https://doi.org/10.26599/JICV.2026.9210085

Research Article

Learning to drive from naturalistic trajectories: Offline reinforcement learning for safe speed guidance at signalized intersections

Xiaoyu Shi[1], Yuhuan Lu[2], Zihao Sheng[3], Jian Zhang[4], Sikai Chen[3,✉], Heye Huang[5], Tiantian Chen[1,✉]

[1]*Cho Chun Shik Graduate School of Mobility, Korea Advanced Institute of Science and Technology, Daejeon 34141, Republic of Korea*

[2]*State Key Laboratory of Internet of Things for Smart City, University of Macau, Macau 999078, China*

[3]*Department of Civil and Environmental Engineering, University of Wisconsin–Madison, Madison WI 53706, USA*

[4]*School of Transportation, Southeast University, Nanjing 210096, China*

[5]*State Key Laboratory of Automotive Safety and Energy, Tsinghua University, Beijing 100084, China*

✉ Corresponding authors.

E-mail: S. Chen, sikai.chen@wisc.edu; T. Chen, nicole.chen@kaist.ac.kr

Received: December 30, 2025; Revised: March 15, 2026; Accepted: May 4, 2026

Abstract

This study proposes an offline reinforcement learning framework based on Critic Regularized Regression (CRR) to optimize speed guidance at signalized intersections under mixed traffic conditions. Using real-world trajectory data combined with signal phase and timing (SPaT) information, the framework learns safe and efficient driving policies without requiring online interactions. A structured data-processing pipeline converts raw vehicle trajectories into

Markov Decision Process (MDP) format, effectively encoding vehicle motion states and signal timing information to facilitate realistic decision-making. SUMO simulation experiments demonstrate substantial improvements over rule-based baselines in safety, comfort, and efficiency: time-to-collision increases from 2.75 to 8.53 seconds, jerk is reduced by over 50%, and time headway is consistently maintained at approximately 1.68 seconds. Trajectory visualizations confirm smoother and more adaptive driving behavior. Comparisons with state-of-the-art approaches including Model Predictive Control (MPC), Behavior Cloning (BC), Twin Delayed Deep Deterministic Policy Gradient (TD3), and Batch Constrained Q-learning (BCQ) highlight CRR's stable performance across multiple evaluation metrics. Ablation studies reveal the critical role of different components in ensuring robust policy behavior, while communication loss simulations demonstrate framework resilience. The method's computational efficiency, requiring considerably less time than MPC, and robustness to communication failures reinforce its practicality for real-time deployment.

Keywords

Offline Reinforcement Learning, signalized intersections, speed guidance, Markov Decision Process, trajectory data processing, Critic Regularized Regression, mixed traffic flow

1 Introduction

Urban growth has led to sustained increases in travel demand, which in turn elevate both congestion levels and road users' exposure to safety-critical conflicts. As key nodes where multiple streams compete for priority, signalized intersections often become network bottlenecks and strongly influence system-wide efficiency and safety (Yong et al., 2015; Milakis et al., 2017; Soteropoulos et al., 2019). In parallel, developments in cloud computing, data analytics, and the Internet of Vehicles (IoV) have enabled Intelligent Transportation Systems (ITS) to support connectivity-assisted operations (Xu et al., 2025; Shi et al., 2025). Through Vehicle-to-Infrastructure (V2I) and Vehicle-to-Vehicle (V2V) communication, vehicles and controllers can exchange motion states and signal status in real time, creating opportunities to mitigate queuing and reduce crash risk. With such technologies being deployed, traffic increasingly

operates under mixed autonomy, in which connected automated vehicles (CAVs) share the roadway with human-driven vehicles (HDVs) (Dong et al., 2024; Wu et al., 2024; Gao et al.,2025; Liu et al., 2023).

Meanwhile, onboard sensors and connected controllers continuously generate high-resolution trajectory and signal datasets, including NGSIM (U.D. of Transportation), HighD (Krajewski et al., 2018), US-Accident (Moosavi et al., 2019), and UCF-SST-CitySim (Zheng et al., 2023). These resources have supported data-driven studies on signal operation, surrogate risk assessment, and driver behavior (Wei et al., 2022; Yang et al., 2021; Zhang et al., 2025). Building on this foundation, V2I-enabled systems such as Green Light Optimal Speed Advisory (GLOSA), Advisory Speed Limit (ASL) (Wu et al., 2015; Jiang et al., 2017; Yao et al., 2018), Cooperative Adaptive Cruise Control (CACC), Intelligent Speed Adaptation (ISA), and Connected Vehicle–Infrastructure Cooperation (CVIC) aim to shape approach trajectories to improve safety and reduce delay (Güvenç et al., 2012; Yang et al., 2017; Li et al., 2014; Xia et al., 2013; Rakha and Kamalanathsharma, 2011). Existing approaches are commonly framed as (i) rule-based strategies (e.g., GLOSA) and (ii) optimization-based controllers (e.g., MPC (Wang et al., 2018), Dynamic Programming (DP) (Sun et al., 2020; Ma and Yang, 2021), and Pontryagin's Maximum Principle (PMP)). Rule-based methods can be overly rigid, while optimization-based methods may become computationally difficult for real-time deployment.

Deep reinforcement learning (DRL) provides a promising alternative by enabling agents to learn effective strategies directly through interaction with their environment (Han et al., 2025). Similar to human learning, vehicle agents adapt via trial and error, gradually improving their ability to balance safety, maneuverability, and comfort while maximizing cumulative rewards (Guo et al., 2021; Shi et al., 2018, Qi et al., 2016. Zhang et al., 2022). This paradigm is particularly well-suited to complex, dynamic, and nonlinear decision-making problems. Autonomous driving control is typically nonlinear and interaction-driven, which makes model-free RL a practical choice in many studies. Depending on the action structure, different RL families are adopted. For discrete decisions (e.g., phase selection or high-level maneuvers), value-learning methods such as Q-learning and Deep Q-Networks (DQN) are commonly used

(Mnih et al., 2015). For continuous control (e.g., longitudinal acceleration), policy-gradient methods including Deterministic Policy Gradient (DPG) (Mousavi et al., 2017) and Proximal Policy Optimization (PPO) directly optimize parameterized policies. Actor–critic algorithms (e.g., DDPG, TD3, and Soft Actor-Critic, SAC) combine policy optimization with value estimation and have been applied to speed regulation and adaptive decision-making in dynamic traffic (Dong et al., 2022; Lu et al., 2025; Zhou et al., 2020; Zhu et al., 2020; Shi et al., 2024; Jiang et al., 2023).

Despite their effectiveness, online RL algorithms demand extensive interactions with the environment during training. As illustrated in **Fig.** 1, on-policy RL (a) updates the current policy π_k using data generated by itself, whereas off-policy RL (b) leverages a replay buffer $\mathcal{D}$ that stores samples from all previous policies $\pi_0, \ldots, \pi_k$ to train the next policy π_{k+1}. Although online RL can be effective, it typically requires repeated interaction and exploration during training, which can be impractical for real vehicles due to safety constraints and data-collection cost. Training purely in simulation also raises deployment concerns, as learned behaviors may not transfer reliably to real-world traffic. Offline RL addresses these limitations by learning from pre-collected trajectories and deploying only the finalized policy, thereby reducing the need for risky online exploration and improving the feasibility of learning directly from naturalistic driving data.

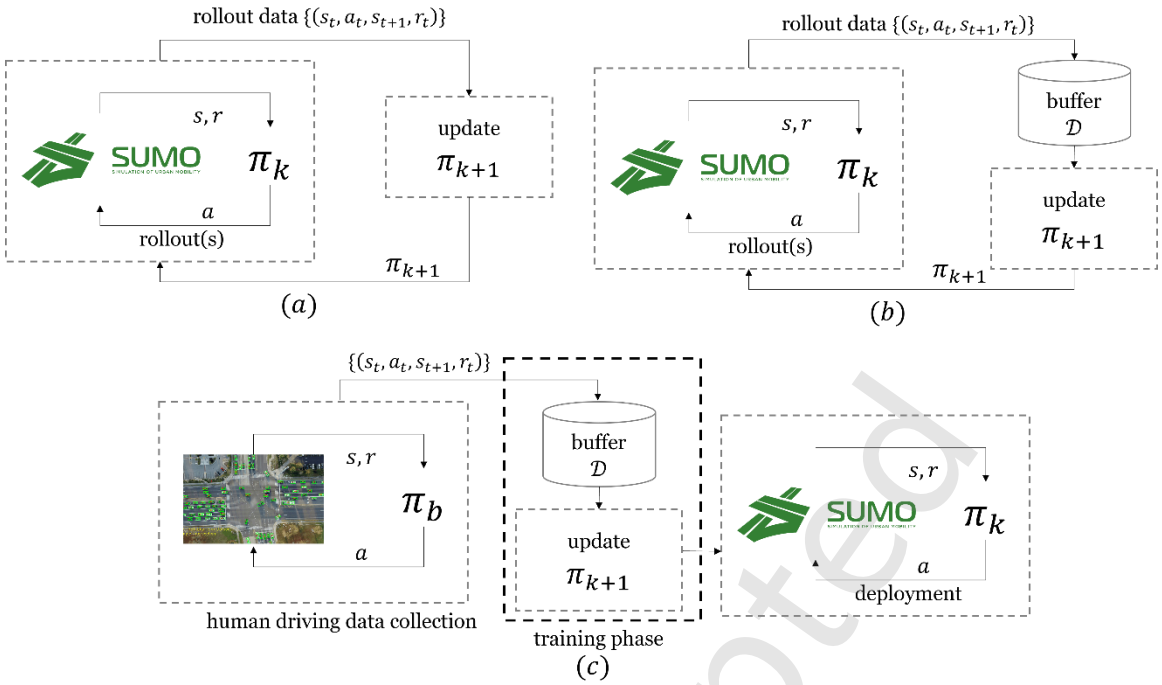


Fig. 1: Pictorial illustration of three RL patterns. (a) online RL, (b) off-policy RL, (c) offline RL; s represents the surrounding state that can be observed, a means the chosen action, r denotes the corresponding reward.

