

Joint longitudinal-lateral trajectory planning for CAVs in mixed traffic at signalized intersections

Xingwei Jiang¹, Meng Li^{1,2}, Qingquan Liu³✉

 **Cite this article:** Jiang X W, Li M, Liu Q Q. *Commun Transp Res* 2026, 6(1): 9640011. <https://doi.org/10.26599/COMMTR.2026.9640011>

ABSTRACT: Mandatory lane changes (MLCs) pose significant challenges to trajectory planning at intersections, where vehicles are required to change lanes mid-block to reach designated turn lanes before the stop bar. MLCs often generate shockwaves that induce increased vehicle delay and fuel consumption, and the presence of human-driven vehicles (HDVs) in mixed traffic further exacerbates this issue. To address these challenges, this study formulates the joint longitudinal-lateral trajectory planning problem in mixed traffic as a multiagent reinforcement learning (MARL) task. We propose SS-MA-PPO, a Simulation-Supervised Multi-Agent Proximal Policy Optimization framework, which guides connected and autonomous vehicles (CAVs) in both acceleration and lane-change decisions. A Simulation-Guided Supervisory Module (SGSM) performs offline trajectory rollouts of human-driver models to assess feasibility and safety, and arbitrates online between rule-based and learned policies. The information of surrounding vehicles is incorporated in the observation to achieve vehicle cooperation, and a transfer learning mechanism is designed to accelerate training. Experiments using a real-world dataset from Langfang, China demonstrate that SS-MA-PPO outperforms both conventional and MARL baselines across various evaluation metrics. Ablation experiments verify the substantial effectiveness of the proposed SGSM module, vehicle cooperation, and transfer learning, achieving enhanced performance and faster training convergence.

KEYWORDS: mandatory lane changes (MLCs); connected and autonomous vehicles (CAVs); longitudinal-lateral trajectory planning; mixed traffic; multiagent reinforcement learning (MARL)

1 Introduction

The rapid rise in vehicle ownership has led to severe traffic congestion in urban areas, where signalized intersections often serve as critical bottlenecks (Gao et al., 2025; Li et al., 2023). The mismanagement of these intersections can significantly diminish overall traffic efficiency and safety and lead to increased emissions (Leng et al., 2025; Liu et al., 2025; Yang et al., 2026; Ye et al., 2019). Fortunately, advancements in connected and autonomous vehicles (CAVs) have created opportunities to improve traffic conditions at intersections. On the one hand, vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communications enable perception information for the entire traffic stream to be shared among all connected vehicles (Chen et al., 2025). On the other hand, automated driving technology supports precise vehicle trajectory control, fostering more rational driving behaviors (Liu et al., 2023a, 2023b).

Vehicle trajectory planning is a major topic in traffic control with CAVs (Li et al., 2023). Most studies adjust acceleration and deceleration so that vehicles traverse intersections during green phases and achieve smooth stop-and-go transitions. Early efforts in longitudinal planning for CAVs typically rely on advanced car-following models, such as cooperative adaptive cruise control (CACC) (Milanés et al., 2014; Zhou et al., 2022), which enables

the stable formation of CAV platoons. At intersections, these platoon control methods must be adapted to comply with traffic signals (Chen et al., 2022; Zhao et al., 2018). In addition to platoon-based control, several studies have focused on longitudinal control for individual vehicles. Optimal control models have been formulated for each individual vehicle and solved by approaches such as the shooting heuristic algorithm (Ma et al., 2017; Zhou et al., 2017), dynamic programming (Wang et al., 2022a; Xiong and Jiang, 2022), or commercial solvers (Yao and Li, 2021). These methods demonstrate the potential for reducing vehicle delay and energy consumption.

Such studies generally assume the absence of lane changes and concentrate on optimizing longitudinal trajectories within a single lane (Wu and Qu, 2022). However, lane-changing maneuvers are inevitable at intersections. Because vehicles must reach designated turn lanes before the stop line (Wu et al., 2021) yet may start in arbitrary lanes, mandatory lane changes (MLCs) naturally arise. This compulsory requirement can nonetheless exert a substantial negative impact on traffic. Even one discretionary lane change can affect multiple surrounding vehicles (Jokhio et al., 2024), disrupt their normal car-following behavior for tens of seconds (He et al., 2023a; Zheng, 2014), and trigger traffic oscillations and capacity reductions (Laval and Daganzo, 2006). The problem is magnified

¹ Department of Civil Engineering, Tsinghua University, Beijing 100084, China. ² State Key Laboratory of Intelligent Green Vehicle and Mobility, Tsinghua University, Beijing 100084, China. ³ College of Mechanical and Vehicle Engineering, Chongqing University, Chongqing 400044, China.

✉ Corresponding author. E-mail: liuqq@cqu.edu.cn

Received: September 23, 2025; Revised: November 27, 2025; Accepted: January 6, 2026

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

at intersections, where MLCs are often required to be executed within short, congested approach lanes, further degrading traffic efficiency and safety (Ali et al., 2019; Pan et al., 2016).

To mitigate the adverse effects of MLCs at intersections, recent studies have integrated both longitudinal and lateral maneuvers into multivehicle trajectory-planning frameworks. Some model predictive control (MPC) approaches utilize mixed-integer programming (MIP) to jointly optimize acceleration profiles and lane-changing sequences (Liu et al., 2024; Lu et al., 2019; Yang et al., 2024). Others use discretized state-action representations for CAV platoons, executing lane changes after coordinating longitudinal spacings among vehicles (Cai et al., 2022; Duan et al., 2023; Liu et al., 2023a; Wu et al., 2021). However, these studies are confined to fully CAV environments. In mixed traffic containing both human-driven vehicles (HDVs) and CAVs, accurately modeling HDV behavior—particularly lane-changing maneuvers—remains challenging (Guo et al., 2021). Consequently, integrated longitudinal-lateral trajectory planning for multiple vehicles in mixed traffic is still scarce (Ma et al., 2021; Ying and Feng, 2024).

In summary, these optimization-based methods show promise for enhancing trajectory smoothness and intersection capacity under various conditions. However, accurately modeling the car-following and lane-changing behaviors of HDVs with explicit models is difficult, and their performances are highly sensitive to model accuracy (Lee and Choi, 2023; Zhou et al., 2023). Moreover, optimization models often entail high computational complexity, which impedes real-time implementation in complex scenarios (Li and Li, 2019).

Recently, multiagent reinforcement learning (MARL) has emerged as a promising solution for multivehicle trajectory planning (Haydari and Yilmaz, 2022; Peng et al., 2021; Zhang et al., 2024). By learning control strategies directly through interactions with the environment, MARL obviates explicit human-behavior modeling and requires only limited online computational effort. MARL-based methods have been successfully applied to scenarios such as on-ramp merging (Chen et al., 2023) and highway driving (He et al., 2023b; Leng et al., 2025; Zhang et al., 2023). For signalized intersections, MARL has also been widely employed. Zhou et al. (2020) trained a DDPG-based model in a circular intersection environment. Wang et al. (2022b) integrated a simulation-based prediction module into an MARL controller to reduce stopping and harsh deceleration before the stop line. Leng et al. (2025) modeled discretionary lane changes of individual vehicles across multiple intersections. Nevertheless, most existing MARL-based approaches overlook MLCs at intersections, as they often assume vehicles are initially positioned in their target lanes. In addition, the “black-box” nature of deep reinforcement learning raises concerns about robustness and reliability. Finally, MARL often struggles to converge, particularly when the number of agents is large.

To bridge this research gap, we propose a MARL-based approach, focusing on multivehicle trajectory planning in mixed-traffic environments at signalized intersections, where MLCs are explicitly considered. Specifically, we propose SS-MA-PPO, a simulation-supervised multiagent proximal policy optimization algorithm that controls both longitudinal acceleration and lateral lane-changing maneuvers. The proposed approach integrates a supervisor module to enhance robustness and a transfer-learning mechanism to facilitate convergence. Our main contributions are summarized as follows:

1) We utilize a MARL approach for the longitudinal-lateral trajectory-planning task. The formulation explicitly incorporates the mandatory lane-changing requirement induced by

intersection lane guidance. This approach is applied to mixed-traffic environments containing both CAVs and HDVs, and HDV behaviors are implicitly captured through interactions in the training environment.

2) We design a simulation-guided supervisory module (SGSM) that arbitrates between the learned MARL controller and a complementary rule-based controller. SGSM enforces safety during exploration and deployment, improves robustness, and significantly improves the overall performance of the algorithm.

3) We enrich each agent’s state with observations of surrounding vehicles to promote cooperation and higher-quality decisions. In addition, we employ a transfer-learning mechanism in which agents are first pretrained in a simplified, low-density scenario and then transferred to complex, real-world traffic environments. This combination accelerates policy convergence and improves training efficiency and final performance in real-world scenarios.

The remainder of this study is organized as follows. Section 2 reviews the fundamentals of reinforcement learning and surveys key paradigms in cooperative MARL. Section 3 frames joint longitudinal-lateral vehicle control at signalized intersections as a partially observable MARL problem, details the learning framework, and specifies the neural network architecture adopted for the shared parameter. Section 4 introduces the Simulation-Guided Supervisory Module, which accelerates training and enforces safety during exploration. Section 5 describes the experimental setup and compares the proposed method with rule-based and MARL baselines on a real-world intersection dataset. Finally, Section 6 concludes the study and outlines future research directions.

2 Preliminaries

2.1 Reinforcement learning fundamentals

Reinforcement learning (RL) is a paradigm in which an agent learns to make sequential decisions by interacting with an environment to maximize cumulative rewards.

The RL process is typically modeled as a Markov decision process (MDP), characterized by the tuple $(\mathcal{S}, \mathcal{U}, P, r, \gamma)$, where $\mathcal{S}$ denotes the set of environment states, $\mathcal{U}$ denotes the set of admissible control actions, $P(\hat{s}|s, u)$ is the transition probability of moving from state s to state $\hat{s}$ when action u is taken, $r(s, u)$ is the immediate reward obtained after executing action u in state s , and $\gamma \in [0, 1]$ is the discount factor that balances immediate and future rewards.

At each time step t , the agent observes the current state s_t , selects a control action u_t according to the policy $\pi(u_t|s_t)$, receives a reward r_t , and transitions to the next state s_{t+1} .

The objective is to find an optimal policy π^* that maximizes the expected discounted cumulative return:

$$J(\pi) = E_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, u_t) \right] \quad (1)$$

Among these methods, proximal policy optimization (PPO) has gained prominence because of its stability and sample efficiency.

The PPO updates the policy by maximizing a clipped surrogate objective that prevents overly large parameter changes:

$$L^{\text{CLIP}}(\theta) = E_t \left[\min \left(r_t(\theta) \hat{A}_t, \text{CLIP} \left(r_t(\theta), 1 - \epsilon, 1 + \epsilon \right) \hat{A}_t \right) \right] \quad (2)$$

where $r_t(\theta) = \frac{\pi_\theta(u_t|s_t)}{\pi_{\theta_{old}}(u_t|s_t)}$ is the probability ratio between the

new and old policies, $\hat{A}_t$ is the estimated advantage, and ϵ is the clipping hyperparameter.

2.2 Multi-agent reinforcement learning

MARL extends the classical reinforcement learning framework to environments involving multiple learning agents that interact simultaneously. Compared to the single agent setting, MARL introduces additional challenges, such as nonstationarity of the environment, credit assignment among agents, and the need for cooperation or competition under partial observability (Amato, 2024).

A fundamental categorization of MARL approaches can be made along two axes. First, in cooperative MARL, agents aim to maximize a shared return, often under constraints that require explicit coordination. In noncooperative or competitive settings, agents optimize their own rewards, sometimes at the expense of others (Du, 2024; Littman, 1994). Second, approaches can also be distinguished by whether interagent communication is fixed and centralized during training, dynamic, and learned, or completely decentralized and implicit.

With the growth of deep learning, several deep MARL algorithms have been proposed.

