Offline Multi-Agent Reinforcement Learning with a Physics-Informed World Model for Cooperative Mixed Traffic Control

Lu Liu
College of Transportation, Tongji University
Shanghai, China, 201804
Email: luliu0720@tongji.edu.cn

Chi Xie
College of Transportation, Tongji University
Shanghai, China, 201804
Email: chi.xie@tongji.edu.cn

Xi Xiong*
College of Transportation, Tongji University
Shanghai, China, 201804
Email: xi_xiong@tongji.edu.cn

Total Number of Pages: 20

Submitted: August 18, 2026
*Corresponding author

ABSTRACT

This study investigates cooperative control of connected and automated vehicles (CAVs) at partially observable highway bottlenecks in mixed traffic, aiming to mitigate congestion without relying on complete global traffic states or online trial-and-error. We propose a physics-informed world model-based offline multi-agent reinforcement learning framework that reconstructs a physically interpretable global traffic state from local CAV observation-action histories, with coupled macroscopic-microscopic traffic dynamics providing physics-based supervision. A probabilistic ensemble world model learns traffic-state transitions and system rewards, while model disagreement quantifies epistemic uncertainty. Multi-step imagined rollouts with pessimistic rewards and uncertainty-driven truncation are then used for offline policy learning. Experiments in a SUMO-based on-ramp bottleneck using approximately 1×10^6 offline transitions show that physics supervision improves state reconstruction and world-model prediction accuracy.

INTRODUCTION

With the rapid development of automated driving and vehicular networking technologies, connected and automated vehicles (CAVs) are evolving from passive traffic participants into active control agents. Unlike conventional traffic control strategies that rely on roadside infrastructure, CAVs can directly regulate surrounding driving behavior and local traffic evolution through their own control actions, providing a new approach for link-level traffic flow control (Lu et al. 2022). This capability is particularly important for highway bottlenecks, such as on-ramp merging and lane-drop areas, where intensified vehicle interactions and capacity variations can amplify local speed disturbances, leading to queue formation and capacity degradation (Chung et al. 2007). Although recent studies have explored cooperative CAV control to improve traffic efficiency, most of them assume that the global traffic state is available or that the observations of all CAVs can collectively cover the entire traffic system. This assumption is difficult to satisfy at highway bottlenecks, where CAVs are constrained by limited sensing and communication ranges and can only observe partial surrounding traffic conditions.

To enable effective coordination among multiple CAVs for mitigating traffic congestion, this paper proposes a physics-informed world model-based offline multi-agent reinforcement learning framework for mixed traffic at bottlenecks with limited global visibility. As shown in Fig. 1, we consider a typical on-ramp merging scenario where human-driven vehicles (HDVs) and CAVs coexist. The blue shaded area denotes the CAV control zone upstream of the merging point. Once entering this zone, each CAV obtains local traffic information within a limited range through onboard sensors and vehicle-to-everything (V2X) communication. Based on these partial observations and historical interactions, CAVs execute longitudinal acceleration control to actively regulate upstream traffic flow. In contrast to fully automated traffic, mixed traffic involves highly stochastic car-following and merging behaviors of HDVs, making the traffic response to CAV control difficult to predict (Xiong et al. 2024). Therefore, online trial-and-error learning on real roads may introduce potential safety risks. To address this issue, this paper focuses on learning a world model from historical observation-action data of CAVs to capture bottleneck traffic evolution and support reliable CAVs cooperative control.

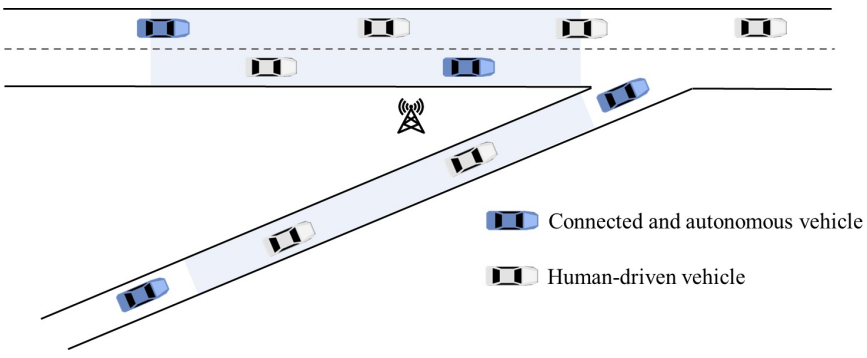


Figure 1 Mixed-traffic on-ramp bottleneck with the CAV control zone

Early efforts in CAV control were dominated by rule-based approaches, which designed control policies based on vehicle spacing, relative speed, and leading-vehicle motion information (Liao et al. 2022). To extend CAV control from individual vehicle regulation to traffic-level coordination, many studies have combined macroscopic-microscopic coupled models with model

predictive control (Qiu and Du 2023). However, the performance of these methods depends on the modeling accuracy of driving behaviors and traffic demand variations (Li et al. 2020). Moreover, as the number of CAVs increases, the optimization scale grows rapidly, leading to substantial communication and computational burdens (Feng et al. 2021).