The availability of comprehensive historical traffic state and action datasets has facilitated the development of data-driven offline RL paradigms, which effectively leverage existing data to learn optimized policies. As a result, offline RL has increasingly been applied to autonomous driving control tasks. For example, Hu and Li (Hu and Li, 2022) developed an energy-efficient driving strategy for hybrid autonomous vehicles using offline RL techniques, while Dai et al. (Li et al., 2016) employed these methods to train efficient traffic signal controllers. Various combinations of human-driving, exploratory, and online-buffer datasets have been explored to identify optimal policies using offline RL algorithms (Fang et al., 2022). Benchmarks indicate that offline RL approaches can reliably maintain safe distances and prevent collisions across diverse scenarios, including highway traffic, lane reductions, and cut-in maneuvers (Lee et al., 2024). However, existing offline RL methods for autonomous driving often neglect the critical role of SPaT data and surrogate safety measures. To address this gap, our work extracts these features from real-world driving data to enhance safety performance. While human driving data may lack the precision of expert-annotated or simulated datasets, carefully designed models and algorithms can transform such suboptimal data into valuable training resources. This process enables the identification of both beneficial features

and inherent vulnerabilities in human driving patterns, ultimately supporting the development of robust driving policies (Dai et al., 2022; Zhang et al., 2019; Bhaskar et al., 2011; Li and Ban, 2019). Although SPaT-based trajectory datasets and DRL have been widely used in traffic management, prior research has primarily focused on traffic signal control (Xu et al., 2025; Shuvo et al., 2024) rather than leveraging SPaT information to improve vehicle safety performance.

Compared with existing literature, the primary contributions of this paper are as follows. First, we construct a comprehensive MDP dataset derived from real-world human driving trajectories, integrating SPaT data, surrogate safety measures, and comfort metrics. This approach moves beyond traditional crash-based evaluation methods and reduces the reliance on purely simulated training data by learning directly from naturalistic driving trajectories, thereby grounding the policy in real-world driving behavior. Second, we address the challenges of mixed-traffic environments by evaluating our framework under varying CAV Market Penetration Rates (MPRs) and realistic operational constraints. Systematic testing under demanding scenarios, including V2I/V2V communication failures, information delays, and incomplete sensor data, demonstrates the framework's robustness and delineates its operational boundaries, offering critical insights for real-world deployment where ideal conditions cannot be assumed. Third, we perform a comprehensive comparative simulation analysis of the CRR-based offline RL paradigm against classical control strategies, including rule-based GLOSA, optimization-based MPC, online reinforcement learning TD3, and behavior cloning methods. Combined with sensitivity analyses and ablation studies, these evaluations demonstrate the superiority of our offline RL framework and identify the essential components and boundary conditions required for effective deployment in ITS.

The remainder of this paper is organized as follows. Section II reviews the fundamentals of reinforcement learning employed in our framework and the comparative analyses. Section III describes the process of converting raw driving data into the target MDP format. Section IV defines the research scenarios and the components of the MDP. Section

V presents simulation experiments conducted in SUMO and evaluates the proposed approach against several state-of-the-art (SOTA) methods. Finally, Section VI provides concluding remarks.

2 Preliminaries

2.1 Deep Reinforcement Learning

DRL combines the decision-making capabilities of reinforcement learning with the representation power of deep neural networks, making it effective for problems involving high-dimensional inputs and outputs. The mathematical foundation rests on the MDP framework, characterized by the tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$. Here, $\mathcal{S}$ encompasses all possible states, $\mathcal{A}$ represents available actions, $P(s' \mid s, a)$ describes transition dynamics from state s to s' given action $a, R(s, a)$ quantifies immediate feedback, and $\gamma \in [0,1]$ weighs future versus immediate outcomes. Within the DRL paradigm, an agent sequentially observes state $s_t \in \mathcal{S}$ at each timestep t, executes action $a_t \in \mathcal{A}$ following policy $\pi(a \mid s)$, and obtains reward r_t. The learning objective centers on maximizing expected long-term return:

$$G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k} \tag{1}$$

The Q-function $Q^\pi(s, a)$ encodes expected cumulative rewards when initiating from state s with action a, then adhering to policy π :

$$Q^\pi(s, a) = \mathbb{E}^\pi[G_t \mid s_t = s, a_t = a] \tag{2}$$

Value-based DRL methods focus on estimating $Q(s, a)$ to inform action selection. The recursive Bellman formulation provides:

$$Q(s, a) = \mathbb{E}\left[r + \gamma \max_{a'} Q(s', a') \mid s, a\right] \tag{3}$$

where s' denotes the subsequent state. This equation enables iterative refinement of $Q(s, a)$ through sampled experience. For large or continuous state spaces, neural networks serve as function approximators for $Q(s, a)$,

enabling generalization. Value-based techniques form a cornerstone of numerous DRL algorithms by optimizing the action-value function to derive effective policies through selecting actions that maximize $Q(s, a)$.

Policy-based approaches directly optimize the parameterized policy $\pi_\theta(a \mid s)$ by tuning θ to maximize expected return $J(\pi_\theta) = \mathbb{E}^{\pi_\theta}[G_t]$. The gradient computation follows:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}^{\pi_\theta}[\nabla_\theta \log\, \pi_\theta(a \mid s) Q^{\pi_\theta}(s, a)] \tag{4}$$

Actor-critic architectures merge benefits from both paradigms. The actor component handles policy π_θ updates, while the critic approximates the value function. Through temporal difference error minimization, the critic reduces estimation variance and stabilizes policy learning.

2.2 Offline Reinforcement Learning

Offline RL optimizes a policy using a static dataset of previously logged transitions, without collecting additional experience during training. This paradigm is attractive in safety-critical settings where online exploration is costly, risky, or operationally infeasible. A central challenge is dataset–policy mismatch: when the learned policy queries state–action regions that are weakly covered (or absent) in the logged data, value estimates can become unreliable due to out-of-distribution extrapolation. Accordingly, many offline RL methods introduce explicit regularization toward the behavior policy and adopt conservative critic learning to improve robustness under distribution shift.

3 Data Preparation

The offline reinforcement learning approach is evaluated using the open-source UCF-SST-CitySim dataset (Zheng et al., 2023), which comprises vehicle trajectories captured via drone footage across multiple locations in the United States. For this study, data from the intersection of University Boulevard and Alafaya Trail near the University of Central Florida in Orlando are employed. This intersection consists of nine lanes per direction and eight distinct signal phases, accommodating both left-turn and through movements for all cardinal directions. The dataset is accompanied by a detailed lane map and coordinate system.

3.1 Dataset Description

The raw dataset contains 4,035 vehicle trajectories along with 60 minutes of signal timing data, recorded at 30 frames per second. Each trajectory entry includes the vehicle's unique identifier, position coordinates, speed, and lane assignment, as summarized in Table 1. For the present analysis, only southbound trajectories exceeding 20 meters in length are selected, yielding a refined set of 750 trajectories. The 20-meter threshold is applied to exclude very short trajectory fragments, typically arising from vehicles entering or exiting the camera's field of view at the boundary, that do not capture a sufficient approach distance for meaningful speed guidance decisions. Signal timing and trajectory data are then integrated, with frames synchronized and variables converted into an MDP format to support agent training.

Table 1: Description of Trajectory CSV Columns

Name	Description
frameNum	Frame index at 30 FPS
carId	Unique vehicle identifier
boundingBox1Xft	X-coordinate of vehicle's bounding box
speed	Vehicle speed (m/s)
laneId	Lane identifier

3.2 Feature Extraction and Data Processing

Observations for the agent are defined as tuples (s_t, a_t, r_t, s_{t+1}), where s_t includes the distance to the stop line, vehicle speed, signal timing variables and phase, as well as the relative position, speed, and acceleration differences between the ego vehicle and a potential leading vehicle. The distance to the stop line is computed by comparing the bounding box positions of the ego vehicle with the stop line location. Signal-related variables are obtained by merging trajectory data with signal timing information using aligned frame IDs; in cases of inconsistencies, average signal durations and phases are used to adjust and extend the signal cycle to ensure consistency. Acceleration a_t is

calculated from changes in speed over consecutive time intervals. To identify the leading vehicle for the ego vehicle in each frame, all vehicles in the same lane are considered, and the leading vehicle is defined as the one that is both closer to the stop line than the ego vehicle and nearest to it. Once identified, the relative position, speed, and acceleration differences are computed. If no leading vehicle is present, the relative position is set to 1000 m, a value far exceeding the communication and sensing range of the intersection scenario, while the speed and acceleration differences are set to zero.

The reward r_t is defined based on travel distance, jerk, and time-to-collision (TTC) at each frame. Travel distance is calculated as the change in bounding box position over a single time step, while jerk is derived from the corresponding change in acceleration. TTC is computed using the positions and speeds of the ego vehicle and its leading vehicle. Frames are sampled at 0.5-second intervals, reducing the original 30 FPS data to 2 Hz. This interval balances computational efficiency with sufficient temporal resolution for capturing vehicle dynamics at intersection approach speeds. Furthermore, 0.5 seconds aligns with the simulation step length used in SUMO evaluation, ensuring consistency between training data and deployment conditions.

4 Methodology

4.1 Research Scenario

This study investigates a scenario in which a connected and automated vehicle (CAV) navigates a signalized intersection within an intelligent transportation system. As illustrated in **Fig.** 2, V2I communication enables transmitting devices to provide the vehicle with SPaT information, as well as data on the state of preceding vehicles. Using the received information, the CAV adjusts its acceleration based on its current position and speed, generating an optimized speed profile and trajectory. The primary objective is to allow the vehicle to cross the intersection smoothly, minimizing abrupt acceleration or deceleration, and ideally passing during the green phase to improve both safety and ride comfort.