Multiagent Deep Deterministic Policy Gradient (MADDPG) is one of the earlier deep MARL methods following the centralized training with decentralized execution (CTDE) framework. It allows each agent to have its own policy while employing a centralized critic that observes the observations and actions of all agents (Lowe et al., 2017). This helps stabilize training in partially observable settings but is limited in scalability and sample efficiency.

Multi-Agent A2C (MAA2C) is a multiagent actor-critic method in which the critic learns a joint state-value function (Papoudakis et al., 2021). It extends the on-policy A2C framework by using a centralized critic that takes the global environment state as input.

Recently, Independent Proximal Policy Optimization (IPPO) has gained attention for its improved stability and scalability. It extends the popular PPO algorithm to the multiagent setting by treating all other agents as part of the environment (Witt et al., 2020). It supports parameter sharing and modular policy structures, making it particularly suitable for large-scale and partially cooperative tasks (Wang et al., 2024).

Although centralized joint learning mitigates partial observability and nonstationarity, it suffers from a joint action space that grows exponentially with the number of agents (Yu et al., 2021). In contrast, decentralized architectures are simpler to build and scale. Therefore, we choose IPPO as the core learning framework due to its proven ability to handle multiagent coordination, especially in environments such as traffic control where real-time, decentralized decision-making is essential. Conceptually, IPPO can be regarded as a multiagent implementation of PPO. Therefore, in the remainder of this study, we refer to it as MA-PPO for consistency.

2.3 Problem setting

The traffic environment is modeled at a microscopic level in SUMO under the following unified setting. Signal control operates on a fixed-time plan: each approach follows the four-phase timing diagram in Fig. 1, and signal splits remain constant throughout training and evaluation, with no adaptive phase adjustments. Vehicles enter the network as a Poisson arrival process; their

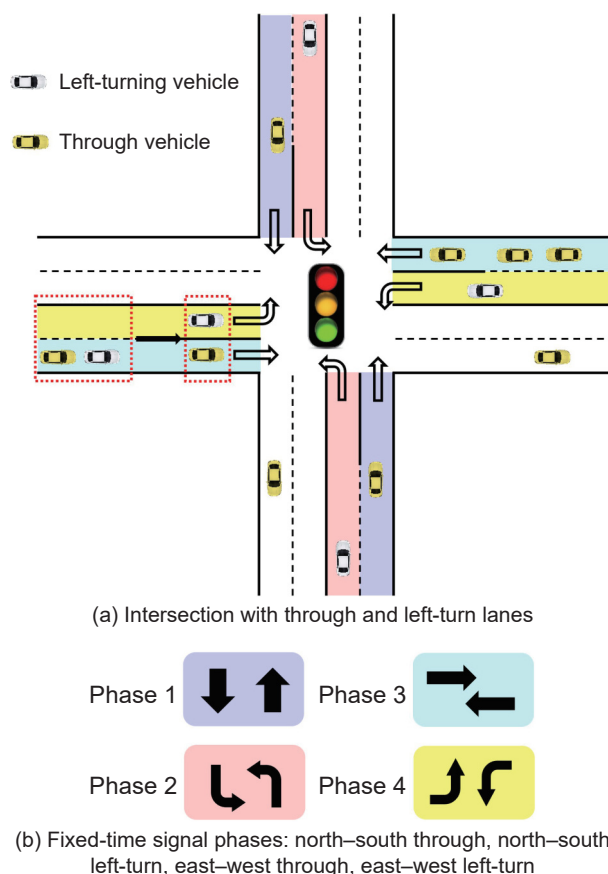


Fig. 1 Visualization of the experimental setup.

initial lanes are randomly assigned among feasible options, while target lanes are prescribed, so a subset of vehicles must execute mandatory lane changes. Every CAV is equipped with onboard sensing and V2X communication and, at each decision step, receives its own kinematic state, a high-fidelity indication of the current signal phase, and the most recent broadcasts of neighboring CAVs' positions, speeds, and intended high-level actions. HDV states are inferred from onboard perception, but their future actions are not observable.

3 Joint longitudinal-lateral trajectory planning at signalized intersections as a MARL problem

To illustrate the overall framework of our proposed MARL system for integrated longitudinal-lateral trajectory planning, we provide a schematic diagram in Fig. 2. Fig. 2 summarizes how CAVs and HDVs interact within the environment, the information flow among agents, and the key modules involved in state observation, action selection, and reward computation.

3.1 Formulation of the Markov decision process

In this section, we first model the mixed-traffic environment at signalized intersections as a partially observable Markov decision process (POMDP), where control actions are denoted by u to distinguish them from the physical acceleration variable a used elsewhere in this study. We then detail our MARL algorithm, which explicitly incorporates the control actions of surrounding vehicles into each agent's observation. We also present the design of our reward function.

We consider a model-free multiagent network represented by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each agent ($i \in \mathcal{V}$) communicates with neighbors $j \in \mathcal{N}_i = \{j \mid (i, j) \in \mathcal{E}\}$ via edges ε_{ij} . Let $\mathcal{S}$ and $\mathcal{U}$

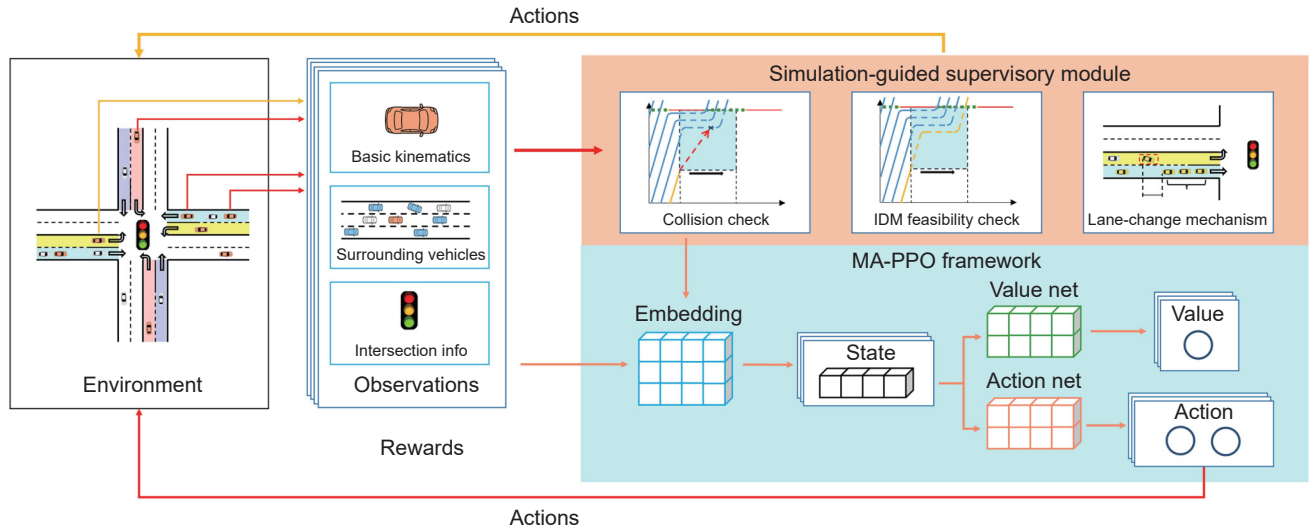


Fig. 2 Overview of the proposed multiagent framework for joint longitudinal-lateral trajectory planning in mixed traffic at signalized intersections.

denote the global state space and the joint control-action space, respectively. The environment evolves according to the transition distribution $P(s'|s, u)$ and a reward function $R(s, u)$.

A decentralized MARL framework is adopted, with each CAV acting as an individual agent. Because each agent has a limited sensing or communication range, it observes only a subset of the global state, resulting in a POMDP. Formally, we write:

$$(\mathcal{S}, \{\mathcal{O}_i\}_{i \in \mathcal{V}}, \{\mathcal{U}_i\}_{i \in \mathcal{V}}, P, \{R_i\}_{i \in \mathcal{V}}, \gamma) \quad (3)$$

Let $s \in \mathcal{S}$ be the global state. Each agent i receives a local observation $o_i \in \mathcal{O}_i = \mathcal{O}_i(s)$ and selects $u_i \in \mathcal{U}_i$. The environment evolves according to $P(s'|s, u)$, and agent i obtains reward R_i .

3.1.1 Observation space

Each agent i maintains a local observation $o_i(t) \in \mathcal{O}_i$, made up of four components.

- 1) Basic kinematics: current speed $v_i(t)$, acceleration $a_i(t)$, distance to the intersection $d_i(t)$, and estimated arrival time $e_i(t)$.
- 2) Intersection-related info: signal phase of the current lane and its derivative, lane offset.
- 3) Surrounding vehicles: relative distance, velocity differences, and accelerations of neighbors. In this work, neighbors are defined as five slots within the observation range selected by longitudinal proximity: the same-lane leader, the left-front vehicle, the left-rear vehicle, the right-front vehicle, and the right-rear vehicle. Fig. 3 illustrates an example in which the orange car is the ego agent and the blue cars are observable neighbors. Crucially, we also include each neighbor's most recent high-level action, such as a lane-change command or acceleration decision to foster coordinated maneuvers. For a nonexistent neighbor, we encode a large sentinel distance with zero relative speed, acceleration, and intent. For an existing HDV neighbor, kinematic features (relative distance and relative speed) are observed and retained, whereas unknown action features are padded with zeros.

- 4) Collision avoidance: A binary indicator $c_i(t)$ showing whether a collision with a preceding vehicle is imminent in the near future.

We denote the joint observation by $o(t) = \{o_1(t), \dots, o_N(t)\}$, constructed from each agent's local observation under partial observability.

3.1.2 Action space

Each agent i selects a control action $u_i(t) \in \mathcal{U}_i$ from a hybrid

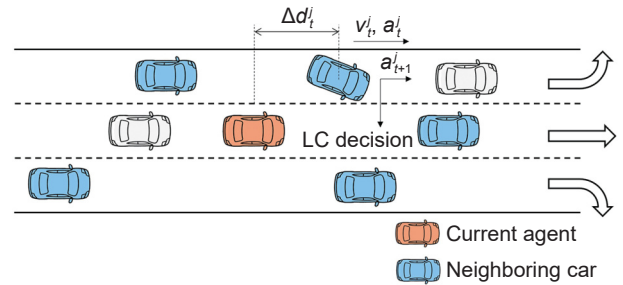


Fig. 3 Illustration of the local neighborhood observed by a CAV agent. The orange vehicle is the ego agent, and the blue vehicles are observable neighbors. Unoccupied neighbor slots or unknown HDV control actions are zero-padded in the observation vector.

discrete-continuous set. In particular, the discrete component indicates whether the vehicle changes to the left lane, remains in the current lane, or changes to the right lane, whereas the continuous component specifies a longitudinal acceleration a_i^{acc} . Formally, we define:

$$\mathcal{U}_i^{lat} = \{\text{LaneChangeLeft}, \text{KeepLane}, \text{LaneChangeRight}\} \quad (4)$$

$$\mathcal{U}_i^{long} = \{a_i^{acc} \in \mathbb{R}\} \quad (5)$$

$$\mathcal{U}_i = \mathcal{U}_i^{lat} \times \mathcal{U}_i^{long} \quad (6)$$

Once an action is chosen, the simulation environment (SUMO) applies the corresponding steering and throttle commands for a short time horizon. Consequently, the joint action space is $\mathcal{U} = \mathcal{U}_1 \times \dots \times \mathcal{U}_N$.

3.1.3 Reward function

A central component of our MARL formulation is the design of a multiterm reward function that promotes efficient, safe, and well-coordinated driving behaviors at signalized intersections. In particular, we incorporate speed tracking, lane-offset penalty, failed lane-change penalty, signal-aware arrival reward, and collision avoidance into a single scalar reward for each CAV agent. Let

$$r_i(t) = w_{\text{speed}} r_i^{\text{speed}}(t) + w_{\text{offset}} r_i^{\text{offset}}(t) + w_{\text{lc}} r_i^{\text{lc}}(t) + w_{\text{coll}} r_i^{\text{coll}}(t) + w_{\text{arriv}} r_i^{\text{arriv}}(t) \quad (7)$$

where $r_i^{\text{speed}}(t)$ encourages traveling at or below a target speed,

$r_i^{\text{offset}}(t)$ penalizes large lane offsets, $r_i^{\text{lc}}(t)$ penalizes failed lane-change attempts, $r_i^{\text{coll}}(t)$ rewards collision-free states, and $r_i^{\text{arriv}}(t)$ incentivizes timely passage through intersections during green phases. The nonnegative weights w_{speed} , w_{offset} , w_{lc} , w_{coll} , w_{arriv} control the relative importance of each term.

