1 = 1$, $\zeta_2 = 2$, $\zeta_3 = 1$, and $\zeta_4 = 1.5$.

To evaluate the effectiveness of the proposed strategy in decongesting the bottleneck, we compare three performance metrics.

(1) Average travel delay (ATD): ATD is defined as the average travel delay of all vehicles traversing the 16 km corridor in the experiment, calculated as the difference between the actual travel time and the travel time under ideal conditions.

(2) Average traffic speed (ATS): ATS is defined as the average speed of all vehicles on the roadway over the duration of the experiment.

(3) Congestion length (CL): CL is defined as the maximum congestion length observed on the roadway during the experiment.

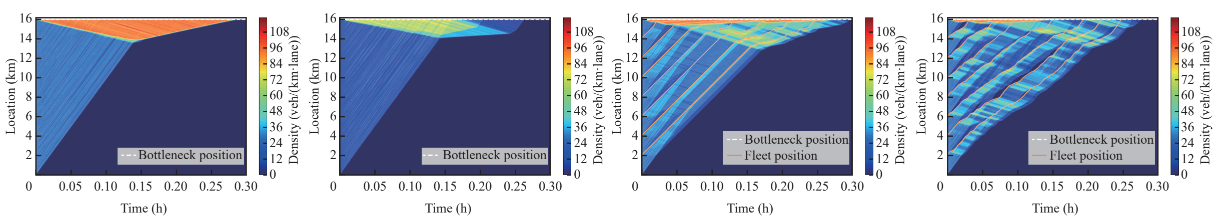
The average performance of the randomized traffic simulation under four strategy cases, across different traffic demand levels ranging from 1800 to 2600 veh/(h·lane), is presented in Table 2. Values are reported as mean ± standard deviation over five independent runs with different random seeds. To better illustrate the traffic dynamics on the bottleneck segment under different regulation cases, the traffic density contour plots along the roadway are presented in Figs. 6 and 7 for demand levels of 2600 veh/(h·lane) and 2200 veh/(h·lane), respectively. In the density map, the color transition from red to blue represents decreasing traffic density. In addition, the bottleneck location is highlighted using white dashed lines, while the fleet positions are highlighted using orange solid lines.

Compared with the NC case, the proposed RLFRS consistently improves all three performance metrics across different demand levels. In terms of ATD, RLFRS achieves a reduction of approximately 5%–23%, with more significant improvements observed under high-demand conditions (e.g., a reduction from 568 to 457 s at 2600 veh/(h·lane)). Meanwhile, the ATS is increased, and the CL is markedly reduced, with the most notable improvement observed at 2600 veh/(h·lane). From the traffic density contour plots, it can be observed that, compared to NC (Fig. 6(a)), the proposed strategy (Fig. 6(d)) substantially reduces both the spatial extent and duration of

Table 2 Comparative performance across different strategy cases. The bold is the best result. The arrows indicate the preferred direction of each metric: "↓" means lower values are better, and "↑" means higher values are better.