To establish a performance benchmark and define a theoretical upper bound, we first construct a baseline model under idealized conditions, assuming:

- Perfect V2I and V2V communication, with complete availability of SPaT information;
- Real-time access to vehicle state data, including position, speed, and acceleration;
- Negligible communication latency and sensor noise.

Although these assumptions represent an idealized scenario, they provide a necessary baseline for assessing the framework's theoretical capabilities and serve as a reference for quantifying performance degradation under realistic conditions.

Acknowledging the challenges of real-world deployments, we systematically relax these idealized assumptions to evaluate the framework's robustness. Section V presents comprehensive ablation studies that examine:

- Communication disruptions: Performance under partial or complete V2I/V2V disconnections, reflecting real-world network instabilities;
- Information delay: Effects of varying communication latencies typical in vehicular networks;
- Incomplete information: Operation under missing or corrupted sensor or SPaT data;
- Mixed traffic conditions: Interaction with non-connected human-driven vehicles, where complete traffic state information is unavailable.

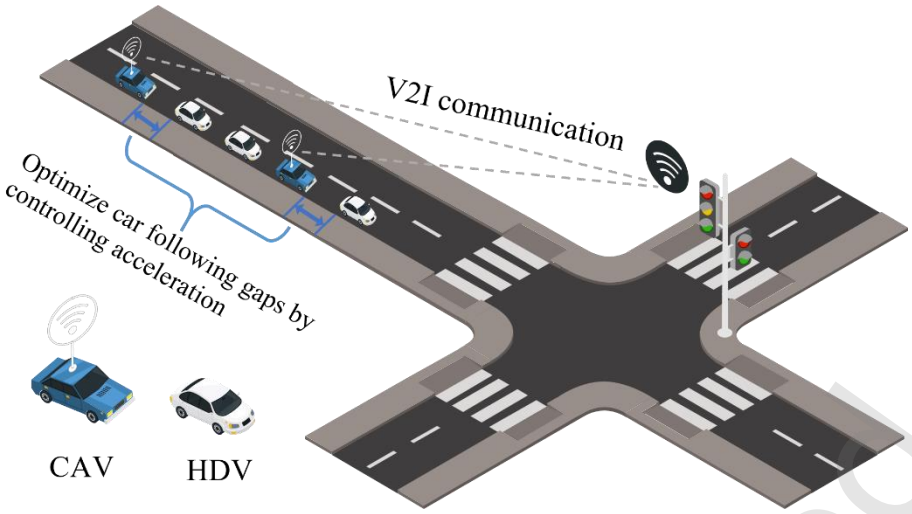


Fig. 2: Research Scenario of mixed traffic flow with CAVs and HDVs

Table 2: Parameters of the IDM Car-Following Model

Parameter (Symbol)	Value
Minimum gap when standing (s_0)	2.5 m
Maximum acceleration (a)	2.6 $\mathrm{m/s^2}$
Emergency deceleration ($b_{\max}$)	9.0 $\mathrm{m/s^2}$
Maximum deceleration (b)	4.5 $\mathrm{m/s^2}$
Maximum speed (v_0)	14.0 $\mathrm{m/s}$
Startup delay (T)	0.0 s
Acceleration exponent (δ)	4.0
Internal step length when computing follow speed(Δt)	0.25 s

4.2 Intelligent Driver Model (IDM)

The IDM represents a well-established car-following model that captures human driving dynamics by considering preferred velocity and safe spacing requirements (Treiber et al., 2000). This research employs IDM to simulate HDV behavior. The model generates continuous acceleration outputs based on current velocity, gap to the front vehicle, and closing rate. For vehicle α, acceleration $\dot{v}_\alpha$ is computed as:

$$\dot{v}_\alpha = a_\alpha \left(1 - \left(\frac{v_\alpha}{v_\alpha^0}\right)^\delta - \left(\frac{s^*(v_\alpha, \Delta v_\alpha)}{s_\alpha}\right)^2\right) \tag{5}$$

In this expression, a_α signifies maximum acceleration capacity, v_α indicates present velocity, v_α^0 represents target velocity, δ controls acceleration sensitivity, s_α denotes current spacing to the leader, and $s^*(v_\alpha, \Delta v_\alpha)$ specifies desired spacing dependent on both velocity v_α and relative speed Δv_α.

The desired spacing function $s^*(v, \Delta v)$ takes the form:

$$s^*(v, \Delta v) = s_0 + s_1 \sqrt{\frac{v}{v_0}} + Tv + \frac{v\Delta v}{2\sqrt{ab}} \tag{6}$$

Here, s_0 sets minimum standstill distance, s_1 adjusts for velocity-dependent spacing, T defines safe time gap, and b indicates comfortable deceleration. All parameters adopt SUMO default values listed in Table 2.

4.3 Critic Regularized Regression (CRR)

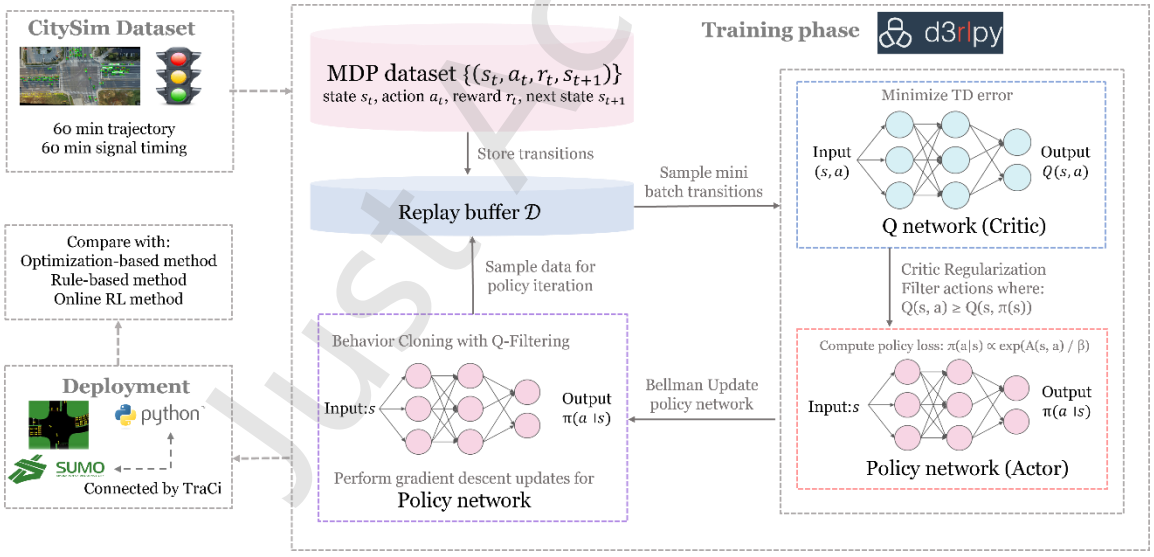


Fig. 3: Framework of the Critic Regularized Regression Algorithm

CRR is an offline reinforcement learning algorithm developed to address the limitations of conventional off-policy RL in offline settings (Wang et al., 2020). CRR employs a critic-based filtering mechanism to select high-quality actions from the dataset while discarding those with low critic evaluations. This approach mitigates the overestimation of Q-values commonly observed in offline RL. Formally, the policy update can be expressed as:

$$\arg\max_{\pi} \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\mathcal{B}}[f(Q_\theta, \pi, \mathbf{s}, \mathbf{a}) \log \pi(\mathbf{a} \mid \mathbf{s})] \tag{7}$$

By constraining the policy to actions within the dataset distribution, CRR improves both stability and performance. CRR's critic-weighted filtering mechanism is well-suited for learning from naturalistic human driving data, which inherently contains a mixture of driving qualities. Unlike Behavior Cloning, which treats all demonstrated behaviors equally, CRR evaluates each state–action pair through its learned critic and selectively reinforces actions with above-average Q-values. **Fig.** 3 illustrates the CRR framework and its application to our problem, where the d3rlpy offline DRL library is used for implementation.

4.4 Markov Decision Process (MDP)

4.4.1 State Space

The agent's state is represented as an 8-dimensional vector:

$$s = [d, v, t_{\text{cur}}, t_{\text{next}}, p_{\text{cur}}, \Delta d, \Delta v, \Delta a] \tag{8}$$

where d denotes the distance to the stop line, v the current speed, t_{cur} the remaining duration of the current signal phase, and t_{next} the time until the next green phase (set to 0 during green). The variable p_{cur} indicates the current signal phase (1 for green, 0 otherwise). Under this encoding, the yellow phase is treated as a non-green state, adopting a safe approach that encourages the vehicle to prepare for stopping rather than attempting to clear the intersection during the transition period. The temporal distinction between yellow and red is still captured by the remaining duration variable t_{cur}. The terms $\Delta d, \Delta v$, and Δa represent the relative position, speed, and acceleration differences between the ego vehicle and a potential leading vehicle.

4.4.2 Action Space

To guarantee safe and comfortable operation, CAVs modulate their velocity responding to real-time observations. The MDP formulation employs a continuous action domain, avoiding local optimum issues common in discrete-action frameworks. Given safety constraints and physical limitations, policy-generated acceleration cannot directly

control vehicle motion and needs regulation. During execution, vehicles face speed limits and acceleration/deceleration bounds:

$$0 \leq v(t) \leq v_m, d_m \leq a(t) \leq u_m \tag{9}$$

where v_m denotes maximum road velocity, while d_m and u_m represent peak deceleration and acceleration respectively. Furthermore, acceleration must satisfy car-following safety requirements to prevent rear-end collisions. Denoting a_t as policy-recommended acceleration and $a_{\text{IDM}}(t)$ as car-following model acceleration, current timestep velocity is bounded by:

$$v(t) = \max\big(\min\big(v_m, v(t-1) + \min(a(t), a_{\text{IDM}}(t))\big), 0\big) \tag{10}$$

This formulation ensures that actual velocity remains within safe operational bounds while respecting both policy guidance and collision avoidance constraints. The $\min(a(t), a_{\text{IDM}}(t))$ selects the more conservative acceleration. When the policy recommends acceleration exceeding the IDM safe threshold, the IDM value prevails. Conversely, when the policy recommends stronger deceleration than IDM, the policy's command is followed, as more aggressive braking increases the safety margin.