1) Speed tracking term. We define a piecewise speed reward:

$$r_i^{\text{speed}}(t) = \begin{cases} -1, & v_i(t) > v_{\text{target}} \\ \frac{v_i(t)}{v_{\text{target}}}, & \text{otherwise} \end{cases} \quad (8)$$

2) Lane-offset penalty. When a CAV is far from its ideal lane or forcing a lane change near the intersection, we apply a penalty that increases with the lane offset and how close the vehicle is to the intersection. Concretely, define

$$r_i^{\text{offset}}(t) = -|\Delta d_i(t)| \alpha_i(t) \quad (9)$$

where $\Delta d_i(t)$ represents the lateral offset of CAV i from its recommended lane center, and $\alpha_i(t) \in [0,1]$ is a linear factor proportional to the approach progress. $\alpha_i(t)$ is computed as

$$\alpha_i(t) = 1 - \frac{d_i(t)}{L_{\text{lane}}}, \alpha_i(t) \in [0,1] \quad (10)$$

where L_{lane} is the maximum approach distance. Therefore, lane-offset penalties become more severe if the vehicle is both misaligned and close to the intersection.

3) Failed lane-change penalty. We further track whether an attempted lane change is actually completed. If a CAV attempts a lane change but fails (e.g., due to conflicts or insufficient lateral gap), we assign a negative reward:

$$r_i^{\text{lc}}(t) = \begin{cases} -1, & \text{if an attempt fails} \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

4) Collision avoidance term. To ensure safety, each CAV is given a collision indicator $c_i(t) \in \{0,1\}$, which is set to 1 if a predicted collision with the leading vehicle is imminent. We then define:

$$r_i^{\text{coll}}(t) = \begin{cases} -1, & \text{if } c_i(t) = 1 \\ 1, & \text{otherwise} \end{cases} \quad (12)$$

This component strongly discourages actions that may lead to collisions while rewarding collision-free motion in the short horizon.

5) Signal-aware arrival reward. We adopt a piecewise function to encourage vehicles to arrive during favorable signal phases. The estimated arrival time is calculated by

$$T_{\text{arr}} = t_{\text{now}} + \frac{d_i}{v_i} \quad (13)$$

where t_{now} is the current simulation time, d_i is the distance from the intersection, and v_i is the current speed of the ego vehicle. As illustrated in Fig. 4, if T_{arr} falls within any of the first three green intervals, the reward linearly decays from 1 (at the start of the first green) to 0 (by the end of the third green). Otherwise, a value of -1 is assigned. This design motivates vehicles to adjust their trajectories to capitalize on early green windows, thereby reducing delay and queuing effects at the intersection.

Such a time-sensitive mechanism not only promotes efficient arrivals during green phases but also provides a strong disincentive for timing that would coincide with red signals or occur beyond the designated windows.

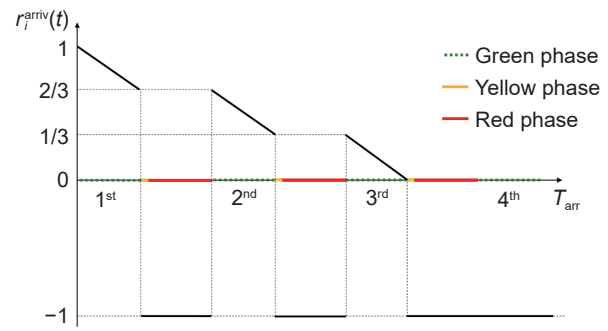


Fig. 4 Piecewise reward function that transitions from 1 to 0 across three green windows and assigns -1 outside these intervals.

6) Overall computation. The aggregated system reward $R(s(t), u(t))$ for multiagent training can be the sum:

$$R(s(t), u(t)) = \sum_{i=1}^N r_i(t) \quad (14)$$

By explicitly including speed, lane usage, collision avoidance, and signal-aware arrivals, this reward function steers the CAVs to learn behaviors that jointly optimize traffic flow, safety, and intersection throughput in mixed-traffic settings.

3.1.4 Transition probability

HDVs follow the intelligent driver model (IDM) (Treiber et al., 2000) for longitudinal motion and LC2013 (Erdmann, 2015) for lane-change decisions, while CAVs rely on the MARL policy for high-level control. The lower-level dynamic laws, including steering and acceleration, are then executed by the simulation environment without requiring explicit knowledge of transition probabilities in the learning algorithm.

At each decision step, every CAV proposes a MARL action $u_{i,\text{MARL}}$, while the SGSM may issue a rule action $u_{i,\text{rule}}$ when a scenario trigger is satisfied.

The environment-stepping (executed) action is therefore:

$$u_{i,\text{env}} = \begin{cases} u_{i,\text{rule}}, & \text{if SGSM intervenes} \\ u_{i,\text{MARL}}, & \text{otherwise} \end{cases} \quad (15)$$

where $u_{\text{env}} = \{u_{i,\text{env}}\}_{i \in \mathcal{V}}$. The transition dynamics depend only on the executed joint action:

$$P(s' | s, u_{\text{env}}) \quad (16)$$

3.2 MA-PPO for CAVs

To solve the above POMDP, we adopt MA-PPO, an extension of the single-agent PPO paradigm, to cooperative or mixed cooperative-competitive multiagent settings. Below, we first introduce the basic PPO objective and then describe the MARL adaptation and parameter-sharing scheme.

In single-agent PPO, the policy $\pi_\theta(u|s)$ is updated by maximizing a clipped surrogate objective (Schulman et al., 2017):

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t \left[\min \left(r_t(\theta) \hat{A}_t, \text{CLIP} \left(r_t(\theta), 1 - \varepsilon, 1 + \varepsilon \right) \hat{A}_t \right) \right] \quad (17)$$

where $r_t(\theta) = \frac{\pi_\theta(u_t|s_t)}{\pi_{\theta_{\text{old}}}(u_t|s_t)}$ is the probability ratio between the new and old policies, $\hat{A}_t$ is an estimator of the advantage function, and ε is a hyperparameter. This clipping mechanism prevents excessively large policy updates that often destabilize learning.

We adopt a decentralized training and execution paradigm

based on MA-PPO with parameter sharing. All CAVs share a common policy and value parameterization, while each agent acts solely on its own local observation. This design preserves decentralized decision making for real-time traffic control and improves sample efficiency via parameter sharing across homogeneous agents.

Concretely, we set $\theta_1 = \theta_2 = \dots = \theta_N = \theta$ and define a shared policy:

$$\pi_{\theta}(u_i | o_i) \text{ for all } i \in \{1, \dots, N\} \quad (18)$$

The value baseline is also local and shared, so advantage estimates are computed per agent from trajectories under partial observability.

We use generalized advantage estimation (GAE) with local TD residuals:

$$\hat{A}_i(t) = \sum_{k=0}^{K-1} (\gamma \lambda)^k \delta_i(t+k) \quad (19)$$

$$\delta_i(\tau) = r_i(\tau) + \gamma V_{\theta}(o_i(\tau+1)) - V_{\theta}(o_i(\tau)) \quad (20)$$

where $\lambda \in [0,1]$ controls the bias-variance trade-off of GAE (Schulman et al., 2015). Policy updates follow PPO's clipped surrogate objective, with each agent independently optimizing its own policy based on its local observations while sharing common parameters across agents.

This MA-PPO with a parameter sharing setup scales well with the number of agents while retaining stable on-policy updates and decentralized execution. Empirically, we find it sufficiently expressive for coordination in our traffic scenarios.

A crucial aspect of our MA-PPO design is the explicit inclusion of surrounding vehicles' actions in the observation space. Specifically, each agent i observes its own local kinematics—speed, acceleration, and distance to the intersection—together with the intersection signals and the most recent control actions of its neighboring vehicles (Fig. 3). Given a batch of experience tuples $\{o_i(t), u_i(t), r_i(t), o_i(t+1)\}$ collected from parallel episodes, the algorithm performs multiple epochs of gradient ascent on the total objective:

$$\max_{\theta} \sum_{i=1}^N E_i [L_i^{\text{CLIP}}(\theta)] \quad (21)$$

where $L_i^{\text{CLIP}}(\theta)$ is the analog of Eq. (17) for agent i , typically summed or averaged across the entire batch.

During learning, we update the MARL policy and value functions only from time steps where $u_i^{\text{env}} = u_i^{\text{MARL}}$ for the corresponding agent; SGSM-controlled steps ($u_i^{\text{env}} = u_i^{\text{rule}}$) are excluded from MARL updates.

During training, each CAV agent applies its learned policy to determine high-level actions (lane changes or speed adjustments). HDVs follow the IDM and LC2013 models, contributing to the environment's dynamics. While standard and widely used, these models may not fully capture the heterogeneity of human driving, which may cause mismatches relative to real traffic. Thus, in Section 5.4, we will evaluate the impact of HDV model heterogeneity.

At every step, we compute individual rewards $r_i(t)$ and update the objective. By adopting parameter sharing, we handle a potentially large number of CAVs without exploding the dimensionality of policy parameters. This approach supports complex intersection layouts, high traffic density, and forced lane changes that demand cooperative interactions.

In summary, MA-PPO provides a robust framework for learning decentralized policies that exploit local information and receive stable advantage estimates from a joint critic.

3.3 DNN settings

We implement both the policy and the value networks as feedforward neural networks with a shared architecture for all agents, following the parameter-sharing scheme commonly used in MARL. Specifically, each network takes as input a real-valued observation vector and processes it through two fully connected layers with ReLU activation and layer normalization, as illustrated in Fig. 5.

The policy network consists of two fully connected layers. The first linear layer maps the input observation of size `obs_dim` to a 256-dimensional hidden representation, which is subsequently normalized by a layer normalization (LayerNorm) operator and activated by ReLU. The second linear layer maps the 256-dimensional output to a 128-dimensional feature vector, which is likewise normalized and passed through ReLU.

Following these hidden layers, the policy network branches into two heads. The continuous acceleration head outputs $\mu \in R$ via a linear projection. A trainable parameter $\log \sigma$ serves as the global log-standard deviation for the acceleration distribution. Thus, the network represents a Gaussian $\mathcal{N}(\mu, \sigma^2)$ for continuous longitudinal control. The discrete lane-change head outputs unnormalized logits in R^3 , corresponding to three discrete lane-change actions $\{-1, 0, +1\}$. A softmax operation is later applied to these logits to produce the probability of each lateral action. Hence, the policy network jointly learns a hybrid action space comprising continuous (acceleration) and discrete (lane change) components.

Under our setup, we estimate a local value function $V(o_i)$ for each agent from its own observation o_i . All agents share the same value-network parameters. The value network mirrors the policy backbone. Specifically, the value network also contains two fully connected layers of dimensions 256 and 128, each followed by LayerNorm and ReLU.

The final value head is a linear projection from the 128-dimensional output to a scalar, which serves as the estimate of $V(o_i)$. We apply a squeeze operation to reduce the last dimension, thus producing a single floating-point value per input observation. All linear layers are initialized with the Xavier (Glorot) scheme, and LayerNorm biases are set to zero.