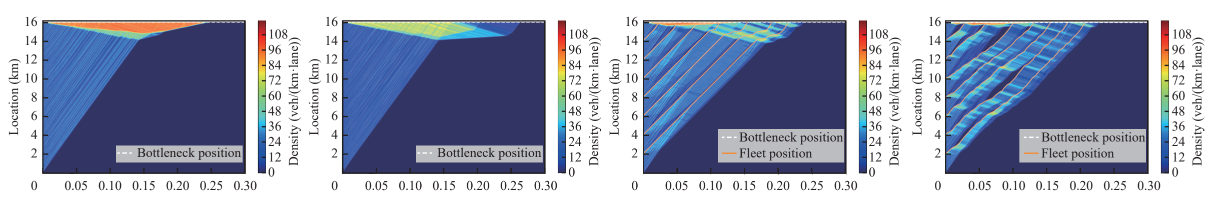
Demand level (veh/(h·lane))	Average travel delay (s) ↓				Average traffic speed (km/h) ↑				Congestion length (km) ↓			
	NC	NCCS	FPS	RLFRS	NC	NCCS	FPS	RLFRS	NC	NCCS	FPS	RLFRS
1800	194±8	**176±7**	185±7	178±5	61.1±0.5	**80.9±0.9**	76.1±0.4	80.6±1.5	0.58±0.05	**0.20±0.06**	0.30±0.01	**0.20±0.09**
1900	223±10	214±14	215±12	**212±12**	58.2±0.6	76.3±1.1	74.4±1.3	**78.2±1.3**	0.75±0.03	0.33±0.05	0.27±0.02	**0.22±0.06**
2000	263±8	258±16	249±6	**240±14**	54.4±0.2	66.5±1.7	68.8±0.9	**75.6±2.1**	0.80±0.07	0.65±0.13	0.35±0.05	**0.26±0.11**
2100	319±10	311±10	283±7	**265±12**	51.7±0.2	60.6±1.5	62.7±0.8	**72.3±1.1**	1.08±0.03	1.12±0.05	0.35±0.06	**0.29±0.06**
2200	369±10	355±13	327±6	**299±8**	48.2±0.4	58.0±1.9	59.7±1.3	**68.1±2.0**	1.22±0.09	1.31±0.05	0.46±0.05	**0.41±0.03**
2300	434±11	415±14	379±12	**333±13**	46.1±0.3	51.8±2.8	53.4±1.3	**65.3±3.2**	1.39±0.05	1.55±0.12	0.62±0.07	**0.42±0.09**
2400	453±11	419±10	421±8	**387±10**	44.7±0.4	51.6±2.2	48.7±1.9	**60.7±1.2**	1.63±0.05	1.71±0.08	0.64±0.11	**0.42±0.10**
2500	511±5	492±17	453±3	**429±15**	42.5±0.2	49.1±1.8	47.6±1.0	**58.5±3.8**	1.88±0.05	1.93±0.09	0.91±0.14	**0.47±0.11**
2600	568±10	533±13	497±8	**457±18**	40.9±0.3	47.3±1.4	45.3±0.8	**54.9±2.0**	2.14±0.03	2.28±0.08	1.12±0.12	**0.49±0.08**

congestion (red regions). By regulating traffic speeds in the upstream region, the CAV fleets prevent excessive inflow from overwhelming the bottleneck capacity, keeping the inflow to the bottleneck within a reasonable range, thereby mitigating capacity drop and maintaining relatively higher throughput. Moreover, the structured fleet formation induces vehicles to shift toward inner lanes in advance, reducing disorderly mandatory lane changes near the bottleneck and further improving traffic flow efficiency. Additionally, during training, we only sampled demand levels of 1800, 2000, 2200, 2400, and 2600 veh/(h·lane). Nevertheless, the proposed method demonstrates strong performance when tested on unseen demand levels (1900, 2100, 2300, and 2500 veh/(h·lane)), indicating a certain level of generalization capability.



(a) No control (b) Non-cooperative CAV regulation strategy (c) Fixed-parameter strategy (d) The proposed strategy

Fig. 6 The contour plots of traffic flow density on the highway segment under different strategy cases under the demand level of 2600 veh/(h·lane). Warmer colors indicate higher traffic density. White dashed lines denote the bottleneck location, and orange solid lines indicate fleet positions.



(a) No control (b) Non-cooperative CAV regulation strategy (c) Fixed-parameter strategy (d) The proposed strategy

Fig. 7 The contour plots of traffic flow density on the highway segment under different strategy cases under the demand level of 2200 veh/(h·lane). Warmer colors indicate higher traffic density. White dashed lines denote the bottleneck location, and orange solid lines indicate fleet positions.

The FPS can be regarded as a partial ablation of the proposed method, where the structured fleet formation is retained but without RL-based real-time decision-making and parameter updates. As shown in Table 2, FPS is able to achieve certain improvements over the no-control case, indicating that the deployment of structured CAV fleets alone can positively influence traffic flow. However, compared with RLFRS, its performance remains limited across all metrics. This is because the fixed fleet parameters cannot adapt to evolving traffic conditions, leading to suboptimal regulation under varying demand levels. In contrast, the proposed RLFRS dynamically adjusts fleet speed and spacing based on real-time traffic states, enabling more flexible and responsive control, and thereby achieving consistently greater improvements in delay reduction, speed enhancement, and congestion mitigation.