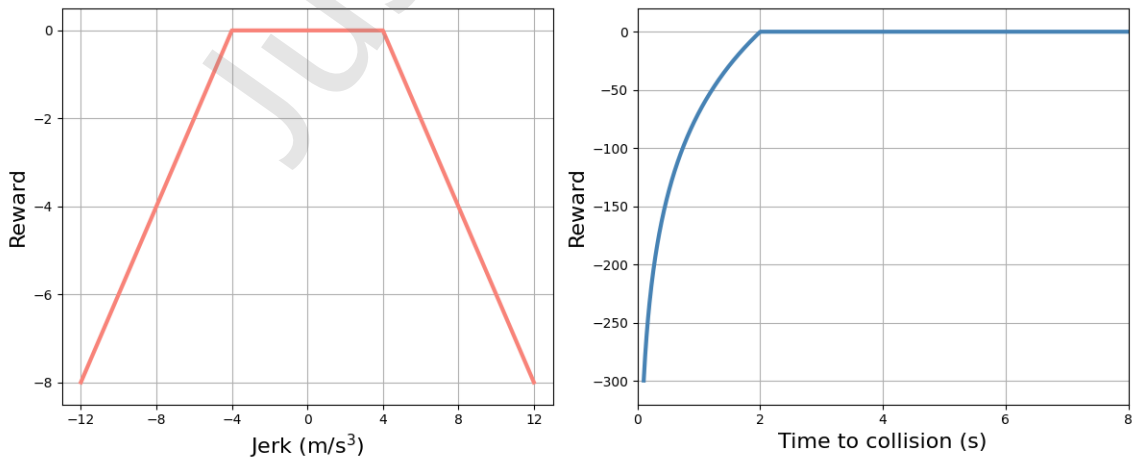


Fig. 4: Reward Function Incorporating TTC and Jerk

4.4.3 Reward Function

The reward function defines the optimization objective for the reinforcement learning algorithm. This study is designed to balance driving efficiency, comfort, and safety.

Specifically, the reward encourages minimizing traffic delays and promoting smooth driving behavior, while ensuring safe interactions with other vehicles.

First, because travel delay can only be fully assessed after the trip is complete, the travel distance within a single time step is used as a proxy for travel time. Under this formulation, this component of the reward function is denoted as r_t^e.

Second, in transportation engineering, "jerk", the rate of change of acceleration, is a key metric for motion smoothness and ride comfort. High jerk values correspond to rapid acceleration changes, which can cause passenger discomfort; hence, controlling jerk is critical for improving ride quality. Jerk is defined as:

$$\text{Jerk} = \frac{dA}{dt} \tag{11}$$

where A denotes acceleration and t time. A predefined threshold u_h represents the maximum allowable jerk to maintain passenger comfort and a smooth trajectory. Accordingly, this component of the reward function is formulated as:

$$r_t^u = \begin{cases} -(|u(t)| - u_h), & \text{if } |u(t)| > u_h \\ 0, & \text{if } |u(t)| \le u_h \end{cases} \tag{12}$$

This implies that when the jerk exceeds the predefined threshold, a penalty is applied within the reward function. Following previous studies, the threshold value in this work is set to 4 m/s³ (Zhao et al., 2018).

Third, surrogate safety measures (SSMs) are real-time metrics used to evaluate crash risk and are widely applied in highway and urban traffic simulations. Unlike traditional safety assessments, which depend on historical crash data,

such as comparing crash rates before and after infrastructure changes, SSMs enable proactive evaluation of potential conflicts.

However, in autonomous driving scenarios, actual crash data are often scarce, necessitating alternative safety metrics. One widely adopted measure is TTC, originally introduced by Hayward (Hayward, 1972), which quantifies the time remaining before a following vehicle would collide with a leading vehicle if both maintain their current speeds.

TTC is calculated as:

$$\mathrm{TTC} = \frac{X_{\mathrm{lead}} - X_{\mathrm{follow}} - L_{\mathrm{lead}}}{V_{\mathrm{follow}} - V_{\mathrm{lead}}} \tag{13}$$

where:

- X_{lead} and X_{follow} represent the positions of the leading and following vehicles, respectively;
- V_{lead} and V_{follow} denote the speeds of the leading and following vehicles;
- L_{lead} is the length of the leading vehicle.

In other studies, a more conservative threshold of 2 seconds has been adopted to further enhance safety (Li et al., 2017; Jeong et al., 2017).

Accordingly, this component of the reward function is defined as:

$$r_t^s = \begin{cases} 100 \times \ln\left(\frac{\mathrm{TTC}}{2}\right), & \text{if TTC} < 2 \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

This formulation penalizes the agent as TTC approaches 2 seconds, with smaller TTC values resulting in larger negative penalties.

Finally, the overall reward function is formulated as:

$$r_t = \omega_1 r_t^e + \omega_2 r_t^u + \omega_3 r_t^s \tag{15}$$

where ω_1, ω_2 and ω_3 are weighting coefficients that balance the contributions of the efficiency, comfort, and safety components. **Fig.** 4 illustrates how TTC and jerk values influence the overall reward.

5 Simulation Analysis

5.1 Simulation Settings

The safe-driving policy was trained using an MDP-formatted dataset derived from the UCF-SST CitySim dataset. Experiments were conducted in a SUMO-based microscopic simulation environment (SUMO v1.21.0), with Python 3.12.4 used to implement both the simulation pipeline and policy training. TraCI was employed for programmatic control of simulation runs, enabling automated parameter configuration and scenario management. All experiments were executed on a personal workstation featuring an Intel Core i9-14700K CPU and an NVIDIA GeForce RTX 4060 GPU. The target intersection, University Boulevard and Alafaya Trail, was reconstructed to evaluate the learned policy under diverse traffic conditions. Performance was assessed using safety-, comfort-, and efficiency-related metrics. For lateral maneuvers, both CAVs and HDVs were governed by SUMO's built-in LC2013 lane-changing model.

To optimize algorithm performance, a combination of Optuna search and grid search was employed to identify the hyperparameter set that maximizes training returns. The hyperparameter group includes the learning rate for the policy function (actor), learning rate for the Q function (critic), batch size, discount factor (γ), and temperature parameter (β). The optimal hyperparameter set was determined after 100 trials, with each trial's average return computed over five runs using different random seeds to mitigate stochastic variability. Both the actor and critic networks consist of two fully connected hidden layers with 256 units each, employing ReLU activation functions. Key hyperparameters for the proposed algorithm and comparative methods are summarized in Table 3.

Test vehicles traveled from the northern to the southern approach of the intersection. The simulation incorporated three signal phases for this direction, with a total cycle time of 90 seconds: 31 s for green, 4 s for yellow, and 55 s for red. Departure lanes for CAVs were randomized, and departure times were uniformly distributed across the signal

cycle to create dynamic traffic scenarios and evaluate the robustness of the learning algorithm. The reward function included penalties and weights to guide agent learning, with a penalty of -5 applied when a CAV's speed dropped below 0.5 m/s. Weight parameters were set as $\omega_1 = 1$ and $\omega_2 = 1$, and $\omega_3 = 1$. The sensitivity of performance to different weight combinations is systematically examined in Section 5.4 (Fig. 11), with a default departure speed of 0 m/s.

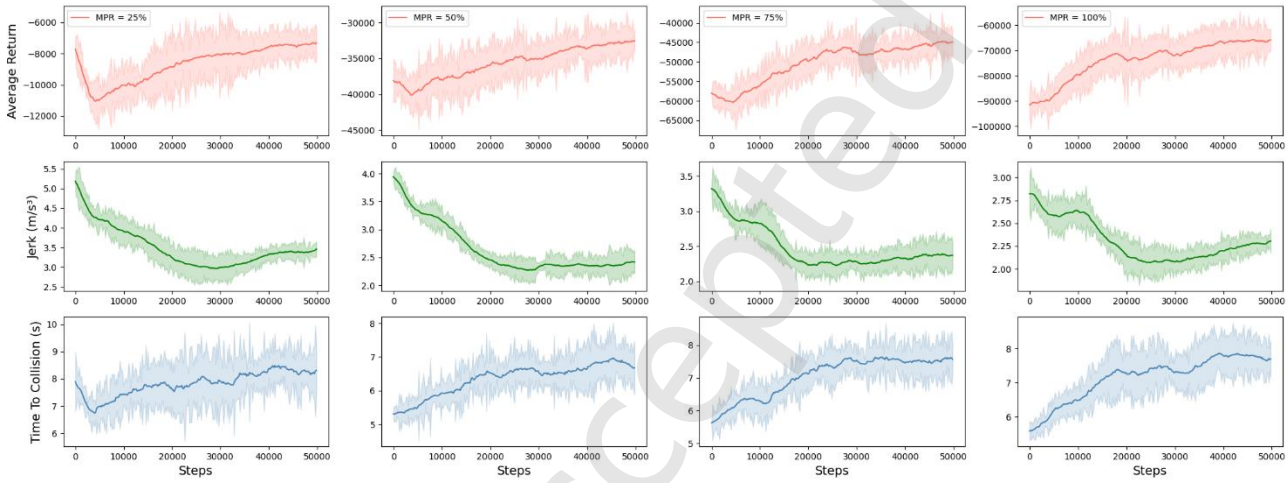


Fig. 5: Trends of Average Return, TTC, and Jerk During Training (solid curves: mean; shaded areas: standard deviation across five trials)

Fig. 5 illustrates the trends of average return, TTC, and jerk across five independent training rounds using different random seeds. During training, the policy was recalled and evaluated after each epoch, with a total of 250 epochs and 200 steps per epoch. Training performance was further assessed under mixed traffic scenarios with MPR of 0.25,0.5, and 0.75, with the 0.25 MPR scenario showing the largest fluctuations and deviations in performance metrics. Overall, both average return and safety metrics demonstrated convergence after approximately 20,000 training steps.