3.4 Pretraining and fine-tuning via transfer learning

To reduce unsafe and inefficient exploration at the beginning of training and to speed up convergence, we employ a transfer-learning protocol that initializes the policy and value networks from a simpler source task and then fine-tunes them on the target task. Pretraining is used only to provide an initialization. All learning on the target task is performed with on-policy PPO using

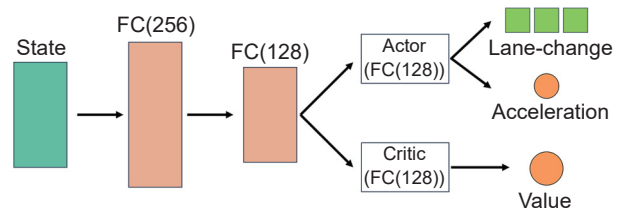


Fig. 5 Network architecture for the shared policy and value functions. The input is an observation vector processed by two shared fully connected layers. The policy head outputs logits for three lane-change actions and a Gaussian mean for longitudinal acceleration; the value head outputs a scalar value estimate.

trajectories collected online in the target environment. The overall pipeline is illustrated in Fig. 6.

We construct a synthetically generated low-density intersection scenario (arrival rate ≈ 1600 veh/h) in SUMO, where only a single CAV agent is controlled and all other vehicles behave as HDVs. The observation and action spaces are kept shape-consistent with the target task: neighbor-related features are retained in the observation vector, and if a specific neighbor or its action is unavailable (e.g., HDVs), the corresponding entries are zero-padded as in Section 3.1.1. We reuse the same reward terms so that the optimization objective matches the target environment. Pretraining produces policy and value parameters θ_{src}, ϕ_{src} for initialization.

The pretrained networks are then warm-started on the real-world mixed-traffic intersection dataset with multiple simultaneously controlled CAV agents. We adopt the same loss and setup described in Section 3 while enabling parameter sharing across agents. All layers are fine-tuned end-to-end; the observation now includes actual neighboring CAV actions when available (HDV actions remain zero-padded). This preserves the input-output interface across domains and allows the multiagent policy to quickly adapt to denser traffic and MLCs.

4 Simulation-guided supervisory module

Although MARL can effectively manage both longitudinal and lateral trajectory planning for multiple vehicles, its training process can become much more complex and slower to converge under high traffic volumes or high penetration rates of autonomous vehicles. This slowdown arises because a large number of interacting agents creates a highly dynamic environment, which in turn leads to more intricate reward landscapes and increased nonstationarity. As a result, learning stability and convergence speed can be significantly affected, potentially leading to suboptimal performance or policy oscillations.

To contain these risks, we design a Simulation-guided supervisory module, SGSM. At every decision step, SGSM tests whether a standard human-driver model (HDM) can handle the scene (such as IDM(Treiber et al., 2000) combined with LC2013(Erdmann, 2015)). If so, the system keeps the HDM motion; otherwise, it calls the MARL policy.

4.1 Trajectory prediction unit

The first layer of the SGSM is the trajectory prediction unit. This unit looks ahead over a fixed horizon and evaluates two key conditions (Fig. 7). The first condition assumes that all vehicles

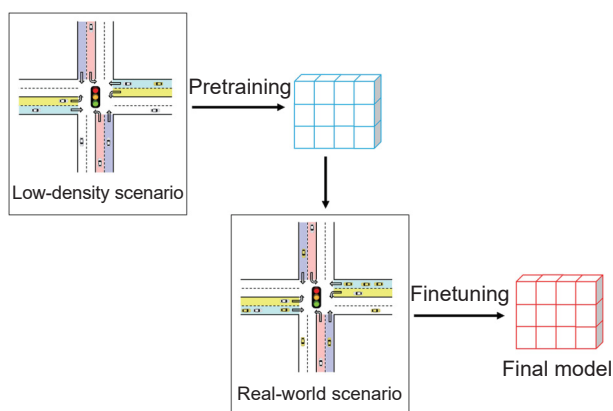


Fig. 6 Transfer-learning pipeline: A simple SUMO scenario (left) is used to pretrain the model, which is then transferred to a complex, denser multiagent scenario (right) and fine-tuned to obtain the final model.

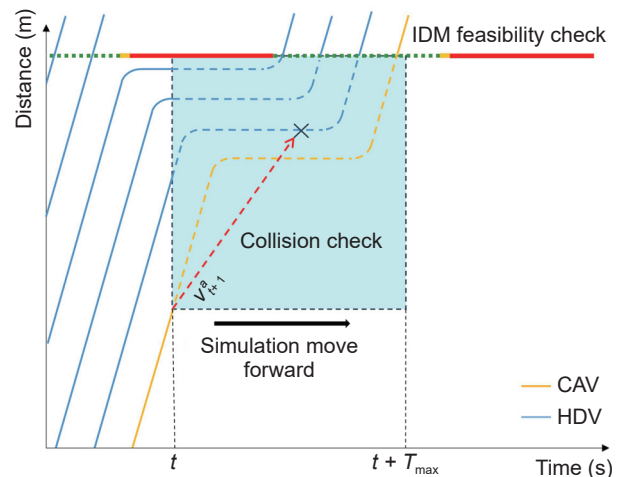


Fig. 7 Workflow of the trajectory prediction module. In the first pass, all vehicles (including the CAV) apply IDM + LC2013. The second pass replays the preceding vehicles' IDM trajectories but checks the collision risk if the CAV maintains its MARL next-step speed.

follow the IDM and LC2013 models. If the controlled CAV can pass through the intersection during the green light without experiencing a stop-and-go situation, it is considered that the HDM is sufficient to control the vehicle. In the second condition, the unit replays the trajectories of all the leaders while allowing the CAV to follow the speed proposed by the MARL policy. If the gap between the CAV and the leading vehicle becomes negative, indicating a potential collision, the step is marked as risky.

The pseudocode provided in Algorithm 1 outlines the trajectory prediction and collision check process. Initially, the trajectory prediction unit sets the feasibility of HDM control to true and the collision risk to false. The algorithm then proceeds in two passes. In the first pass, all vehicles follow the IDM and LC2013 models over the prediction horizon. The unit continuously propagates the positions of the leaders and the CAV. The trajectory of each leader is stored for future reference. If the current CAV is unable to pass through the intersection within the green light duration, HDM control is considered insufficient. The system also checks for any stop-and-go situations. If the current CAV experiences a stop and later resumes motion, HDM feasibility is set to false. In the second pass, the leaders' trajectories are replayed, and the CAV maintains the speed determined by the MARL policy. The algorithm computes the gap between the CAV and the leaders. If at any point this gap becomes zero or negative, indicating a potential collision, the collision risk flag is set to true, and the process is terminated early. The output of the trajectory prediction unit consists of two binary flags: HDM feasibility and collision risk. These flags help the system decide whether to continue using HDM control or switch to the MARL policy for further decision-making.

4.2 Safety regulator

The second SGSM layer acts as a rule-based safety regulator that integrates offline flags with lane discipline rules, as shown in Algorithm 2. First, for lane enforcement, if a CAV approaches the tail of the correct-lane queue while still occupying a nontarget lane, the regulator triggers an immediate MLC (Jiang et al., 2023). In our setting, lane changes are prohibited within a solid lane segment and are permitted only in dashed segments. Specifically, when the agent is within a solid lane segment, the regulator overrides the trigger distance by setting DistanceToCorrectLaneTail to w , which immediately triggers MLC upstream of the solid line.

Algorithm 1 Trajectory prediction and collision check

Input: Vehicle ID veh_id , horizon T_{max} , step Δt , IDM params $\{s_0, T, v_0, a_{max}, b_{comf}, \delta\}$, signal state σ_t , distance to the stop line d_t

Output: HDM_feasible (bool), collision_risk (bool)

```

1: HDM_feasible  $\leftarrow$  True; collision_risk  $\leftarrow$  False
2:  $t \leftarrow 0$ ; stopped_then_moved  $\leftarrow$  False
3: Cache  $\mathcal{T}_{leaders} \leftarrow \emptyset$ 
4: while  $t < T_{max}$  do
5:   for each leader vehicle in the same lane do
6:     Propagate leader's trajectory by IDM + LC2013 with step  $\Delta t$ 
7:     if  $\sigma_t \in \{\text{RED}, \text{YELLOW}\}$  and  $d_t \leq d_{th}$  then
8:       set leader  $v \leftarrow 0$ 
9:     end if
10:    Append current leader state to  $\mathcal{T}_{leaders}$ 
11:   end for
12:   Update ego speed  $v_{ego}(t)$ 
13:   Update flag stopped_then_moved if  $v < v_{th}$  for  $\geq 2$  s and later  $> v_{th}$ 
14:   if PassWithinGreen(ego) = False or stopped_then_moved = True then
15:     HDM_feasible  $\leftarrow$  False
16:   end if
17:    $t \leftarrow t + \Delta t$ 
18: end while
19:  $t \leftarrow 0$ 
20: while  $t < T_{max}$  do
21:   for each leader vehicle in the same lane do
22:     Retrieve leader state at time  $t$  from  $\mathcal{T}_{leaders}$ 
23:   end for
24:   Set ego speed  $v_{ego}(t) \leftarrow v_{MARL}(t)$ 
25:   Compute net gap  $g(t)$ 
26:   if  $g(t) \leq 0$  then
27:     collision_risk  $\leftarrow$  True
28:   break
29:   end if
30:    $t \leftarrow t + \Delta t$ 
31: end while
32: return HDM_feasible, collision_risk

```

Second, for control selection, the vehicle continues with IDM + LC2013 for the current step whenever that model remains feasible; otherwise, it executes the MARL-prescribed action. Fig. 8 visualizes the enforced MLC rule.

By combining multistep trajectory checks with a lane-correction strategy, it reduces scenario diversity so that the MARL policy can focus on difficult cases.

5 Experiments and results

5.1 Experimental setup

We utilize a real-world dataset, hereafter referred to as the Langfang (LF) dataset, collected from a four-leg intersection located in Langfang, Hebei Province, China. As shown in Fig. 1, each arm in the intersection has two lanes, consisting of a dedicated left-turn lane and a through lane. The approach length

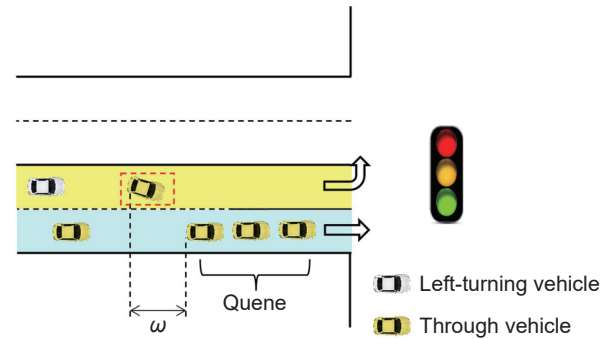


Fig. 8 Forced lane changes when a straight-moving vehicle in a nontarget lane approaches the queue tail within ω .

Algorithm 2 Rule-based safety regulator with MARL integration

Input: MARL actions $\mathbf{A}$, CAV set $\mathcal{V}$, enforcement distance ω , per-vehicle flags HDM_feasible $[\cdot]$, current lane index $l_i(t)$, target lane index l_i^* , lane-change command toward the target lane c_i^{force}

Output: Modified actions $\mathbf{A}$

```

1: for all  $veh\_id \in \mathcal{V}$  do
2:   if  $l_{veh\_id}(t) \neq l_{veh\_id}^*$  then
3:      $d_{tail} \leftarrow \text{DistanceToCorrectLaneTail}(veh\_id)$ 
4:      $d_{stop} \leftarrow \text{DistanceToStopLine}(veh\_id)$ 
5:     if  $d_{stop} < \ell_{solid}$  then
6:        $d_{tail} \leftarrow \omega$ 
7:     end if
8:     if  $d_{tail} \leq \omega$  then
9:        $\mathbf{A}[veh\_id] \leftarrow c_{veh\_id}^{force}$ 
10:    end if
11:  end if
12:  if HDM_feasible[ $veh\_id$ ] = True then
13:     $\mathbf{A}[veh\_id] \leftarrow \text{IDM\_LC2013\_Action}(veh\_id)$ 
14:  end if
15: end for

```

for each leg, measured from the entry point to the stop line, is set to 300 m. In each approach, the final 30 m upstream of the stop line is separated by a solid lane marking between the two lanes, and lane changes are prohibited. To provide a clearer understanding of the experimental setting, a satellite image of the actual intersection is shown in Fig. 9.