The NCCS demonstrates relatively stronger performance under low demand levels (1800–2000 veh/(h·lane)). On the one hand, this is partly because NCCS controls all CAVs in the mixed traffic flow, accounting for 20% of all vehicles, whereas the proposed method uses only a small subset of CAVs, up to about 8%. On the other hand, under relatively low traffic demand (where lane 2 can still accommodate the additional traffic diverted from the closed lane 1), the upstream no-lane-changing setting adopted by NCCS provides an advantage. The mechanism of NCCS can be interpreted as a "flow block": each CAV moderately restricts the inflow near the bottleneck entrance to prevent congestion formation. As reported in Ref. [15], in such scenarios, the no-lane-changing setting leads to higher throughput, because allowing other vehicles to change lanes may disrupt the CAVs' flow-blocking effect and weaken the control performance.

However, as traffic demand increases, the performance of NCCS deteriorates significantly. This is understandable, since NCCS was originally designed and tested in Ref. [15] for a 4-2-1 lane-drop scenario, with the highest tested demand being 875 veh/(h·lane). By contrast, this paper considers one closed lane but with much higher demand. In this case, the no-lane-changing setting becomes a major disadvantage under high demand. When demand exceeds 2100 veh/(h·lane), the CL under NCCS even becomes larger than that under the NC baseline. As shown in Fig. 6(b), congestion is highly uneven across lanes: the closed lane quickly becomes severely congested, the adjacent lane experiences moderate congestion, while the innermost lane remains uncongested. This suggests that balancing and regulating lane-wise traffic distribution is critical under high-demand conditions. In particular, it is important to avoid severe congestion in one lane while another lane remains underutilized. Our proposed method not only regulates inflow, but also emphasizes "flow redistribution". Specifically, CAVs cooperate as multi-lane fleets to reshape the lane-wise traffic distribution upstream of the bottleneck and guide traffic toward the open lanes. This enables the proposed method to maintain better performance under high demand.

D. Ablation Study

To further evaluate the contribution of different components in the proposed framework, three ablation settings are considered in this study.

(1) Without MPC: The lower-level MPC controller is removed and replaced by a simple rule-based speed control strategy. Specifically, for the CAV in lane 1, the target speed is directly set to the fleet speed generated by the upper-level controller. For fleet CAVs in the other lanes, each CAV first adjusts its motion to track the target spacing relative to the neighboring-lane CAV, and then sets its target speed to the fleet speed.

(2) Without RL: The upper-level RL-based fleet parameter decision module is removed, which is consistent with the setting of the FPS baseline.

(3) Without multi-lane fleet: The multi-lane fleet-based CAV cooperation structure introduced in Section III is removed. Instead, only single-lane platoons are adopted as the mobile actuators for Lagrangian traffic control, which is a commonly used setting in previous studies such as Ref. [17]. Specifically, three consecutive CAVs in lane 1 form one platoon, and a total of 8 such platoons are introduced in the traffic flow (consistent with the setting of $N_{\mathrm{m}} = 8$ in the proposed method). The platoon speed is determined by RL in this setting. Except for the modified action space, all other method settings, training hyperparameters, and testing configurations remain identical to those used in the proposed framework.

The ablation experiments are conducted under three traffic demand levels, namely 2000, 2300, and 2600 veh/(h·lane), and the comparison results are shown in Fig. 8.

For the "without RL" setting, the main limitation is that the fleet parameters cannot be dynamically adjusted according to the real-time macroscopic traffic state and information. As a result, the fleets cannot fully exploit their capability to regulate traffic flow under varying traffic conditions. Significant performance degradation can be observed in terms of ATD, ATS, and CL. In particular, under high traffic demand, "without RL" achieves the worst ATS among all settings. This is mainly because the fleet speed and spacing parameters remain fixed and cannot be adaptively updated to maintain traffic speed at an appropriate level upstream of the bottleneck.

For the "without MPC" setting, the performance degradation is most noticeable under relatively low traffic demand. This is mainly because the fleet CAVs are no longer governed by an effective cooperative tracking controller, making it difficult for the fleets to accurately and rapidly track the parameters generated by the upper-level controller. Moreover, the coarse rule-based tracking strategy may lead to unreasonable vehicle maneuvers in some situations (e.g., sudden deceleration), which can introduce additional disturbances into the traffic flow and negatively affect traffic stability.