Table 3: Parameters Used in the Algorithms

Parameter	Value
CRR	
actor learning rate	0.0001
critic learning rate	0.0001

batch size	256
discount factor (γ)	0.99
temperature value (β)	0.5
target network synchronization coefficient	0.001
target network update interval	300
number of critic networks	2
BC	
learning rate	$1e-4$
BCQ	
actor learning rate	0.001
critic learning rate	0.001
imitator learning rate	0.001
batch size	256
discount factor (γ)	0.99
TD3	
actor learning rate	0.0003
critic learning rate	0.0003
batch size	256
discount factor (γ)	0.99
target network synchronization coefficient	0.005
number of critic networks	2

5.2 Overall Performance in Safety, Comfort, and Efficiency

5.2.1 Empirical Cumulative Distribution

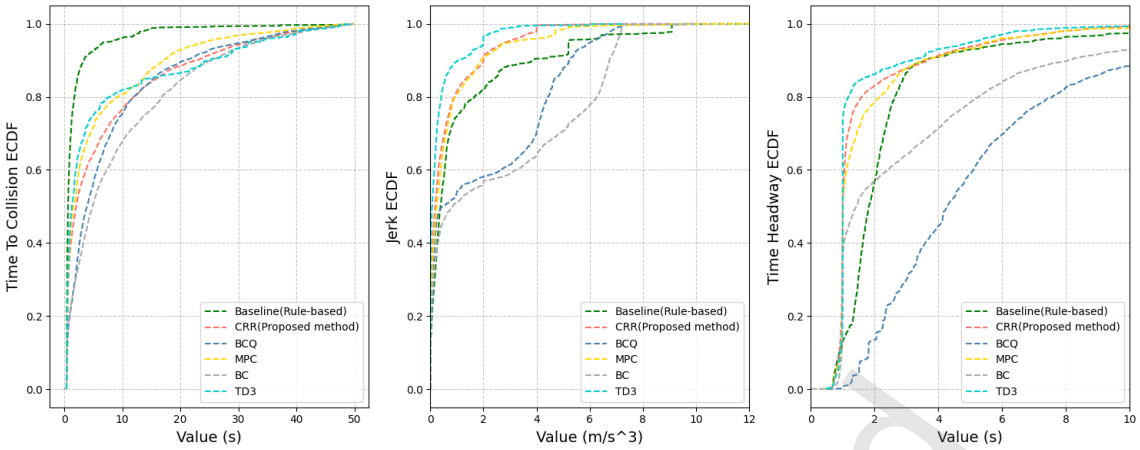


Fig. 6: Empirical Cumulative Distributions of TTC, Jerk, and Time Headway

Fig. 6 presents the empirical cumulative distribution functions (ECDFs) of three key performance metrics: time headway, TTC, and jerk, comparing the CRR-based method with the rule-based GLOSA baseline, optimization-based MPC, BCQ, BC, and the online RL-based TD3 method.

In terms of TTC, the CRR-based method substantially outperforms the rule-based approach in avoiding dangerous situations, such as TTC values below 10s, indicating a lower collision risk and improved overall safety. The average TTC increases to 8.53s, compared to 2.75s for the rule-based method. Regarding jerk, the CRR method achieves a superior balance between safety and comfort. Its ECDF curve shows lower jerk values than the rule-based method and exhibits similar optimization performance to MPC, reflecting smoother and more comfortable driving maneuvers. Specifically, the average jerk is reduced from $1.17\ \mathrm{m/s^3}$ in the rule-based method to $0.58\ \mathrm{m/s^3}$ in CRR, representing a 50% reduction in discomfort. For time headway, the CRR-based method consistently maintains reasonable values across the distribution, demonstrating its ability to preserve efficient inter-vehicle distances. In contrast, both GLOSA and MPC exhibit higher headway values, indicative of less efficient driving. While all four methods produce few instances of dangerously short headways (less than 1s), only the rule-based method generates a notable number of inefficient headways exceeding 2s. The CRR method achieves an average time headway of 1.68s, effectively balancing efficiency and safety.

Computational efficiency was also compared between CRR and MPC using 10,000 sampled decision instances. On average, CRR requires 0.0097s per decision, whereas MPC requires 0.0399s, approximately four times longer than CRR.

Table 4: Mean, Standard Deviation, Coefficient of Variation, and near-miss rate

Method	TTC Mean	TTC Std	TTC CV	Jerk Mean	Jerk Std	Jerk CV	TH Mean	TH Std	TH CV	near-miss rate (TTC<2s)	near-miss rate (TTC<0.5s)
BC	9.325	10.738	1.15	2.585	2.798	1.08	4.036	7.749	1.92	30.25%	12.82%
BCQ	7.679	9.608	1.25	2.046	2.277	1.11	5.810	6.587	1.13	31.90%	8.60%
CRR	7.046	10.243	1.45	0.602	0.964	1.60	1.728	2.342	1.36	49.43%	14.70%
GLOSA	1.873	4.420	2.36	1.168	1.971	1.69	2.802	5.681	2.03	83.17%	33.76%
MPC	5.601	8.725	1.56	0.693	1.199	1.73	1.898	3.203	1.69	52.86%	18.37%
TD3	6.370	10.907	1.71	0.349	0.717	2.05	1.550	1.841	1.19	60.08%	15.72%

Considering that safety is our primary optimization objective, we have added the near-miss rate in Table 4 (TTC < 2s or 0.5s, consistent with the TTC safety threshold used in the reward function or in extreme conditions). The results indicate that no method consistently outperforms the others across all safety-related indicators. CRR records stable safety performance and remains competitive while still reflecting the trade-offs among different safety objectives. Table 4 also presents the mean, standard deviation, and coefficient of variation (CV = Std/Mean) of TTC, jerk, and time headway for all methods. The statistical results further support the overall effectiveness of CRR in achieving a balanced performance.

5.2.2 Sampled Trajectories

Fig. 7: Sampled Trajectories: CRR (Left) vs. Baseline (Right)

Fig. 7 provides a comparative visualization of vehicle trajectories generated by the CRR-based policy and the baseline method (Wang et al., 2023). Vehicles are color-coded according to the active signal light phase, green, yellow, or red, at the time of motion. Trajectories are sampled at 0.5 -second intervals, so the spacing between points reflects the vehicle's instantaneous speed. In the baseline scenario, vehicles exhibit longer inter-point distances during acceleration phases, indicating higher peak speeds compared to those under the CRR policy. This behavior suggests that the baseline method accelerates more aggressively, potentially reducing travel time but at the expense of safety and ride comfort.

In contrast, the CRR-based policy produces smoother, more consistently controlled speed profiles, as reflected in the uniform spacing of trajectory points and restrained acceleration near the intersection. This behavior aligns with the primary objective of the proposed offline RL framework: to prioritize safety and comfort by explicitly penalizing low TTC values and excessive jerk during training. Vehicles under CRR control maintain safer headways and avoid abrupt maneuvers even under varying signal conditions. These trajectory patterns visually demonstrate that the CRR method effectively reduces risk while enabling efficient passage through intersections.

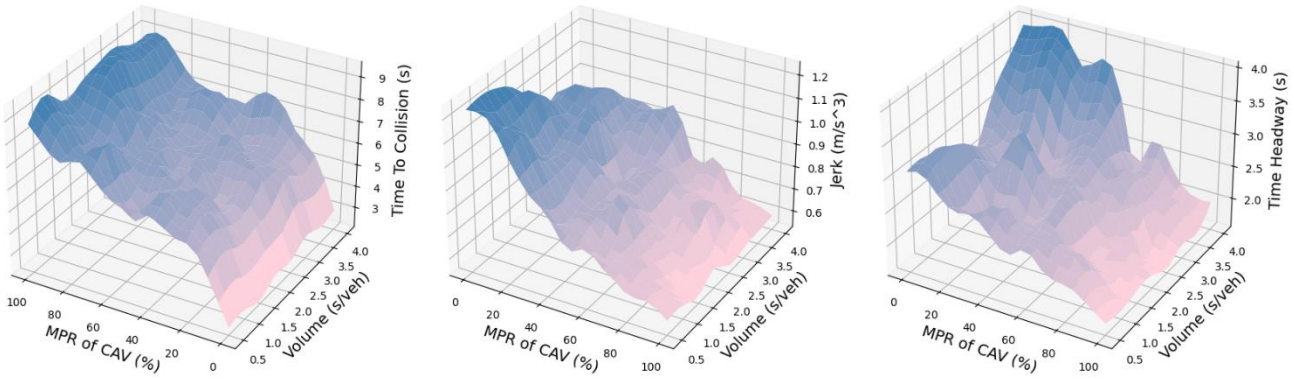


Fig. 8: TTC, Jerk, and Time Headway Across Different Traffic Volumes and MPR Levels (for clearer presentation, the direction of axis in MPR of CAV with TTC is adjusted)