The signal control includes four signal phases: north-south through, north-south left-turn, east-west through, and east-west left-turn. Signal timings are calculated using the Webster method, and the specific timing parameters are listed in Table 1. These timings dictate the length of the green phase for each direction and the duration of the yellow phase, which is set to 2 s for all phases. Table 2 lists key vehicle parameters and default simulation settings.

The traffic flow rate is approximately 2364 vehicles per hour. Based on the predefined penetration rate, a subset of vehicles in the traffic stream is randomly selected to serve as CAVs. The traffic participants consist of both CAVs and HDVs.

The behaviors of HDVs are simulated using the IDM and LC2013 lane-changing model, while all CAVs are controlled by the learned SS-MA-PPO policy within the control area. To preserve realism, we do not strictly prohibit every potential HDV violation. In our intersection testbed, each leg is 300 m long, and we define a control zone of length R upstream of the stop line. In



Fig. 9 Satellite image of the real-world intersection in Langfang, Hebei Province, China.

Table 1 Signal timing parameters for traffic control

Signal phase	Time (s)
North-south through	49
North-south left-turn	30
East-west through	46
East-west left-turn	65
Yellow time (all phases)	2

Table 2 Vehicle behavior parameters used in simulation

Parameter	Value
Vehicle length (m)	5
Vehicle width (m)	2
Minimum gap (m)	2
Maximum speed (m/s)	15
Maximum acceleration (m/s^2)	1.5
Maximum deceleration (m/s^2)	1.5
Car-following model	IDM
Lane-changing model	LC2013
Minimum time headway (s)	2
Driver imperfection coefficient (SUMO sigma)	0.2

the default setting, we set $R = 300$ m, corresponding to full-range control over each approach.

System performance is assessed by four macroscopic indicators, defined as follows:

- 1) Waiting time (s): The total time a vehicle travels slower than 0.1 m/s between entering the approach and crossing the stop line;
- 2) Delay (s): The additional travel time of a vehicle with respect to its free-flow travel time over the same distance;
- 3) Fuel (ml): The fuel volume consumed per vehicle, computed with the polynomial energy consumption model in (Yu et al., 2021).
- 4) Stop-and-go count: stop recorded when the speed stays below v_{stop} for more than t_{stop} after at least t_{move} of motion. In the experiments, we set $v_{\text{stop}} = 0.1$ m/s and $t_{\text{stop}} = t_{\text{move}} = 2$ s.

Fig. 1 illustrates the overall experimental environment. The algorithm uses the hyperparameters summarized in Table 3. All experiments are conducted on a server equipped with an Intel® Xeon Platinum 8336C CPU and an NVIDIA GeForce RTX 4090 GPU.

5.2 Algorithm comparison

In this subsection, we compare the performance of the proposed

Table 3 Hyperparameter settings for training

Parameter	Value
Learning rate	3.00×10^{-4}
Episodes	150
Batch size	256
Clip range, ϵ	0.20
Value function coefficient	1.50
Entropy coefficient	0.01
Discount, factor, γ	0.90
GAE parameter, λ	0.95

SS-MA-PPO with several representative baselines, including SS-MADDPG, SS-MAA2C, and a purely rule-based method (IDM + LC2013).

Fig. 10 shows that in the absence of CAVs, traffic oscillations caused by red lights affect nearly all vehicles approaching the intersection. Due to excessive queues in the target lane, some vehicles are forced to stop frequently in the middle of the road in search of lane-changing opportunities, introducing secondary disturbances, for example, vehicle #17. Table 4 summarizes the evaluation results of all MARL algorithms after training. The proposed SS-MA-PPO consistently outperforms all baseline methods under various CAV penetration rates.

Fig. 11 illustrates vehicle trajectories in a selected lane under different CAV penetration rates, controlled by our proposed method. Red trajectories represent CAVs, and blue trajectories represent HDVs. Discontinuous trajectories represent lane-changing maneuvers. Fig. 12 provides a visual comparison of performance trends as the penetration rate increases.

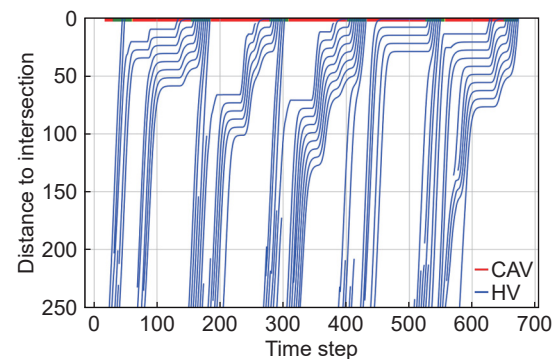


Fig. 10 Performance under IDM (0% CAV Penetration).

Table 4 Performance comparison under different CAV penetration

Penetration	Method	Waiting time (s)	Delay (s)	Fuel (ml)	Stop-and-go count
0%	IDM + LC2013	60.90	161.79	103.80	1.94
	SS-MAA2C	39.83	143.85	95.48	1.21
20%	SS-MADDPG	35.77	147.54	92.28	1.12
	SS-MA-PPO	33.65	138.83	89.33	1.02
40%	SS-MAA2C	23.23	143.42	90.86	1.03
	SS-MADDPG	23.05	142.43	90.66	1.03
	SS-MA-PPO	20.71	128.89	83.93	0.87
60%	SS-MAA2C	17.64	132.52	86.58	0.89
	SS-MADDPG	17.90	141.68	90.17	1.01
	SS-MA-PPO	16.95	121.83	81.73	0.83
100%	SS-MAA2C	9.97	122.15	82.36	0.86
	SS-MADDPG	11.22	134.40	87.37	0.90
	SS-MA-PPO	9.98	109.77	75.91	0.81

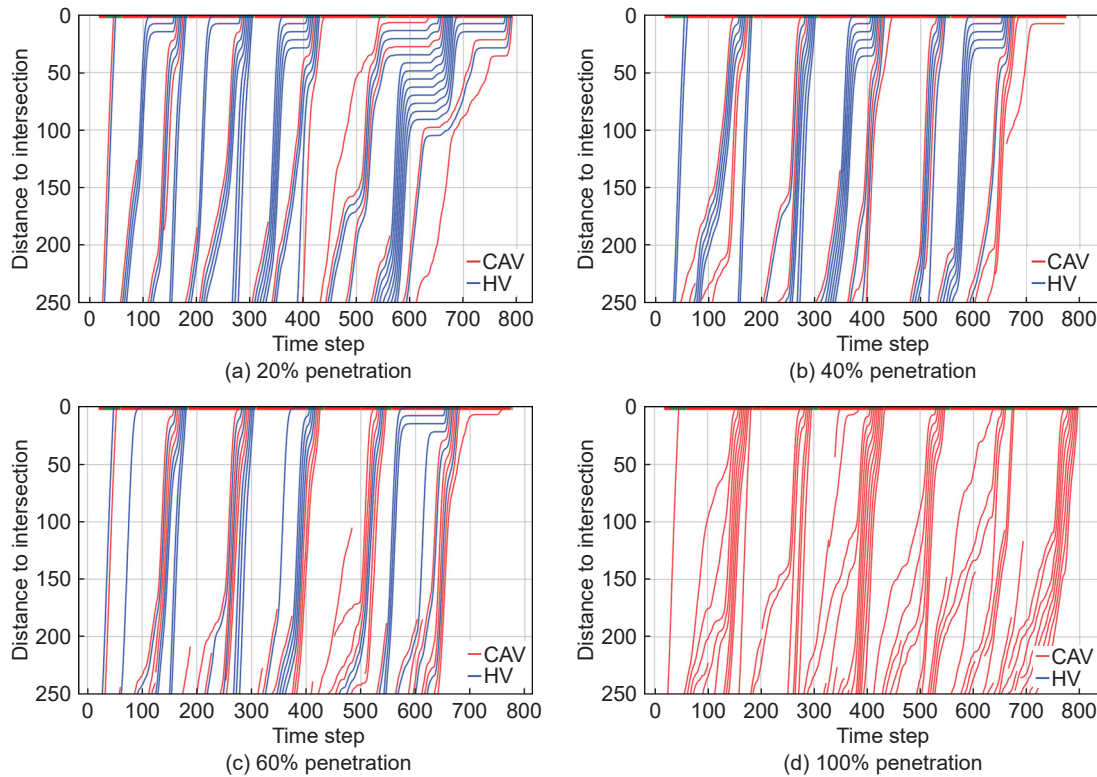


Fig. 11 Vehicle trajectories under different CAV penetration rates: (a) 20% penetration, (b) 40% penetration, (c) 60% penetration, and (d) 100% penetration.

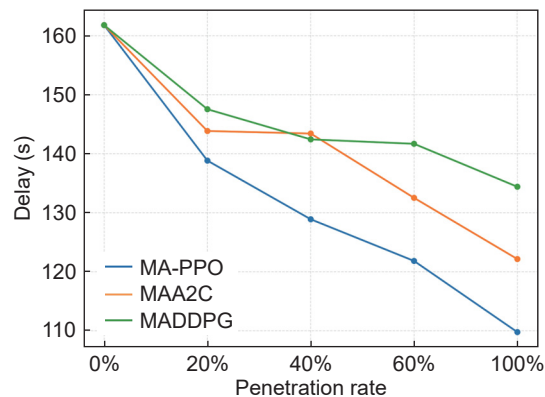


Fig. 12 Delay comparison for SS-MA-PPO, SS-MAA2C, and SS-MADDPG under different CAV penetration rates. The horizontal axis shows the penetration rate (0%, 20%, 40%, 60%, 100%), and the vertical axis indicates the average delay.

As illustrated in Figs. 11a and 11b, the introduction of a small proportion of CAVs significantly suppresses these oscillations. This is attributed to the proactive behavior of CAVs, which adjust their speed before reaching the intersection and perform timely lane changes earlier upstream. These actions not only ensure that CAVs pass through the intersection during green lights but also help following vehicles do so more smoothly. However, minor oscillations induced by HDVs still persist.

In high-penetration scenarios, as shown in Figs. 11c and 11d, oscillations caused by stop-and-go behavior and frequent lane changing are almost eliminated. This demonstrates the superior coordination and efficiency achieved by SS-MA-PPO in mixed traffic environments.

5.3 Ablation studies

To clarify the contribution of each algorithmic component, we perform three targeted ablation studies. Unless otherwise noted,

all experiments are evaluated under the same traffic demand, roadway geometry, and random seeds as the main results.

5.3.1 Impact of lane-changing decisions

To isolate the effects of longitudinal and lateral trajectory planning, we compare the following three approaches under various CAV penetration rates:

- 1) Accel only: The agent executes acceleration actions from SS-MA-PPO, while lane-changing actions follow LC2013;
- 2) Lane only: The agent executes lane-changing actions from SS-MA-PPO, while acceleration actions follow IDM;
- 3) Full control: The agent executes both acceleration and lane-changing actions from SS-MA-PPO.