For the "without multi-lane fleet" setting, the method exhibits the worst ATD and CL among all settings. This is mainly because, without the proposed multi-lane fleet struc-

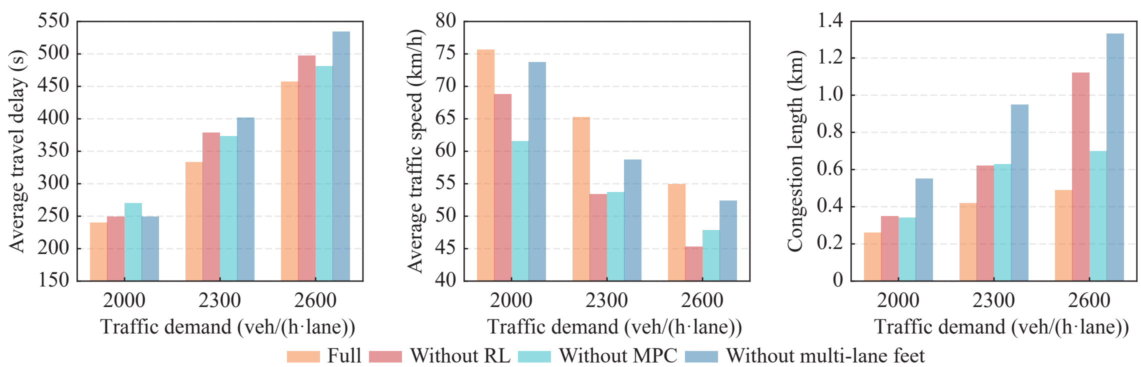


Fig. 8 Ablation study results under traffic demands of 2000, 2300, and 2600 veh/(h·lane). These three subplots report ATD, ATS, and CL, respectively. The proposed full framework is compared with three variants: without RL, without MPC, and without multi-lane fleet.

ture, the control effect of the single-lane platoons is largely limited to the lane where the platoons are deployed, making it difficult to effectively influence the other uncontrolled lanes. Consequently, the method cannot effectively regulate the overall multi-lane traffic flow or redistribute traffic from the closed lane to the remaining open lanes. As a result, congestion continues to develop and propagate, leading to severe traffic delay and congestion accumulation.

E. Analysis of Automation-Level Impact on Regulation Performance

In Lagrangian traffic control, CAVs in mixed traffic serve as mobile actuators for traffic regulation; therefore, the availability of CAVs has a significant impact on control performance. Under different levels of system automation, the CAV penetration rate affects the availability of CAVs in mixed traffic and consequently influences the number of CAV fleets that can be formed, i.e., the number of available actuators in the proposed control framework. To evaluate the effect of automation levels on the proposed method, experiments are conducted under different numbers of available CAV fleets, corresponding to low, medium, and high fleet availability conditions. Since the number of available fleets is closely related to the proportion of available CAVs in traffic, this analysis can also provide insight into the effectiveness of the proposed framework under different CAV penetration levels. The results are presented in Table 3.

Table 3 Comparison of regulation performance under different numbers of available fleets and traffic demand levels. The arrows indicate the preferred direction of each metric: "↓" means lower values are better, and "↑" means higher values are better.

Demand level (veh/(h·lane))	Number of fleets N_m	ATD↓ (s)	ATS↑ (km/h)	CL↓ (km)
1800	4	187	78.10	0.26
	8	178	80.59	0.20
	12	176	80.66	0.16
2200	4	320	65.31	0.50
	8	299	68.10	0.41
	12	291	68.47	0.22
2600	4	501	46.16	0.78
	8	457	54.91	0.49
	12	452	55.12	0.46

Across the three considered traffic demand levels, the regulation performance of the proposed method consistently improves as the number of available fleets increases. Under low demand conditions (1800 veh/(h·lane)), increasing the number of fleets from 4 to 12 yields only marginal performance gains. This suggests that when traffic demand is low, a relatively small number of fleets is sufficient to accomplish the required regulation task. In contrast, under higher demand levels, congestion is more likely to occur and propagate. In such cases, involving a larger number of fleets significantly enhances regulation effectiveness. Specifically, in terms of ATD, under low demand (1800 veh/(h·lane)), the improvements achieved by 8-fleet and 12-fleet cases over the 4-fleet case are only 4.8% and 5.9%, respectively. However, under high demand conditions, the improvements increase to 8.9% and 9.8%. Regarding CL, the 12-fleet case reduces congestion length by 56% and 41% compared to the 4-fleet case under demand levels of 2200 and 2600 veh/(h·lane), respectively.