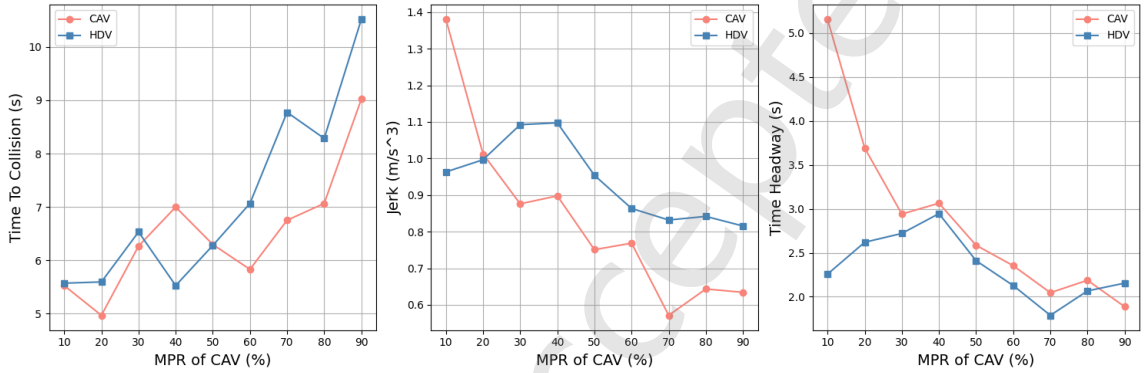


Fig. 9: Comparison of HDVs and CAVs Under Different MPR Levels

5.3 Performance Under Mixed Traffic Flow

Fig. 8 illustrates the impact of increasing the MPR of CAVs on traffic dynamics. The first subplot shows that TTC rises with higher MPR, indicating improved safety, with the effect being particularly pronounced at lower traffic volumes. The second subplot shows a decrease in jerk values as MPR increases, indicating smoother driving and fewer abrupt maneuvers, particularly under lighter traffic conditions. In the third subplot, average time headway decreases as MPR increases, suggesting more efficient utilization of road capacity. At elevated MPR levels, even under higher traffic volumes, CAVs are able to maintain closer yet safe inter-vehicle distances, resulting in increased throughput and improved traffic flow. Overall, higher MPR of CAVs significantly enhances traffic safety, comfort, and efficiency, with these benefits observable across varying traffic volumes.

Fig. 9 illustrates the impact of CAV MPR on both CAVs and HDVs across the three key metrics. In the first subplot, TTC generally increases with higher MPR for both vehicle types, indicating enhanced safety. When MPR is below 50%, the difference in TTC between CAVs and HDVs is negligible. However, at higher MPR levels, HDVs exhibit higher TTC than CAVs, suggesting that CAVs facilitate safer spacing for HDVs and help reduce collision risk. In the second subplot, jerk values decline as MPR increases, reflecting smoother driving with fewer abrupt maneuvers. Notably, CAVs maintain significantly lower jerk than HDVs, especially when MPR exceeds 30%. The third subplot shows a decrease in average time headway as MPR rises, indicating more efficient utilization of road capacity. CAVs exhibit a more pronounced reduction in headway than HDVs, highlighting their ability to maintain close yet safe following distances, thereby enhancing traffic throughput.

5.4 Ablation Study and Weight Analysis

5.4.1 Feature Importance Analysis

Fig. 10 illustrates the effects of various ablation studies on three key performance metrics. Each ablation represents a model design choice that examines the contribution of specific state features or reward components to overall policy performance. Ablation 1 (RelVel-), which removes relative velocity and acceleration features, evaluates the contribution of leading vehicle motion information to policy performance. This results in increased median TTC values but larger variance, indicating inconsistent behavior. Without relative speed awareness, the vehicle tends to overcompensate for safety, leading to inefficient or overly cautious responses. Ablation 2 (SPaT-), which excludes SPaT-related information, evaluates the importance of signal phase and timing information in the state representation. This produces a similarly wide TTC distribution and slightly degraded time headway performance, highlighting the importance of signal coordination for consistent intersection behavior. Ablation 3 (Jerk-) removes the jerk penalty from the reward function, corresponding to scenarios in which communication delays or dropped updates prevent accurate tracking of prior acceleration. As shown in the middle plot, this causes a sharp increase in both the median and variance of jerk, reflecting more abrupt acceleration and deceleration and reduced ride comfort. Ablation 4

(SlowPen-) disables the low-speed penalty, representing conditions where communication robustness is insufficient to guarantee real-time speed updates. This leads to higher variability in time headway, indicating inconsistent gap keeping and reduced efficiency.

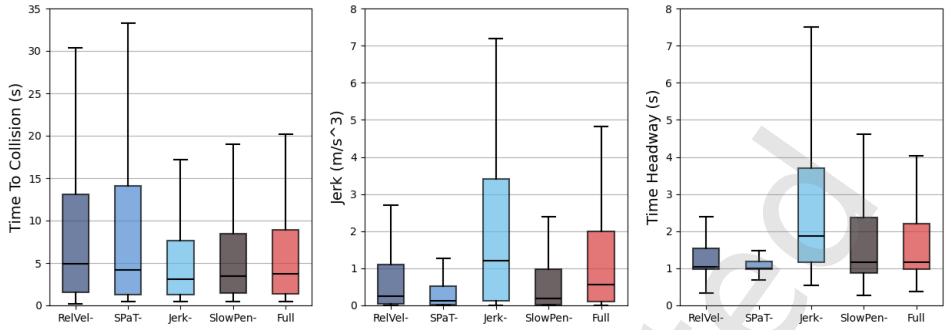


Fig. 10: Box Plots of Ablation Study Results

Fig. 11 illustrates the impact of varying reward function weights on driving behavior. Here, weight1 corresponds to the importance assigned to TTC and jerk, while weight2 corresponds to the weight for travel distance. Increasing weight1 promotes safer and smoother driving, whereas a higher weight 2 enhances efficiency but introduces instability, leading to riskier maneuvers. The most extreme behavior occurs when weight 1 is low and weight 2 is high, resulting in the lowest TTC and highest jerk, indicative of aggressive acceleration and late braking that compromise both safety and comfort. These observations highlight the critical importance of appropriately balancing reward weights to ensure stable and effective driving policies in mixed traffic scenarios.

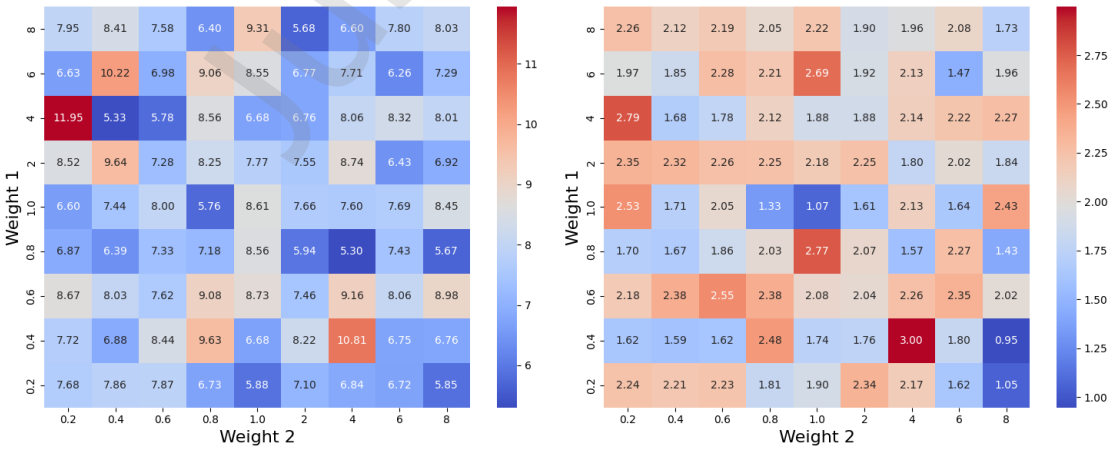


Fig. 11: Performance with Different Reward Weighting Parameters (TTC on the left and Jerk on the right)

5.4.2 Robustness Under Degraded Conditions

This subsection evaluates the robustness of the fully trained CRR policy under simulated communication failures at inference time. Two scenarios were tested: (1) V2V communication loss, where the leading vehicle features are replaced by default values ($\Delta d = 1000, \Delta v = 0, \Delta a = 0$) at inference time; and (2) V2I communication loss, where the SPaT features are replaced by default values ($t_{\text{cur}} = 0, t_{\text{next}} = 0, p_{\text{cur}} = 0$) at inference time. The results are summarized in the following table.

Table 5: Safety metrics under communication loss

Scenario	Mean TTC	Mean Jerk	Mean TH
Original CRR	7.046	0.602	1.728
CRR without SPaT	7.739	0.499	1.629
CRR without V2V	7.299	0.563	1.646

The results indicate that the CRR policy maintains stable performance under both communication failure scenarios. The findings suggest that the policy has learned sufficiently general driving strategies during training that it can operate without severe degradation when specific information channels are disrupted at deployment time.

6 Conclusion

This paper proposes an offline reinforcement learning framework based on the CRR algorithm to enhance vehicle speed guidance at signalized intersections under mixed traffic conditions. By leveraging historical trajectory and SPaT data, the method effectively addresses safety, comfort, and efficiency concerns without requiring online interaction, thereby reducing the dependence on simulated training environments and incorporating real-world driving patterns into the learned policy.

To support learning, naturalistic vehicle trajectories were preprocessed into a structured MDP format. This preprocessing involved aligning trajectory data with signal timing, computing vehicle dynamics such as speed, acceleration, jerk, and TTC, and identifying leading vehicles to extract relative motion features. Such comprehensive data preparation enabled the learning agent to perceive the complex state space of traffic scenarios. Extensive

simulations in SUMO demonstrated the superiority and stability of the CRR-based policy compared to other approaches, including rule-based GLOSA, model-based MPC, online RL TD3, offline RL BCQ, and Behavior Cloning. Quantitative results indicate that CRR significantly improves safety, comfort, and efficiency, while trajectory visualizations confirm smoother and more stable vehicle behavior across varying departure conditions. Ablation studies further highlight the critical role of both input design and reward shaping: omitting velocity or SPaT features caused unstable or inconsistent behavior, whereas removing jerk or low-speed penalties reduced comfort and following stability. Moreover, the proposed method exhibits notable computational advantages, with a per-decision inference time of 0.0097 seconds, nearly four times faster than MPC, underscoring its suitability for real-time deployment.