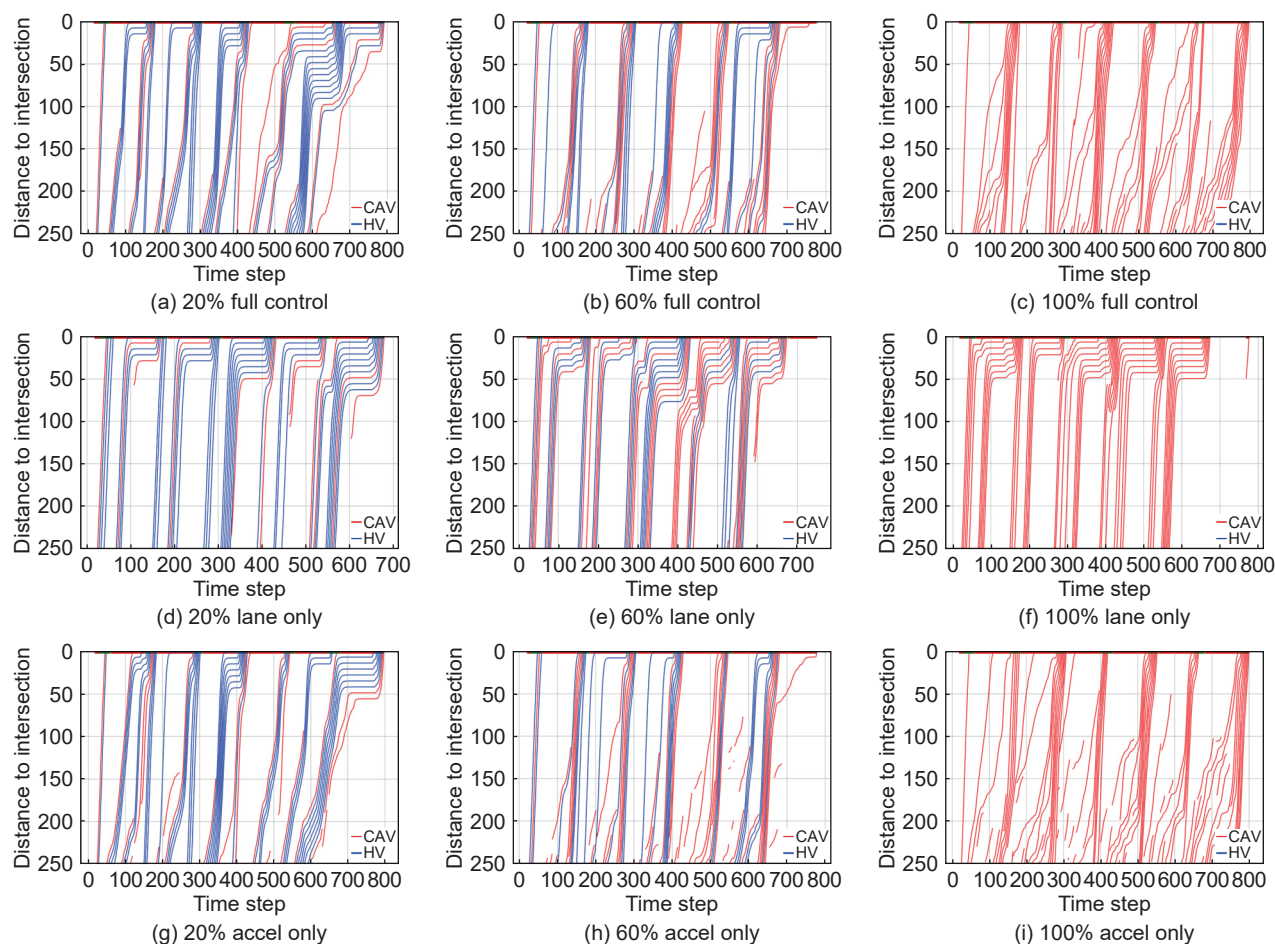
The experimental results are shown in Table 5. Both “Accel only” and “Lane only” outperformed the baseline method (no control), indicating that both longitudinal and lateral control improve traffic flow. However, by controlling both longitudinal acceleration and lateral lane changes, “Full Control” achieved the optimal performance, showing that coordinating both dimensions leads to the best overall performance.

The ratio of ρ_{MARL} and ρ_{rule} is defined as the fraction of environmental steps whose executed control originates from MARL or the rule-based method, averaged over rollouts. The result shows that higher CAV penetration is associated with a lower proportion of MARL-executed actions. Since the MARL controller is activated only in difficult scenarios where vehicles cannot directly pass the stop bar during green phases, a decrease in the ratio of ρ_{MARL} indicates fewer vehicle conflicts. In other words, this result demonstrates that controlling CAVs in both longitudinal and lateral dimensions enhances cooperation, thereby reducing conflicts and ultimately improving overall traffic efficiency.

Fig. 13 further illustrates vehicle trajectories under different control strategies. Compared to the full control setting, the “Accel Only” strategy results in more frequent and fragmented lane

Table 5 Ablation study on control strategies under different penetration rates

Penetration	Control strategy	Waiting time (s)	Delay (s)	Fuel (ml)	Stop-and-go count	ρ_{MARL} (%)	ρ_{rule} (%)
20%	Accel only	38.15	157.35	96.91	1.29	83.39	16.61
	Lane only	61.01	159.63	103.22	1.92	82.90	17.10
	Full control	33.65	139.00	89.33	1.02	79.75	20.25
60%	Accel only	20.27	135.37	87.23	1.23	80.61	19.39
	Lane only	59.80	159.39	103.21	1.92	82.01	17.99
	Full control	16.95	127.03	84.16	0.83	75.39	24.61
100%	Accel only	11.80	119.82	81.79	1.13	74.20	25.80
	Lane only	58.92	159.36	103.09	1.91	79.20	20.80
	Full control	9.43	116.75	78.70	0.81	73.67	26.33

**Fig. 13** Vehicle trajectories under different penetration rates and control strategies. (a–c) Full control at 20%, 60%, and 100%; (d–f) lane only at 20%, 60%, and 100%; (g–i) accel only at 20%, 60%, and 100%.

changes, indicating reduced driving smoothness. Moreover, Fig. 13 shows that since the “Lane Only” strategy does not plan longitudinal maneuvers, vehicles queue before the stop line, which deteriorates traffic efficiency.

5.3.2 Contribution of the SGSM

The SGSM supervises the RL policy and takes control when the learned policy is likely to choose an unsafe or low-value action. Its effect is evaluated at a 40% penetration rate.

Fig. 14 illustrates the comparison of training performance with and without SGSM. As seen in the plot, SS-MA-PPO converges after approximately 75 episodes, while MA-PPO continues learning until approximately episode 150, with significantly worse performance.

This stark difference occurs because as more agents are introduced in the environment, the difficulty of training and the

convergence speed for MA-PPO increase dramatically. The increase in agent interactions leads to more complex dynamics, making it difficult for MA-PPO to efficiently learn and converge in a reasonable number of episodes.

In contrast, SS-MA-PPO benefits from SGSM, which delegates most scenarios that do not require RL intervention to the IDM + LC2013 model, particularly during the early exploration phase. By preventing premature terminations caused by unnecessary risky behaviors, SGSM facilitates more stable and efficient learning.

5.3.3 Cooperative vs. independent learning

We evaluate the effectiveness of sharing neighboring actions. We compare our cooperative variant (SS-MA-PPO) with an independent learner (SS-IMA-PPO), which omits neighbor actions and trains each agent separately.

The evaluation results under different penetration rates are

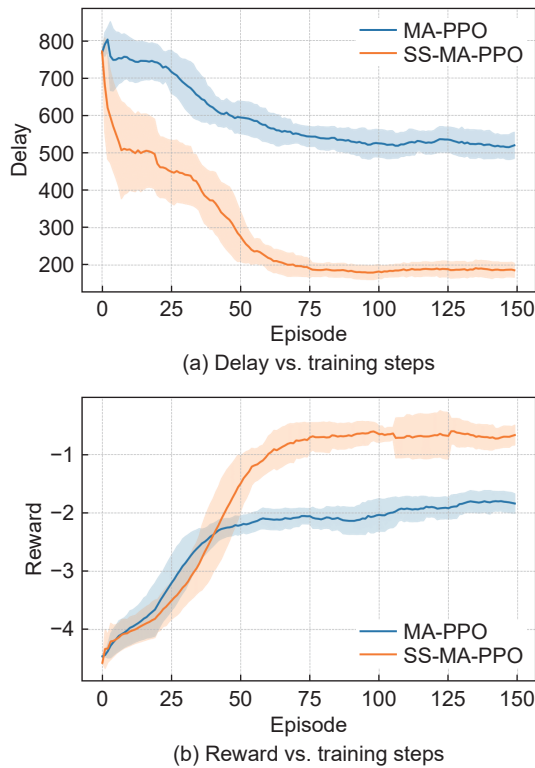


Fig. 14 Training performance with and without SGSM under a 40% penetration rate: (a) delay vs. training steps; (b) reward vs. training steps.

summarized in Table 6. Compared to our method, SS-IMA-PPO converges to worse policies during training, leading to significantly degraded performance in all metrics, particularly in high CAV penetration scenarios. This highlights the advantages of incorporating surrounding vehicle actions and cooperative learning in improving traffic stability and efficiency.

5.3.4 Transfer learning

As shown in Fig. 15, the transfer learning strategy outperforms training from scratch. With transfer learning, the reward starts at a higher level and reaches the final converged state after approximately 30 episodes. Without transfer learning, the reward starts lower and the vehicle delay is higher, taking approximately 100 episodes to finally converge. These results demonstrate that transfer learning significantly accelerates convergence.

5.4 Sensitivity analysis

5.4.1 Control-zone length

First, we perform a sensitivity analysis on the length of the control zone R , which plays an important role in the trajectory planning of CAVs (Liu et al., 2022; Ma et al., 2021). Outside the control zone, CAVs follow the simulator's baseline longitudinal and lane-changing rules, whereas once within the control zone, the SS-MA-

Table 6 Independent vs. cooperative learning under different penetration rates

Penetration	Method	Waiting time (s)	Delay (s)	Fuel (ml)	Stop-and-go count
20%	SS-IMA-PPO	35.82	158.80	97.45	1.20
	SS-MA-PPO	33.65	139.00	89.33	1.02
60%	SS-IMA-PPO	23.20	149.54	93.60	1.13
	SS-MA-PPO	16.95	127.03	84.16	0.83
100%	SS-IMA-PPO	19.03	146.59	93.55	1.10
	SS-MA-PPO	9.43	116.75	78.70	0.81

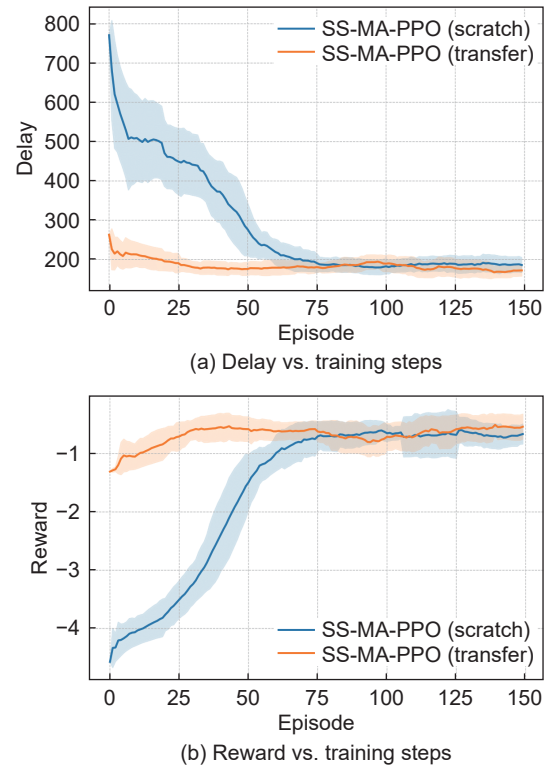


Fig. 15 Comparison of training curves with and without transfer learning: (a) delay vs. training steps; (b) reward vs. training steps. The transfer learning strategy provides better initial performance, faster convergence, and improved final results compared to training from scratch.

PPO policy with SGSM is applied. In this study, we fix all other parameters, set the CAV penetration rate to 40%, and test the length of the control zone $R \in \{100 \text{ m}, 200 \text{ m}, 300 \text{ m}\}$. The corresponding results are reported in Table 7.

The results reveal three main findings. First, short control zones (100 m) substantially degrade performance, with higher waiting time, delay, fuel consumption, and stop-and-go counts. This indicates that activating control too late leaves insufficient distance for lane positioning and approach smoothing. Second, medium and long control ranges (200–300 m) yield similar delay, fuel consumption, and stop-and-go counts, suggesting that once a moderate approach length is available, the marginal benefits of extending the control zone largely plateau. Third, when increasing the control zone from 200 to 300 m, most metrics remain close, but the waiting time improves markedly at 300 m. A reasonable explanation is that the longer control zone enables more effective upstream metering and platoon shaping so that arrivals align better with green phases and stop-line queuing is reduced, without materially changing the overall travel time or energy consumption. Under $R = 200 \text{ m}$, the first 100 m upstream remain uncontrolled. In this relatively uncongested segment, vehicles accelerate according to the baseline rules, reach the intersection earlier relative to the red phase, and consequently experience increased waiting time before the stop bar.