Data-driven methods have therefore attracted increasing attention, among which reinforcement learning (RL) has shown strong potential for cooperative CAV control by learning from interaction feedback (Wang et al. 2023; Li et al. 2023). Online policy exploration in real traffic systems may induce unsafe driving behaviors and incur prohibitive costs (Cang et al. 2021). Offline RL, which learns from pre-collected datasets without continual environment interaction during training, has therefore emerged as a promising direction. Nevertheless, offline RL suffers from distribution shift, as poorly covered states or actions may lead to suboptimal or unsafe decisions (Ran et al. 2023). For example, Fang et al. deployed offline RL using real-world driving data and found that limited data coverage led to unsatisfactory CAV control performance (Fang et al. 2022). Although policy constraints and value regularization have been widely used to improve policy reliability (Huang et al. 2025), CAVs coordination introduces additional challenges: even when each agent selects an in-distribution action, the resulting joint action may still be out of distribution. Moreover, without interactive feedback, effective multi-agent coordination is difficult to guarantee (Zhou et al. 2025). As a result, offline RL has seen limited application in CAV control.

World model-based offline RL offers a promising way to address these limitations. By learning environment dynamics and generating imagined experience, these methods can improve sample efficiency and enhance lookahead prediction capability (Ross and Bagnell 2012). To mitigate model exploitation in offline settings, existing studies often incorporate uncertainty estimation, pessimistic rewards, or penalties for uncertain regions (Wiesemann et al. 2013). However, most world model methods are designed for general control tasks or single-vehicle decision-making and typically rely on relatively complete environment states (Kang et al. 2025). Such an assumption is often difficult to satisfy in practice because of limited sensing and communication capabilities. Therefore, some studies directly concatenate the local observations of multiple CAVs to construct an approximate global state for centralized training (Rauch et al. 2024). However, such representations may introduce substantial information redundancy, and their dimensionality increases with the number of CAVs. Other studies learn low-dimensional latent states to obtain more compact system representations (Hafner et al. 2019). Although latent representations can reduce the modeling complexity, they typically lack explicit traffic-physical semantics, potentially leading to predictions that are inconsistent with actual traffic evolution. In addition, multi-agent systems involve larger joint state-action spaces, making it difficult for finite offline datasets to sufficiently cover the state-action regions that may be visited by the learned policy (Bui et al. 2024). Under such distribution shifts, the policy may exploit prediction errors of the world model in poorly covered regions, thereby generating unreliable imagined experiences (Kidambi et al. 2020).

Based on the above analysis, applying world models to cooperative multi-CAV control in partially observable mixed traffic still faces three interrelated challenges: how to construct an effective global state that captures overall traffic evolution from distributed local observations, how to improve the consistency of the learned world model with traffic-flow physics, and how to mitigate the impact of model prediction errors on imagined policy learning under limited offline data coverage.

This paper proposes a physics-informed world model-based offline multi-agent reinforcement learning framework for mixed-traffic bottlenecks. First, temporal traffic information is ex-

tracted from the local observation-action histories of multiple CAVs, and a structured global traffic state consisting of traffic density, average speed, and microscopic CAV states is reconstructed through traffic-cell-based feature aggregation. Unlike directly concatenating local observations or learning latent states without explicit physical semantics, the proposed method incorporates a coupled macroscopic-microscopic traffic model to provide physics-based consistency supervision for state reconstruction, enabling the reconstructed state to explain local observations while better reflecting traffic-flow evolution. Based on this reconstructed state, a probabilistic ensemble world model is developed to learn traffic-state transitions and system rewards under joint CAV control, while epistemic uncertainty is quantified through prediction disagreement among ensemble members. Multi-step imagined rollouts are then performed in the learned traffic world model, together with uncertainty-aware pessimistic rewards and rollout truncation, to reduce the risk of exploiting unreliable model predictions in poorly covered regions. Finally, cooperative multi-CAV policies are learned under the centralized training with decentralized execution (CTDE) paradigm, allowing each CAV to make decisions based solely on its own local observation history during execution, without online trial-and-error or access to the true global traffic state. A mixed-traffic on-ramp scenario is constructed using simulation of urban mobility (SUMO) to evaluate the proposed method in terms of world-model prediction accuracy, cooperative control performance, and sensitivity to different traffic demands and CAV penetration rates.

The remainder of this paper is organized as follows. Section 2 introduces the coupled macroscopic-microscopic traffic model and the Dec-POMDP formulation. Section 3 presents the proposed physics-informed world model-based offline multi-agent reinforcement learning method. Section 4 describes the experimental setup and analyzes the results. Finally, Section 5 concludes the paper.