In summary, this study demonstrates the feasibility and advantages of applying offline reinforcement learning to vehicle speed guidance at signalized intersections within connected and automated vehicle environments. Future work will explore the integration of CRR with cooperative vehicle systems to further enhance practical applicability and facilitate broader deployment across intelligent transportation systems.

Appendix. A Model Predictive Control (MPC)

The MPC framework optimizes vehicle behavior by considering safety, comfort, and motion smoothness (Zhao et al., 2018). Optimization proceeds by minimizing a cost function combining instantaneous and terminal costs across a prediction horizon:

$$J = \int_0^{t_f} L(X(t), u(t))dt + \theta\left(X\left(t_f\right)\right) \quad (A1)$$

where J represents total cost, $L(X(t), u(t))$ captures instantaneous cost at each moment, and $\theta\left(X\left(t_f\right)\right)$ quantifies terminal cost reflecting deviation from desired final state.

Instantaneous cost incorporates safety, comfort, and travel efficiency at each timestep. Section 4 provides detailed definitions for TTC, jerk, and the IDM model. The formulation is:

$$L(X(t), u(t)) = -q_1 \cdot \text{ TTC } + q_2 \cdot \text{ Jerk } - q_3 \cdot T_{\text{dist}} \tag{A2}$$

where T_{dist} measures traveled distance per timestep, and q_1, q_2, q_3 are weighting coefficients (defaulting to unity) balancing different objectives. Terminal cost $\theta\left(X(t_f)\right)$ penalizes deviations in position, velocity, and acceleration from target values at the endpoint:

$$\theta\left(X(t_f)\right) = p_1\left(x(t_f) - \hat{x}_{t_f}\right)^2 + p_2\left(v(t_f) - \hat{v}_{t_f}\right)^2 + p_3\left(a(t_f) - \hat{a}_{t_f}\right)^2 \tag{A3}$$

Coefficients p_1, p_2, p_3 (defaulting to unity) balance positional, velocity, and acceleration errors. Target terminal values $\hat{x}_{t_f}, \hat{v}_{t_f}, \hat{a}_{t_f}$ correspond to downstream stop line position, maximum allowable velocity ($14.0\ \text{m/s}$), and zero acceleration.

The terminal horizon t_f is determined as the earliest feasible moment for green-phase intersection passage:

$$t_f = \max(t_f', t_g) \tag{A4}$$

where t_f' represents minimum travel duration computed using maximum acceleration ($2.6\ \text{m/s}^2$, matching IDM). Green signal onset time t_g within cycle k follows:

$$t_g = \begin{cases} T_k^g & \text{if } t_f' \in [T_k^g, T_k^r] \\ T_{k+1}^g & \text{if } t_f' \in [T_k^r, T_{k+1}^g] \end{cases} \tag{A5}$$

where T_k^g and T_k^r mark green and red phase initiation times for cycle k respectively.

Appendix. B Green Light Optimal Speed Advisory (GLOSA)

The GLOSA system is implemented through the SUMO glosa device settings, which enables vehicles to receive real-time traffic signal phase information and adjust their speeds accordingly. When approaching a signalized intersection, a vehicle equipped with GLOSA evaluates the remaining time before the next phase change and determines whether a speed adaptation maneuver is necessary. Specifically:

- If the vehicle is likely to arrive at a red light before it turns green, a deceleration maneuver is initiated, ensuring the vehicle reaches the stop line just as the signal changes.
- Conversely, if the vehicle is approaching a green light with a remaining phase duration shorter than its estimated arrival time, a controlled acceleration maneuver may be executed to pass through the intersection before the signal changes.

GLOSA is set to a minimum speed threshold of $5\ m/s$ for slowdown maneuvers and a maximum speed factor of 1.1 to limit excessive acceleration. These parameters ensure a balance between compliance with traffic signals and driving comfort.

Replication and data sharing

The corresponding codes of MDP dataset process flow and policy training are available at https://ets-data.sciopen.com/. Dataset ID is ETS-Data-26-00018.

Author contributions

Xiaoyu Shi: Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Yuhuan Lu:** Validation, Writing – review & editing, Supervision, Methodology. **Zihao Sheng:** Writing – review & editing, Methodology. **Jian Zhang:** Writing – review & editing, Methodology. **Sikai Chen:** Writing – review & editing, Methodology, Supervision. **Heye Huang:** Writing – review & editing, Methodology, Supervision. **Tiantian Chen:** Writing – review & editing, Supervision, Methodology, Conceptualization, Funding acquisition.

Declaration of competing interest

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

References

Bhaskar, A., Chung, E., Dumont, A.-G., 2011. Fusing loop detector and probe vehicle data to estimate travel time statistics on signalized urban networks. Comput.-Aided Civ. Infrastruct. Eng. 26, 433–450.

Dai, Q., Shen, D., Wang, J., Huang, S., Filev, D., 2022. Calibration of human driving behavior and preference using vehicle trajectory data. Transport. Res. C-Emer. 145, 103916.

Dong, H., Zhuang, W., Wu, G., Li, Z., Yin, G., Song, Z., 2024. Overtaking-enabled eco-approach control at signalized intersections for connected and automated vehicles. IEEE Trans. Intell. Transp. Syst. 25, 4527–4539.

Dong, J., Chen, S., Miralinaghi, M., Chen, T., Labi, S., 2022. Development and testing of an image transformer for explainable autonomous driving systems. J. Intell. Conn. Veh., 5, 235–249.

Fang, X., Zhang, Q., Gao, Y., Zhao, D., 2022. Offline reinforcement learning for autonomous driving with real world driving data. In: Proceedings of the IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), 3417–3422.

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, 24–29.

Guo, Q., Angah, O., Liu, Z., Ban, X., 2021. Hybrid deep reinforcement learning based eco-driving for low-level connected and automated vehicles along signalized corridors. Transport. Res. C-Emer. 124, 102980.

Güvenç, L., Uygan, I.M.C., Kahraman, K., Karaahmetoglu, R., Altay, I., Sentürk, M., et al., 2012. Cooperative adaptive cruise control implementation of team MEKAR at the grand cooperative driving challenge. IEEE Trans. Intell. Transp. Syst. 13, 1062–1074.

Han, L., et al., 2025. Safe and efficient DRL driving policies using fuzzy logic for urban lane changing scenarios. J. Intell. Conn. Veh. 8, 9210054-1–9210054-13.

Hayward, J.C., 1972. Near miss determination through use of a scale of danger. Transp. Res. Rec. 384, 24–34.

Hu, B., Li, J., 2022. A deployment-efficient energy management strategy for connected hybrid electric vehicle based on offline reinforcement learning. IEEE Trans. Ind. Electron. 69, 9644–9654.

Jeong, E., Oh, C., Lee, S., 2017. Is vehicle automation enough to prevent crashes? Role of traffic operations in automated driving environments for traffic safety. Accid. Anal. Prev. 104, 115–124.

Jiang, H., Hu, J., An, S., Wang, M., Park, B.B., 2017. Eco approaching at an isolated signalized intersection under partially connected and automated vehicles environment. Transport. Res. C-Emer. 79, 290–307.

Jiang, X., Zhang, J., Shi, X., Cheng, J., 2023. Learning the policy for mixed electric platoon control of automated and human-driven vehicles at signalized intersection: A random search approach. IEEE Trans. Intell. Transp. Syst. 24, 5131–5143.

Krajewski, R., Bock, J., Kloeker, L., Eckstein, L., 2018. The highD dataset: A drone dataset of naturalistic vehicle trajectories on German highways for validation of highly automated driving systems. In: Proceedings of the 21st International Conference on Intelligent Transportation Systems (ITSC), 2118–2125.

Lee, D., Eom, C., Kwon, M., 2024. AD4RL: Autonomous driving benchmarks for offline reinforcement learning with value-based dataset. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 8239–8245.

Li, L., Lv, Y., Wang, F., 2016. Traffic signal timing via deep reinforcement learning. IEEE/CAA J. Autom. Sin. 3, 247–254.

Li, W., Ban, X., 2019. Connected vehicles-based traffic signal timing optimization. IEEE Trans. Intell. Transp. Syst. 20, 4354–4366.

Li, X., Cui, J., An, S., Parsafard, M., 2014. Stop-and-go traffic analysis: Theoretical properties, environmental impacts and oscillation mitigation. Transport. Res. B-Meth. 70, 319–339.

Li, Y., Li, Z., Wang, H., Wang, W., Xing, L., 2017. Evaluating the safety impact of adaptive cruise control in traffic oscillations on freeways. Accid. Anal. Prev. 104, 137–145.

linLiu, Y., Wu, F., Liu, Z., Wang, K., Wang, F., Qu, X., 2023. Can language models be used for real-world urban-delivery route optimization?. The Innovation, 4, 100520.

Lu, Y., Zhang, Z., Wang, W., Zhu, Y., Chen, T., Al-Otaibi, Y.D., et al., 2025. Knowledge-driven lane change prediction for secure and reliable Internet of Vehicles. IEEE Trans. Intell. Transp. Syst., 26, 14120–14131.

Ma, F., Yang, Y., 2021. Eco-driving-based cooperative adaptive cruise control of connected vehicles platoon at signalized intersections. Transport. Res. D-Tr. E. 92, 102746.