Table 7 Sensitivity to the control zone length R

R (m)	Avg. waiting (s)	Avg. delay (s)	Avg. fuel (ml)	Avg. stop-and-go
100	28.85	157.27	96.21	1.01
200	24.80	129.67	83.94	0.87
300	20.71	128.89	83.93	0.87

5.4.2 HDV model heterogeneity

In the real world, the heterogeneity of human driving behaviors may degrade the performance of CAV trajectory planning strategies (Shang et al., 2024). In the preceding experiments, all HDVs follow the IDM car-following model combined with the LC2013 lane-changing model. To further examine robustness under heterogeneous HDV behaviors, we construct additional scenarios in SUMO in which 20% or 40% of the original IDM + LC2013 HDVs are randomly reassigned to follow the Krauss car-following model (Krauss, 1998) together with the SL2015 lane-changing model. In other words, we implement simulations that include two types of HDV behavior models differing in both longitudinal acceleration and lane-changing decisions while applying the same SS-MA-PPO policies trained in the baseline environment.

Figs. 16 and 17 present the distributions of waiting time and travel time under different levels of HDV heterogeneity. As shown in Figs. 16b and 17b, the median waiting time and travel time of HDVs increase slightly as heterogeneity grows. However, the corresponding metrics for CAVs remain nearly unchanged across

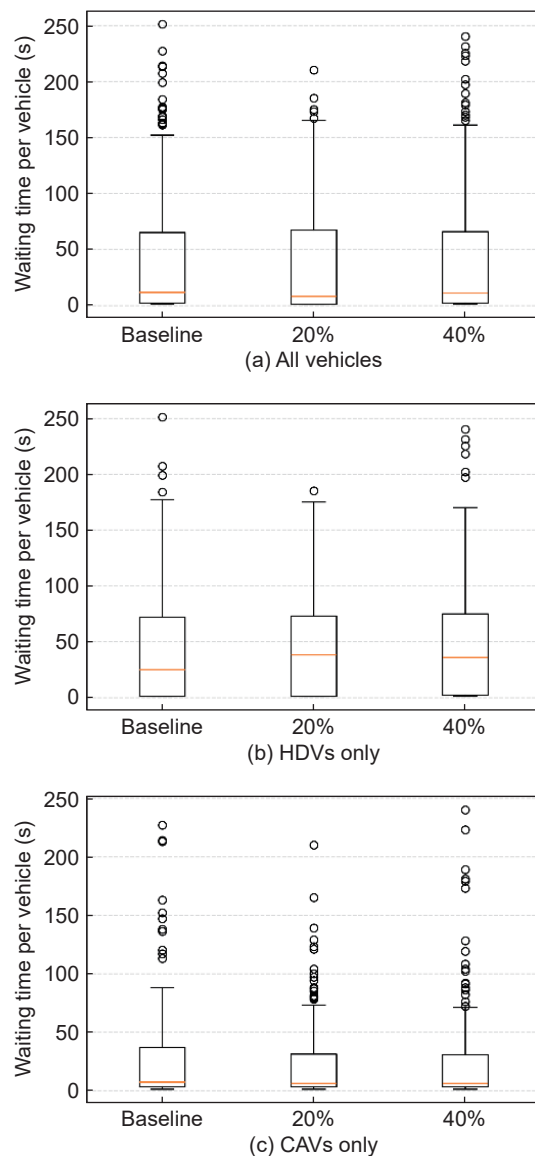


Fig. 16 Waiting time under different levels of HDV model heterogeneity: (a) all vehicles, (b) HDVs only, and (c) CAVs only.

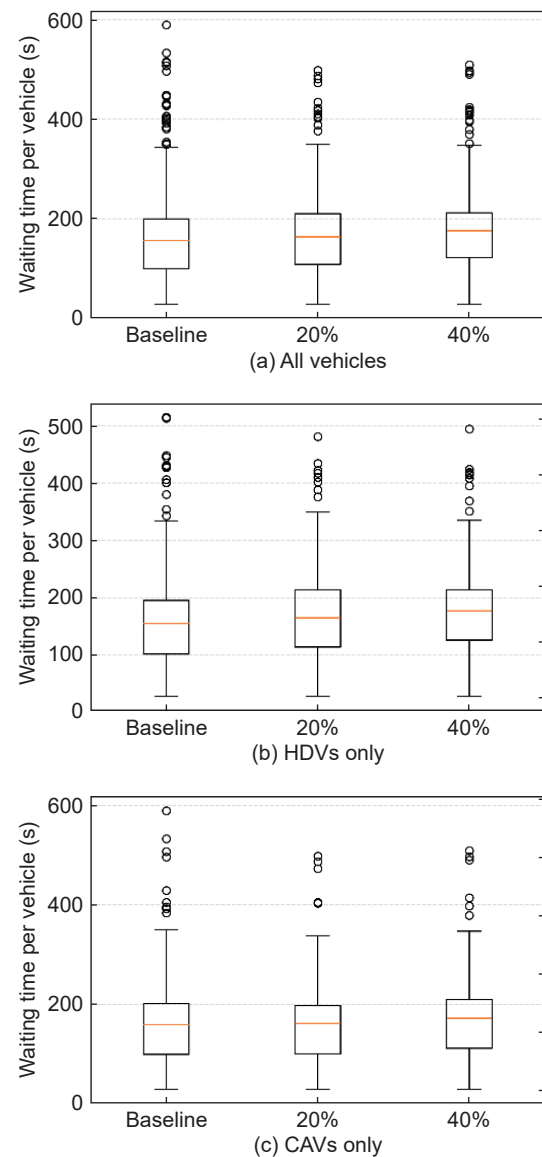


Fig. 17 Travel time under different levels of HDV model heterogeneity: (a) all vehicles, (b) HDVs only, and (c) CAVs only.

the three scenarios (Figs. 16c and 17c), indicating that the performance of the SS-MA-PPO controller is maintained under varying levels of heterogeneity. Therefore, as illustrated, the overall performance of all vehicles remains largely consistent. These findings demonstrate that our method is robust to HDV heterogeneity and that the SGSM-based safety check is not overly sensitive to moderate variations in HDV behavior models.

6 Conclusions

This work presents SS-MA-PPO, a simulation-supervised multiagent proximal policy optimization framework for CAVs at signalized intersections. The method explicitly tackles MLCs by coupling longitudinal and lateral motion in a single decision, embeds SGSM that takes over during hazardous states, and exploits transfer learning to warm-start policies across penetration scenarios. Comprehensive experiments on a real-world dataset from Langfang demonstrate its effectiveness under mixed traffic.

Across penetration rates from 20% to 100%, SS-MA-PPO achieves the lowest delay, waiting time, fuel consumption, and stop-and-go counts, outperforming SS-MADDPG, SS-MAA2C,

and the rule-based IDM + LC2013 baseline. Ablation studies show that combining longitudinal and lateral trajectory planning reduces delay by up to approximately 20% compared with single-axis strategies. Incorporating neighboring actions further cuts delay and fuel, confirming the value of explicit cooperation. Transfer learning shortens convergence and lifts the final reward, while the SGSM avoids unsafe exploration and accelerates training. In summary, SS-MA-PPO provides a practical and scalable solution for cooperative CAV control and offers clear benefits for traffic efficiency and energy use in realistic mixed-traffic conditions.

This study has several limitations that warrant further investigation. First, the current study assumes fixed-time signal timing, while in mixed traffic, adaptive signal control is necessary. Therefore, future work will integrate signal control optimization into the SS-MA-PPO framework. Second, it is a promising topic to extend this method from isolated intersections to corridor- or network-scale scenarios, which could achieve coordination among multiple intersections. Third, a promising direction is to replace IDM + LC2013 with data-driven and hybrid human models, such as inverse-RL surrogates, to enhance realism and generalization. Finally, we note that naive extensions beyond on-policy PPO may incur an action-next-state mismatch. It may be beneficial to use SGSM-controlled steps primarily for critic fitting while masking their contributions to the actor gradients.

Author contributions

Xingwei Jiang: Conceptualization, Methodology, Writing – Original draft. **Meng Li:** Conceptualization, Writing – review & editing, supervision. **Qingquan Liu:** Conceptualization, methodology, Writing – review & editing.

Replication and data sharing

The source codes and replication package are available on ETS data at <https://doi.org/10.26599/ETSD.2025.9190072>. The source code is available at: <https://github.com/Xingwei-Jiang/SS-MA-PPO>.

Acknowledgements

This research is supported by the National Key R&D Program of China (2023YFB4301800), the National Natural Science Foundation of China (Nos. 52325209 and 52272420), the Tsinghua University–Mercedes Benz Joint Institute for Sustainable Mobility, and the Tsinghua–Toyota Joint Research Institute Interdisciplinary Program.

Declaration of competing interest

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

References

Ali, Y., Zheng, Z., Haque, M. M., Wang, M., 2019. A game theory-based approach for modelling mandatory lane-changing behaviour in a connected environment. *Transp Res Part C Emerg Technol*, **106**, 220–242.

Amato, C., 2024. An Initial Introduction to Cooperative Multi-agent Reinforcement Learning. <https://doi.org/10.48550/arXiv.2405.06161>

Cai, M., Xu, Q., Chen, C., Wang, J., Li, K., Wang, J., et al., 2022. Formation control with lane preference for connected and automated vehicles in multi-lane scenarios. *Transp Res Part C Emerg Technol*, **136**, 103513.

Chen, C., Cai, M., Wang, J., Li, K., Xu, Q., Wang, J., et al., 2022. Cooperation method of connected and automated vehicles at unsignalized

intersections: Lane changing and arrival scheduling. *IEEE Trans Veh Technol*, **71**, 11351–11366.

Chen, D., Hajidavalloo, M. R., Li, Z., Chen, K., Wang, Y., Jiang, L., et al., 2023. Deep multi-agent reinforcement learning for highway on-ramp merging in mixed traffic. *IEEE Trans Intell Transp Syst*, **24**, 11623–11638.

Chen, L., Dong, T., Li, X., Xu, X., 2025. Logistics engineering management in the platform supply chain: An overview from logistics service strategy selection perspective. *Engineering*, **47**, 236–249.

Du, Y., 2024. Cooperative Multi-Agent Learning in a Complex World: Challenges and Solutions. Proceedings of the AAAI Conference on Artificial Intelligence, 37(13), 15436–15436.

Duan, L., Wei, Y., Dong, S., Li, C., 2023. Centralized vehicle trajectory planning on general platoon sorting problem with multi-vehicle lane changing. *Transp Res Part C Emerg Technol*, **154**, 104273.

Erdmann, J., 2015. SUMO's Lane-changing model. In: Proceedings of the Simulation of Urban Mobility (SUMO) Conference 2014: Modeling Mobility with Open Data, 105–123.

Gao, Z., Jia, B., Xie, D., Wang, W., Wu, J., 2025. A discussion on the complexity and transit mechanisms of urban traffic systems. *Engineering*, **44**, 26–31.

Guo, Q., Ban, X. J., Aziz, H. M. A., 2021. Mixed traffic flow of human driven vehicles and automated vehicles on dynamic transportation networks. *Transp Res Part C Emerg Technol*, **128**, 103159.

Haydari, A., Yilmaz, Y., 2022. Deep reinforcement learning for intelligent transportation systems: A survey. *IEEE Trans Intell Transp Syst*, **23**, 11–32.

He, J., Qu, J., Zhang, J., He, Z., 2023a. The impact of a single discretionary lane change on surrounding traffic: An analytic investigation. *IEEE Trans Intell Transp Syst*, **24**, 554–563.