To further illustrate these effects, the density contour plots for the 4-fleet and 12-fleet cases under a demand level of 2600 veh/(h·lane) are shown in Fig. 9, while the 8-fleet case is presented in Fig. 6(d). It can be observed that increasing the number of participating fleets effectively increases the number of available actuators, thereby providing greater control flexibility. This results in more refined traffic regulation, smoother traffic flow, and stronger suppression of congestion formation and propagation. Notably, although a larger number of fleets entails higher control overhead, even with 12 available fleets, only approximately 3% of all vehicles in the traffic stream are utilized for control under a demand level of 2600 veh/(h·lane).

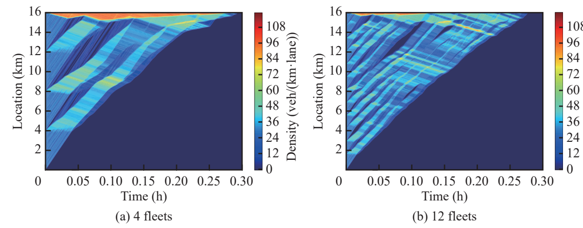


Fig. 9 The contour plots of traffic flow density under the proposed strategy with different numbers of available fleets under the demand level of 2600 veh/(h·lane).

F. Performance Analysis

To assess the contribution of different reward components in Eq. (4), we vary the weight coefficients ζ_1, ζ_2, ζ_3, and ζ_4 and train multiple agents under different reward settings. The training performance and resulting traffic regulation outcomes are then compared to evaluate the impact of reward design. Note that the four terms in the reward function can be grouped into two categories. The first group, r_{thp} and r_{vel}, directly targets the primary objectives of improving bottleneck throughput and increasing average traffic speed. The second group, r_{dist} and r_{cong}, encourages vehicles to move out of the closed lane in advance and penalizes the formation and growth of congestion. Compared with the first group, these terms mainly serve as intermediate guidance that helps the agent discover behaviors conducive to overall performance improvement.

To examine the roles of these two groups, we retrain multiple agents by varying the relative magnitudes of their corresponding weight coefficients. The specific weight configurations and the resulting performance of the retrained agents are summarized in Table 4. Among these configurations, two cases set selected coefficients to zero, effectively removing the corresponding reward terms and serving as ablation cases: one retaining only the first group, which removes the density distribution and congestion penalties ($\zeta_3 = \zeta_4 = 0$), and the other retaining only the second group, which removes the throughput and speed rewards ($\zeta_1 = \zeta_2 = 0$). The training reward curves for these ablation cases are shown in Fig. 10.

It can be observed that when the second reward group is ablated, the agent fails to efficiently discover an effective strategy during training. As shown in Fig. 10(a), the reward fluctuates significantly throughout the training process. This is mainly because the relationship between the agent's actions and the resulting improvements in traffic speed and throughput is highly indirect, making the reward signal weakly informative and difficult for the agent to exploit. As shown in Table 4, because congestion penalties are not considered, the CL metric exhibits the worst performance. When ζ_3 and ζ_4 are reduced rather than completely removed, CL shows improvement. In contrast, appropriately scaled weights ($\zeta_1 = 1$, $\zeta_2 = 2$, $\zeta_3 = 1$, and $\zeta_4 = 1.5$) not only mitigate congestion but also improve ATD and ATS. This is mainly because the inclusion of congestion-related penalties guides the model to efficiently discover effective strategies, and the reduction in congestion further helps to improve the average speed.

Table 4 Performance of agents trained under different reward designs. The bold is the best result. The arrows indicate the preferred direction of each metric: "↓" means lower values are better, and "↑" means higher values are better.

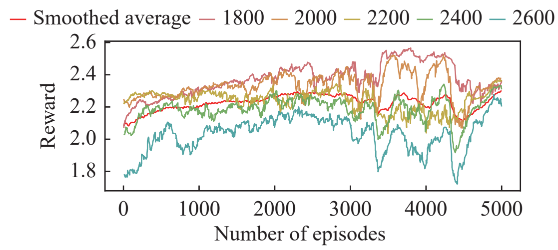
ζ_1	ζ_2	ζ_3	ζ_4	ATD (s) ↓	ATS (km/h) ↑	CL (km) ↓
1.0	2.0	0.0	0.0	486	50.33	1.02
1.0	2.0	0.3	0.5	489	50.69	0.70
1.0	2.0	1.0	1.5	**457**	**54.91**	**0.49**
0.3	0.6	1.0	1.5	483	49.82	0.57
0.0	0.0	1.0	1.5	501	48.13	0.50

On the other hand, when the first reward group is ablated, the reward signal becomes more direct and immediately informative. As shown in Fig. 10(b), the reward quickly enters a stable increasing phase and continues to improve smoothly throughout training. However, because the reward solely penalizes congestion length and the flow distribution on the closed lane, the learned policy tends to reduce congestion by slowing down traffic and preventing vehicles from entering the bottleneck region. As a result, although the congestion length is effectively suppressed, neither the average traffic speed nor the overall throughput is significantly improved. This outcome does not satisfy the fundamental objective of bottleneck decongestion, which aims to enhance traffic efficiency rather than merely restrict inflow.