Milakis, D., van Arem, B., van Wee, B., 2017. Policy and society related implications of automated driving: A review of literature and directions for future research. J. Intell. Transp. Syst. 21, 324–348.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., et al., 2015. Human-level control through deep reinforcement learning. Nature 518, 529–533.

Moosavi, S., Samavatian, M.H., Parthasarathy, S., Ramnath, R., 2019. A countrywide traffic accident dataset. https://doi.org/10.48550/arXiv.1906.05409.

Mousavi, S.S., Schukat, M., Howley, E., 2017. Traffic light control using deep policy-gradient and value-function-based reinforcement learning. IET Intell. Transp. Syst. 11, 417–423.

Qi, X., Wu, G., Boriboonsomsin, K., Barth, M.J., Gonder, J., 2016. Data-driven reinforcement learning-based real-time energy management system for plug-in hybrid electric vehicles. Transp. Res. Rec. 2572, 1–8.

Rakha, H., Kamalanathsharma, R.K., 2011. Eco-driving at signalized intersections using V2I communication. In: Proceedings of the 14th International IEEE Conference on Intelligent Transportation Systems (ITSC), 341–346.

Shi, J., Qiao, F., Li, Q., Yu, L., Hu, Y., 2018. Application and evaluation of the reinforcement learning approach to eco-driving at intersections under infrastructure-to-vehicle communications. Transp. Res. Rec. 2672, 89–98.

Shi, X., Zhang, J., Jiang, X., Chen, J., Hao, W., Wang, B., 2024. Learning eco-driving strategies from human driving trajectories. Physica A 633, 129353.

Shi, X., Chen, S., Jang, K., Chen, T., 2025. Conflict-aware platoon scheduling at unsignalised intersections via traveling salesman problem: A large language model-guided heuristic solution. In: 2025 IEEE 28th International Conference on Intelligent Transportation Systems (ITSC), 1718–1723.

Shuvo, S., Mukherjee, S., Chatterjee, S., Glavaski, S., Vrabie, D., Canayon, G., et al., 2024. Deep multi-agent reinforcement learning for real-world signalized traffic corridor control. In: Proceedings of the International Conference on Machine Learning and Applications (ICMLA), 644–651.

Soteropoulos, A., Berger, M., Ciari, F., 2019. Impacts of automated vehicles on travel behaviour and land use: An international review of modelling studies. Transp. Rev. 39, 29–49.

Sun, C., Guanetti, J., Borrelli, F., Moura, S.J., 2020. Optimal eco-driving control of connected and autonomous vehicles through signalized intersections. IEEE Internet Things J. 7, 3759–3773.

Treiber, M., Hennecke, A., Helbing, D., 2000. Congested traffic states in empirical observations and microscopic simulations. Phys. Rev. E 62, 1805–1824.

U.S. Department of Transportation, n.d. NGSIM. http://www.ngsim.fhwa.dot.gov.

Wang, H., Peng, P., Huang, Y., Tang, X., 2018. Model predictive control-based eco-driving strategy for CAV. IET Intell. Transp. Syst. 13, 323–329.

Wang, Z., Novikov, A., Zolna, K., Merel, J.S., Springenberg, J.T., Reed, S.E., et al., 2020. Critic regularized regression. Adv. Neural Inf. Process. Syst. 33, 7768–7778.

Wang, Z., Zheng, O., Li, L., Abdel-Aty, M., Cruz-Neira, C., Islam, Z., 2023. Towards next generation of pedestrian and connected vehicle in-the-loop research: A digital twin co-simulation framework. IEEE Trans. Intell. Veh. 8, 2674–2683.

Wei, L., Chen, P., Mei, Y., Wang, Y., 2022. Turn-level network traffic bottleneck identification using vehicle trajectory data. Transport. Res. C-Emer. 140, 103707.

Wu, R., Jia, H., Huang, Q., Tian, J., Gao, H., Wang, G., 2024. Multi-lane unsignalized intersection cooperation strategy considering platoons formation in a mixed connected automated vehicles and connected human-driven vehicles environment. IEEE Trans. Intell. Transp. Syst. 25, 1569–1585.

Wu, X., He, X., Yu, G., Harmandayan, A., Wang, Y., 2015. Energy-optimal speed control for electric vehicles on signalized arterials. IEEE Trans. Intell. Transp. Syst. 16, 2786–2796.

Xia, H., Wu, G., Boriboonsomsin, K., Barth, M.J., 2013. Development and evaluation of an enhanced eco-approach traffic signal application for connected vehicles. In: Proceedings of the 16th International IEEE Conference on Intelligent Transportation Systems (ITSC), 296–301.

Xu, T., Pang, Y., Zhu, Y., Ji, W., Jiang, R., 2025. Real-time driving style integration in deep reinforcement learning for traffic signal control. IEEE Trans. Intell. Transp. Syst. 26, 11879–11892.

Xu, Z., Chen, T., Huang, Z., Xing, Y., Chen, S., 2025. Personalizing driver agent using large language models for driving safety and smarter human–machine interactions. IEEE Intell. Transp. Syst. Mag., 17, 96–111.

Yang, D., Wu, Y., Sun, F., Chen, J., Zhai, D., Fu, C., 2021. Freeway accident detection and classification based on the multi-vehicle trajectory data and deep learning model. Transport. Res. C-Emer. 130, 103303.

Yang, H., Rakha, H., Ala, M.V., 2017. Eco-cooperative adaptive cruise control at signalized intersections considering queue effects. IEEE Trans. Intell. Transp. Syst. 18, 1575–1585.

Yao, H., Cui, J., Li, X., Wang, Y., An, S., 2018. A trajectory smoothing method at signalized intersection based on individualized variable speed limits with location optimization. Transport. Res. D-Tr. E. 62, 456–473.

Yong, J.Y., Ramachandaramurthy, V.K., Tan, K.M., Mithulananthan, N., 2015. A review on the state-of-the-art technologies of electric vehicle, its impacts and prospects. Renew. Sust. Energ. Rev. 49, 365–385.

Zhang, J., Jiang, X., Cui, S., Yang, C., Ran, B., 2022. Navigating electric vehicles along a signalized corridor via reinforcement learning: Toward adaptive eco-driving control. Transp. Res. Rec. 2676, 657–669.

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, 291–312.

Zhang, X., Sun, J., Qi, X., Sun, J., 2019. Simultaneous modeling of car-following and lane-changing behaviors using deep learning. Transport. Res. C-Emer. 104, 287–304.

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. Transport. Res. C-Emer. 95, 802–821.

Zheng, O., Abdel-Aty, M., Yue, L., Abdelraouf, A., Wang, Z., Mahmoud, N., 2024. CitySim: A drone-based vehicle trajectory dataset for safety-oriented research and digital twins. Transp. Res. Rec. 2678, 606–621.

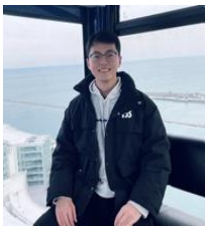
Zhou, M., Yu, Y., Qu, X., 2020. Development of an efficient driving strategy for connected and automated vehicles at signalized intersections: A reinforcement learning approach. IEEE Trans. Intell. Transp. Syst. 21, 433–443.

Zhu, M., Wang, Y., Pu, Z., Hu, J., Wang, X., Ke, R., 2020. Safe, efficient, and comfortable velocity control based on reinforcement learning for autonomous driving. Transport. Res. C-Emer. 117, 102662.

Author biography

Xiaoyu Shi received the B.Sc. and M.Sc. degrees in transportation engineering from Southeast University. She is currently pursuing the Ph.D. degree with the Cho Chun Shik Graduate School of Mobility, Korea Advanced Institute of Science and Technology (KAIST). Her main research interests include reinforcement learning, autonomous vehicle control at intersections.

Yuhuan Lu received the B.Sc. and M.Sc. degrees in transportation engineering from Sun Yat-sen University. He received Ph.D. degree in computer science from the State Key Laboratory of Internet of Things for Smart City (University of Macau). He is also affiliated with the Department of Computer and Information Science at the University of Macau. His research interests lie in knowledge representation, intelligent transportation systems, and autonomous driving.



Zihao Sheng received the B.S. degree in automation from Xi'an Jiaotong University, Xi'an, China, in 2019, and the M.S. degree in control engineering from Shanghai Jiao Tong University, Shanghai, China, in 2022. He is currently pursuing a Ph.D. degree with the Department of Civil and Environmental Engineering, University of Wisconsin–Madison, USA. His main research interests include human-centered AI, autonomous driving, reinforcement learning, and intelligent transportation.

Prof Jian Zhang received the Ph.D. degree from Southeast University, Nanjing, China, in 2011. He is currently the Vice Director of the Research Center for Internet of Mobility, Southeast University. His research interests include transportation applications of mobile phone data, connected vehicles, and the public transportation system. He is also a member of the American Society of Civil Engineers.

Sikai (Sky) Chen received the Ph.D. degree in civil engineering with a focus on computational science and engineering from Purdue University in 2019. He is an Assistant Professor with the Department of Civil and Environmental Engineering, Department of Mechanical Engineering (courtesy), University of Wisconsin–Madison. His research centers around three major themes, such as human users, AI, and transportation. He aims to innovate and develop safe, efficient, sustainable, and human-centered transportation systems using cutting-edge methods and technologies.

Heye Huang received the bachelor's degree in traffic and transportation engineering from Central South University, Changsha, China, in 2018. She received Ph.D. degree from the State Key Laboratory of Automotive Safety and Energy, Tsinghua University, Beijing, China. Her current research interests include connected and automated vehicles, risk assessment, and decision making.

Tiantian Chen received the Ph.D. degree in Civil and Environmental Engineering from the Hong Kong Polytechnic University. She is now an Associate Professor at the Cho Chun Shik Graduate School of Mobility, Korea Advanced Institute of Science and Technology (KAIST). Her research interests include traffic safety, driving simulation, travel behavior, and human-machine interaction.