He, X., Yang, H., Hu, Z., Lv, C., 2023b. Robust lane change decision making for autonomous vehicles: An observation adversarial reinforcement learning approach. *IEEE Trans Intell Veh*, **8**, 184–193.

Jiang, H., Yao, Z., Jiang, Y., He, Z., 2023. Is all-direction turn lane a good choice for autonomous intersections? A study of method development and comparisons. *IEEE Trans Veh Technol*, **72**, 8510–8525.

Jokhio, S., Dürr, M., Bärghman, J., Baumann, M., 2024. Influence of surrounding traffic on lane change dynamics: Insights from a video-based laboratory study. *Transp Res Part F Traffic Psychol Behav*, **105**, 87–98.

Krauss, S., 1998. Microscopic Modeling of Traffic Flow: Investigation of Collision Free Vehicle Dynamics. <https://api.semanticscholar.org/CorpusID:111793371>

Laval, J. A., Daganzo, C. F., 2006. Lane-changing in traffic streams. *Transp Res Part B Methodol*, **40**, 251–264.

Lee, J., Choi, S. B., 2023. Integrated control of steering and braking for path tracking using multi-point linearized MPC. *IEEE Trans Intell Veh*, **8**, 3324–3335.

Leng, J., Sun, C., Dong, H., Li, D., Zhang, C., Chen, P. C. Y., 2025. Adaptive eco-driving with guided speed planning and lane changing through signalized intersections. *IEEE Trans Transp Electr*, **11**, 8365–8376.

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

Li, L., Li, X., 2019. Parsimonious trajectory design of connected automated traffic. *Transp Res Part B Methodol*, **119**, 1–21.

Littman, M. L., 1994. Markov games as a framework for multi-agent reinforcement learning. In: Proceedings of the Eleventh International Conference on International Conference on Machine Learning, 157–163.

Liu, M., Zhao, J., Hoogendoorn, S., Wang, M., 2022. A single-layer approach for joint optimization of traffic signals and cooperative vehicle trajectories at isolated intersections. *Transp Res Part C Emerg Technol*, **134**, 103459.

Liu, Q., Guo, Y., An, Y., Li, M., 2025. Joint lane management and signal optimization for mixed-autonomy intersections: An analytical approach. *Phys A Stat Mech Appl*, **668**, 130556.



- Liu, Q., Lin, X., Li, M., Li, L., He, F., 2023a. Coordinated lane-changing scheduling of multilane CAV platoons in heterogeneous scenarios. *Transp Res Part C Emerg Technol*, **147**, 103992.
- Liu, Q., Zhang, K., Li, M., Chen, X., Lin, X., Li, S., 2024. Integrated optimization of traffic signal timings and vehicle trajectories considering mandatory lane-changing at isolated intersections. *Transp Res Part C Emerg Technol*, **163**, 104614.
- Liu, Y., Wu, F., Liu, Z., Wang, K., Wang, F., Qu, X., 2023b. Can language models be used for real-world urban-delivery route optimization? *Innovation*, **4**, 100520.
- Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., Mordatch, I., 2017. Multiagent Actor-critic for Mixed Cooperative-competitive Environments. <https://doi.org/10.48550/arXiv.1706.02275>
- Lu, G., Nie, Y., Liu, X., Li, D., 2019. Trajectory-based traffic management inside an autonomous vehicle zone. *Transp Res Part B Methodol*, **120**, 76–98.
- Ma, C., Yu, C., Yang, X., 2021. Trajectory planning for connected and automated vehicles at isolated signalized intersections under mixed traffic environment. *Transp Res Part C Emerg Technol*, **130**, 103309.
- Ma, J., Li, X., Zhou, F., Hu, J., Park, B. B., 2017. Parsimonious shooting heuristic for trajectory design of connected automated traffic part II: Computational issues and optimization. *Transp Res Part B Methodol*, **95**, 421–441.
- Milanés, V., Shladover, S. E., Spring, J., Nowakowski, C., Kawazoe, H., Nakamura, M., 2014. Cooperative adaptive cruise control in real traffic situations. *IEEE Trans Intell Transp Syst*, **15**, 296–305.
- Pan, T. L., Lam, W. H. K., Sumalee, A., Zhong, R. X., 2016. Modeling the impacts of mandatory and discretionary lane-changing maneuvers. *Transp Res Part C Emerg Technol*, **68**, 403–424.
- Papoudakis, G., Christianos, F., Schäfer, L., Albrecht, S.V., 2021. Benchmarking Multiagent Deep Reinforcement Learning Algorithms in Cooperative Tasks. <https://doi.org/10.48550/arXiv.2006.07869>
- Peng, B., Keskin, M. F., Kulcsár, B., Wymeersch, H., 2021. Connected autonomous vehicles for improving mixed traffic efficiency in unsignalized intersections with deep reinforcement learning. *Commun Transp Res*, **1**, 100017.
- Schulman, J., Moritz, P., Levine, S., Jordan, M., Abbeel, P., 2015. High-dimensional Continuous Control Using Generalized Advantage Estimation. <https://doi.org/10.48550/arXiv.1506.02438>
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O., 2017. Proximal Policy Optimization Algorithms. <https://doi.org/10.48550/arXiv.1707.06347>
- Shang, Y., Zhu, F., Jiang, R., Li, X., Wang, S., 2024. Trajectory planning at a signalized road section in a mixed traffic environment considering lane-changing of CAVs and stochasticity of HDVs. *Transp Res Part C Emerg Technol*, **158**, 104441.
- Treiber, M., Hennecke, A., Helbing, D., 2000. Congested traffic states in empirical observations and microscopic simulations. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics*, **62**, 1805–1824.
- Wang, Q., Gong, Y., Yang, X., 2022a. Connected automated vehicle trajectory optimization along signalized arterial: A decentralized approach under mixed traffic environment. *Transp Res Part C Emerg Technol*, **145**, 103918.
- Wang, S., Wang, Z., Jiang, R., Yan, R., Du, L., 2022b. Trajectory jerking suppression for mixed traffic flow at a signalized intersection: A trajectory prediction based deep reinforcement learning method. *IEEE Trans Intell Transp Syst*, **23**, 18989–19000.
- Wang, S., Wang, Z., Jiang, R., Zhu, F., Yan, R., Shang, Y., 2024. A multi-agent reinforcement learning-based longitudinal and lateral control of CAVs to improve traffic efficiency in a mandatory lane change scenario. *Transp Res Part C Emerg Technol*, **158**, 104445.
- Witt, C. S. de, Gupta, T., Makoviichuk, D., Makoviychuk, V., Torr, P. H. S., Sun, M., et al., 2020. Is Independent Learning All You Need in the StarCraft Multiagent Challenge. <https://doi.org/10.48550/arXiv.2011.09533>
- Wu, J., Ahn, S., Zhou, Y., Liu, P., Qu, X., 2021. The cooperative sorting strategy for connected and automated vehicle platoons. *Transp Res Part C Emerg Technol*, **123**, 102986.
- Wu, J., Qu, X., 2022. Intersection control with connected and automated vehicles: A review. *J Intel Connect Veh*, **5**, 260–269.
- Xiong, B. K., Jiang, R., 2022. Speed advice for connected vehicles at an isolated signalized intersection in a mixed traffic flow considering stochasticity of human driven vehicles. *IEEE Trans Intell Transp Syst*, **23**, 11261–11272.
- Yang, S., Xu, Y., Wang, P., Li, D., 2024. An efficient rolling-horizon approach for cooperative multi-lane platoon formation with undefined configurations. *IEEE Trans Intell Transp Syst*, **25**, 21608–21621.
- Yang, Y., Zhan, J., Xu, M., Liu, Y., Qu, X., 2026. Toward climate-neutral urban mobility: Understanding shared e-scooter carbon emission patterns through multi-city evidence in Europe. *Transp Res Part A Policy Pract*, **203**, 104736.
- Yao, H., Li, X., 2021. Lane-change-aware connected automated vehicle trajectory optimization at a signalized intersection with multi-lane roads. *Transp Res Part C Emerg Technol*, **129**, 103182.
- Ye, F., Hao, P., Qi, X., Wu, G., Boriboonsomsin, K., Barth, M. J., 2019. Prediction-based eco-approach and departure at signalized intersections with speed forecasting on preceding vehicles. *IEEE Trans Intell Transp Syst*, **20**, 1378–1389.
- Ying, J., Feng, Y., 2024. Infrastructure-Assisted cooperative driving and intersection management in mixed traffic conditions. *Transp Res Part C Emerg Technol*, **158**, 104443.
- Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., et al., 2021. The Surprising Effectiveness of PPO in Cooperative, Multiagent Games. <https://doi.org/10.48550/arXiv.2103.01955>
- Zhang, J., Chang, C., Zeng, X., Li, L., 2023. Multi-agent DRL-based lane change with right-of-way collaboration awareness. *IEEE Trans Intell Transp Syst*, **24**, 854–869.
- Zhang, S., Zhang, J., Yang, L., Chen, F., Li, S., Gao, Z., 2024. Physics guided deep learning-based model for short-term origin-destination demand prediction in urban rail transit systems under pandemic. *Engineering*, **41**, 276–296.
- Zhao, W., Ngoduy, D., Shepherd, S., Liu, R., Papageorgiou, M., 2018. A platoon based cooperative eco-driving model for mixed automated and human-driven vehicles at a signalised intersection. *Transp Res Part C Emerg Technol*, **95**, 802–821.
- Zheng, Z., 2014. Recent developments and research needs in modeling lane changing. *Transp Res Part B Methodol*, **60**, 16–32.
- Zhou, F., Li, X., Ma, J., 2017. Parsimonious shooting heuristic for trajectory design of connected automated traffic part I: Theoretical analysis with generalized time geography. *Transp Res Part B Methodol*, **95**, 394–420.
- Zhou, H., Zhou, A., Li, T., Chen, D., Peeta, S., Laval, J., 2022. Congestion-mitigating MPC design for adaptive cruise control based on Newell's car following model: History outperforms prediction. *Transp Res Part C Emerg Technol*, **142**, 103801.
- Zhou, M., Yu, Y., Qu, X., 2020. Development of an efficient driving strategy for connected and automated vehicles at signalized intersections: A reinforcement learning approach. *IEEE Trans Intell Transp Syst*, **21**, 433–443.
- Zhou, Z., Rother, C., Chen, J., 2023. Event-triggered model predictive control for autonomous vehicle path tracking: Validation using CARLA simulator. *IEEE Trans Intell Veh*, **8**, 3547–3555.



Xingwei Jiang is currently pursuing the concurrent B.Eng. degrees in civil engineering systems and in mathematics and physics at Department of Civil Engineering, Tsinghua University, Beijing, China. His research interests encompass reinforcement learning and intelligent transportation systems.



Meng Li received the B.S. degree in structure engineering from Tsinghua University, Beijing, China, in 2001, the M.S. degree in transportation science and supply chain management and the M.E. degree in transportation engineering from the University of California, Berkeley, USA, in 2003 and 2004, respectively, and the Ph.D. degree in transportation engineering from Nagoya University, Japan, in 2010. He is currently a

Professor with the Department of Civil Engineering, Tsinghua University, working in the fields of transportation big data analysis, intelligent transportation systems, and automated transportation. He is the General Scientific Committee Member and the Topic Area Manager (TAM) of the World Conference on Transport Research Society (WCTRS) and the Committee Member of the Sub-Committee of the Transportation Research Board (TRB), USA. He serves as the Editorial Board Member for the *Journal of Intelligent Transportation Systems* and an Associate Editor for the *Asia-Pacific Journal of Operational Research*, *IEEE Open Journal of Intelligent Transportation Systems*, and *Smart and Resilient Transportation*.



Qingquan Liu received the B.S. degree in automation from Tsinghua University, Beijing, China, in 2020, and the Ph.D. degree in civil engineering (transportation) from Tsinghua University in 2025. He is currently an Associate Professor with the College of Mechanical and Vehicle Engineering, Chongqing University, Chongqing, China. His research interests include cooperative control for connected and autonomous vehicles and traffic optimization in

mixed traffic environments.