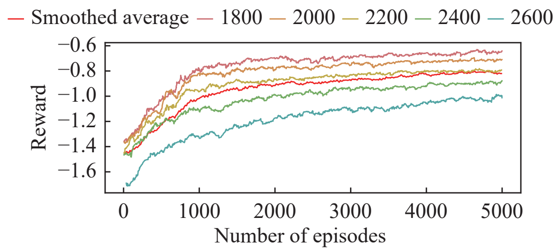
Overall, when all four reward components in Eq. (4) are jointly applied and properly balanced, the first two terms represent the ultimate performance objectives, while the latter two act as shaping signals that guide the learning process. Under their combined effect, the agent not only learns effective regulation strategies more efficiently, but also achieves superior traffic performance.

VI. Conclusions

This paper presents a learning-based Lagrangian traffic control framework for mitigating congestion in multi-lane bottleneck scenarios. By leveraging a small fraction of CAVs organized into cooperative cross-lane fleets, the proposed method enables effective system-level traffic regulation without relying on explicit mechanistic models. The hierarchical architecture integrates an upper-level RL policy for adaptive decision-making with a lower-level MPC controller for real-time motion tracking, thereby bridging macroscopic traffic



(a) Without density distribution and congestion penalties



(b) Without throughput and speed rewards

Fig. 10 Training reward curves under different reward ablation settings. The red curve represents the smoothed average reward across all training episodes, while the colored curves correspond to episodes under different traffic demand levels (1800–2600 veh/(h·lane)).

objectives and microscopic vehicle dynamics. Simulation results demonstrate that the proposed approach consistently improves traffic efficiency across a wide range of demand conditions, achieving significant reductions in travel delay and congestion while maintaining robust performance under dynamic and uncertain traffic environments. Furthermore, the results show that the proposed method is particularly effective under high-demand conditions, where cross-lane coordination significantly suppresses congestion. It also demonstrates strong generalization across unseen demand levels and improved performance with increased CAV fleet availability.

Several directions remain open for further investigation. The current framework adopts a centralized decision-making architecture with a single agent at the traffic control center. This framework may face limitations in scalability and reliability, especially as the number of fleets and the spatial extent of the network increase. A promising direction is to explore multi-agent reinforcement learning frameworks while still achieving cooperative global behavior. Moreover, the generalization capability of the trained policy to unseen environments has not yet been evaluated, including different road geometries, multi-bottleneck configurations, different closed-lane settings, as well as heterogeneous HDV driving behaviors. Future work will investigate methods to improve cross-scenario transferability and robustness under these more complex and realistic traffic conditions.

References

[1] T. T. Nguyen, P. Krishnakumari, S. C. Calvert, H. L. Vu, and H. Van Lint, Feature extraction and clustering analysis of highway congestion, *Transp. Res. Part C: Emerg. Technol.*, 2019, 100, 238–258.

[2] S. Siri, C. Pasquale, S. Sacone, and A. Ferrara, Freeway traffic control: A survey, *Automatica*, 2021, 130, 109655.

[3] K. Yuan, V. L. Knoop, and S. P. Hoogendoorn, Capacity drop: Relationship between speed in congestion and the queue discharge rate, *Transp. Res. Rec.*, 2015, 2491(1), 72–80.

[4] W. Jin, Performance analysis and control design for on-ramp metering of active merging bottlenecks [Online], https://rip.trb.org/view/1441909, 15 April 2026.

[5] Y. Chen and L. Li, *Advances in Intelligent Vehicles*. Amsterdam, The Netherlands: Elsevier, 2013.

[6] C. Creß, Z. Bing, and A. C. Knoll, Intelligent transportation systems using roadside infrastructure: A literature survey, *IEEE Trans. Intell. Transp. Syst.*, 2024, 25(7), 6309–6327.

[7] D. Cao, J. Wu, J. Wu, J. Wu, B. Kulcsár, and X. Qu, A platoon regulation algorithm to improve the traffic performance of highway work zones, *Comput.-Aided Civ. Infrastruct. Eng.*, 2021, 36(7), 941–956.

[8] T. T. Phan, L. B. Le, and D. Ngoduy, A cooperative space distribution method for autonomous vehicles at a lane-drop bottleneck on multi-lane freeways, *IEEE Trans. Intell. Transp. Syst.*, 2022, 23(4), 3710–3723.

[9] D. Cao, X. Wang, L. Li, C. Lv, X. Na, Y. Xing, X. Li, Y. Li, Y. Chen, and F.-Y. Wang, Future directions of intelligent vehicles: Potentials, possibilities, and perspectives, *IEEE Trans. Intell. Veh.*, 2022, 7(1), 7–10.

[10] J. Li, C. Yu, Z. Shen, Z. Su, and W. Ma, A survey on urban traffic control under mixed traffic environment with connected automated vehicles, *Transp. Res. Part C: Emerg. Technol.*, 2023, 154, 104258.

[11] T. Ersal, I. Kolmanovsky, N. Masoud, N. Ozay, J. Scruggs, R. Vasudevan, and G. Orosz, Connected and automated road vehicles: State of the art and future challenges, *Veh. Syst. Dyn.*, 2020, 58(5), 672–704.

[12] G. Piacentini, P. Goatin, and A. Ferrara, Traffic control via moving bottleneck of coordinated vehicles, *IFAC-PapersOnLine*, 2018, 51(9), 13–18.

[13] M. Čičić and K. H. Johansson, Traffic regulation via individually controlled automated vehicles: A cell transmission model approach, in *Proc. 21st International Conference on Intelligent Transportation Systems*, Maui, HI, USA, 2018, 766–771.

[14] E. Vinitsky, K. Parvate, A. Kreidieh, C. Wu, and A. Bayen, Lagrangian control through deep-RL: Applications to bottleneck decongestion, in *Proc. 21st International Conference on Intelligent Transportation Systems*, Maui, HI, USA, 2018, 759–765.

[15] E. Vinitsky, N. Lichtlé, K. Parvate, and A. Bayen, Optimizing mixed autonomy traffic flow with decentralized autonomous vehicles and multi-agent reinforcement learning, *ACM Trans. Cyber-Phys. Syst.*, 2023, 7(2), 13.

[16] G. Piacentini, C. Pasquale, S. Sacone, S. Siri, and A. Ferrara, Multiple moving bottlenecks for traffic control in freeway systems, in *Proc. 18th European Control Conference*, Naples, Italy, 2019, 3662–3667.

[17] G. Piacentini, A. Ferrara, I. Papamichail, and M. Papageorgiou, Highway traffic control with moving bottlenecks of connected and automated vehicles for travel time reduction, in *Proc. 58th Conference on Decision and Control*, Nice, France, 2019, 3140–3145.

[18] M. Cicic, X. Xiong, L. Jin, and K. H. Johansson, Coordinating vehicle platoons for highway bottleneck decongestion and throughput improvement, *IEEE Trans. Intell. Transp. Syst.*, 2022, 23(7), 8959–8971.

[19] S. C. Vishnoi, J. Ji, M. Bahavarnia, Y. Zhang, A. F. Taha, C. G. Claudel, and D. B. Work, Exploring CAV-based traffic control for improving traffic conditions in the face of bottlenecks, in *Proc. 62nd IEEE Conference on Decision and Control*, Singapore, Singapore, 2023, 98–104.

[20] C. Daini, P. Goatin, M. L. D. Monache, and A. Ferrara, Centralized traffic control via small fleets of connected and automated vehicles, in *Proc. European Control Conference*, London, UK, 2022, 371–376.

[21] S. Yang, Y. Xu, and D. Li, Valve fleets: A novel control method for mixed traffic flow regulation with applications in bottleneck segments, *IEEE Trans. Intell. Transp. Syst.*, 2025, 26(12), 23215–23230,

[22] Y. Yan, Y. Feng, J. Wang, H. Zhang, and G. Yin, Decision-making and planning for intelligent vehicle considering human factors: Methods, challenges, and prospects, *IEEE Trans. Intell. Transp. Syst.*, 2026, 27(1), 145–168.

[23] Y. Yan, P. Zhang, C. Du, H. Wang, H. Wang, D. Pi, and Y.-H. Chen, Vehicle trajectory prediction for autonomous driving applications: State-of-the-art review, research challenges, and future directions, *Automot. Innovation*, 2026, 9(2), 263–295.

[24] H. Wei, X. Liu, L. Mashayekhy, and K. Decker, Mixed-autonomy traffic control with proximal policy optimization, in *Proc. IEEE Vehicular Networking Conference*, Los Angeles, CA, USA, 2019, 1–8.

[25] C. Wu, A. R. Kreidieh, K. Parvate, E. Vinitsky, and A. M. Bayen, Flow: A modular learning framework for mixed autonomy traffic, *IEEE Trans. Robot.*, 2022, 38(2), 1270–1286.

[26] X. Liu, E. Apriaskar, and L. Mihaylova, Deep reinforcement learning method for control of mixed autonomy traffic systems, in *Proc. IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, Pilsen, Czech Republic, 2024, 1–6.

[27] P. Ha, S. Chen, J. Dong, and S. Labi, Leveraging vehicle connectivity and autonomy for highway bottleneck congestion mitigation using reinforcement learning, *Transportmetrica A: Transp. Sci.*, 2025, 21(1), 2215338.

[28] Y. Yan, L. Peng, T. Shen, J. Wang, D. Pi, D. Cao, and G. Yin, A multi-vehicle game-theoretic framework for decision making and planning of autonomous vehicles in mixed traffic, *IEEE Trans. Intell. Veh.*, 2023, 8(11), 4572–4587.

[29] L. Liu, M. Wang, M.-O. Pun, and X. Xiong, A multi-agent rollout approach for highway bottleneck decongestion in mixed autonomy, in *Proc. 27th International Conference on Intelligent Transportation Systems*, Edmonton, Canada, 2024, 3377–3382.

[30] Y. Veksler, S. Hornstein, H. Wang, M. L. D. Monache, and D. Urieli, Cooperative cruising: Reinforcement learning-based time-headway control for increased traffic efficiency, arXiv preprint arXiv: 2412.02520, 2024.

[31] Y. Lin, R. Yan, R. Jiang, S. Zheng, and H. Zhao, Reinforcement learning-based control algorithm for connected and automated vehicles on a two-lane highway section with a moving bottleneck, *Urban Lifeline*, 2025, 3(1), 19.

[32] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y. P. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner, and E. Wiessner, Microscopic traffic simulation using SUMO, in *Proc. 21st International Conference on Intelligent Transportation Systems*, Maui, HI, USA, 2018, 2575–2582.

[33] Gurobi Optimization, Gurobi optimizer reference manual [Online], https://www.gurobi.com, 15 April 2026.

Siwen Yang received the BE degree in automation from Central South University, Changsha, China, in 2022, and the ME degree in control engineering from Shanghai Jiao Tong University, Shanghai, China, in 2025. He is currently pursuing the PhD degree in electrical and computer engineering at Purdue University, IN, USA. His research interests include reinforcement learning and intelligent transportation systems.

Guanxiao Li received the MS degree in electrical and computer engineering from Purdue University, West Lafayette, IN, USA, in 2025. She is currently pursuing the PhD degree in electrical and computer engineering at Purdue University, Indianapolis, IN, USA. Her research interests include autonomous driving, reinforcement learning, and foundation model.

Zhitong He received the MS degree in electrical and computer engineering from The Ohio State University, Columbus, OH, USA, in 2021. He is currently pursuing the PhD degree in electrical and computer engineering at Purdue University, Indianapolis, IN, USA. His research interests include autonomous driving systems, motion planning and control, and dynamic traffic systems under uncertainty.

Lingxi Li received the PhD degree in electrical and computer engineering from University of Illinois at Urbana-Champaign, USA, in 2008. He is currently a professor at Elmore Family School of Electrical and Computer Engineering, Purdue University, USA. He has authored/co-authored one book and over 180 research articles in refereed journals and conferences. He is currently serving as a senior/associate editor for six international journals and has served as the general chair, the program chair, the program co-chair, etc. for more than 30 international conferences. His research interests include modeling, analysis, control, and optimization of complex systems, connected and automated vehicles, intelligent transportation systems, intelligent vehicles, discrete event dynamic systems, human machine interactions, and parallel intelligence.