PRELIMINARIES

Decentralized POMDP Formulation

The cooperative control problem of multiple CAVs in mixed-autonomy traffic is formulated as a decentralized partially observable Markov decision process (Dec-POMDP) (Bernstein et al. 2002), represented by the tuple $\langle \mathcal{N}, S, A, P, R, \Omega, O, \gamma \rangle$. $\mathcal{N} = \{1, 2, \ldots, N\}$ denotes the set of CAVs operating within the coordination region. The road segment within the coordination region is partitioned into M homogeneous cells with length Δx, indexed by $\{J\} = \{1, 2, \ldots, M\}$. The global state space is denoted by S, and at time t, the system state $s(t) \in S$ characterizes the underlying traffic state of the mixed-autonomy system. Specifically, the global state is defined as $s(t) = (\rho(t), \bar{v}(t), e(t))$, where $\rho(t) = \{\rho_1(t), \rho_2(t), \ldots, \rho_M(t)\}$ and $\bar{v}(t) = \{\bar{v}_1(t), \bar{v}_2(t), \ldots, \bar{v}_M(t)\}$ denote the traffic density and average speed of the M road cells, respectively. The term $e(t)$ contains the microscopic states of all CAVs, including their positions and velocities.

The joint action space is defined as $A = \prod_{i \in \mathcal{N}} A_i$, where A_i is the action space of CAV i. In this study, the action of each CAV is represented by a continuous longitudinal acceleration command $a_i(t) \in A_i = [a_{\min}, a_{\max}]$, where $a_{\min}$ and $a_{\max}$ denote the minimum and maximum admissible accelerations. The actions of all CAVs jointly form the action vector $a(t) = \{a_i(t)\}_{i \in \mathcal{N}}$. The stochastic evolution of the traffic system is characterized by the transition function $P : S \times A \times S \to [0, 1]$, which captures the combined effects of traffic-flow propagation, CAV control actions, and stochastic human-driving behaviors.

Let $\Omega = \prod_{i \in \mathcal{N}} \Omega_i$ denote the joint observation space, and let $O = \{O_i\}_{i \in \mathcal{N}}$ denote the set

of local observation functions. For each CAV i, the observation function maps the global traffic state to its local observation, $o_i(t) = O_i(s(t))$. The local observation of CAV i is defined as $o_i(t) = \langle x_i(t), v_i(t), \varepsilon_i^{\text{surr}}(t)\rangle$, where $x_i(t)$ and $v_i(t)$ denote the position and velocity of CAV i, respectively, and $\varepsilon_i^{\text{surr}}(t)$ represents the surrounding traffic information perceived within its sensing or communication range. The observations of all CAVs constitute the joint observation vector $o(t) = \{o_i(t)\}_{i\in\mathcal{N}}$. The parameter $\gamma \in [0, 1)$ is the discount factor.

All CAVs share a common team reward function $R : S \times A \to \mathbb{R}$. The immediate reward is designed to improve system-level traffic efficiency while suppressing excessive control actions,

$$r(t) = R(s(t), a(t)) = w_1 \sum_{j=1}^{M} \rho_j(t)\bar{v}_j(t) - w_2 \sum_{i\in\mathcal{N}} a_i^2(t), \tag{1}$$

where the first term approximates the aggregated traffic flow over all road cells, and the second term penalizes aggressive acceleration or deceleration. The parameters $w_1 > 0$ and $w_2 > 0$ balance traffic efficiency and driving smoothness.

At each decision step t, CAV i does not have access to the complete global state $s(t)$. Instead, it selects its control action based on its local interaction history $\tau_i(t) = (o_i(t-H : t), a_i(t-H : t-1))$, according to a decentralized policy, $a_i(t) \sim \pi_i(\cdot \mid \tau_i(t))$. After the joint action $a(t)$ is executed, the traffic system evolves to the next state $s(t + 1)$ according to $P(s(t + 1) \mid s(t), a(t))$, and a team reward $r(t) = R(s(t), a(t))$ is generated. The objective is to learn the optimal decentralized policies $\pi^* = \{\pi_i^*\}_{i\in\mathcal{N}}$ that maximize the expected cumulative discounted return,

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(t)\right]. \tag{2}$$

Traffic Flow Dynamics Prior

To ensure the physical consistency of state inference and traffic evolution prediction, a macroscopic-microscopic coupled traffic flow model is introduced as a physics-informed prior.

At the macroscopic level, the mainstream traffic dynamics are described by the Lighthill-Whitham-Richards (LWR) conservation law (Richards 1956),

$$\partial_t \rho(x, t) + \partial_x q(x, t) = 0, \tag{3}$$

$$q(x, t) = Q(\rho(x, t)) = \rho(x, t)v(x, t), \tag{4}$$

where $\rho(x, t)$, $v(x, t)$, and $q(x, t)$ denote the traffic density, traffic speed, and traffic flow at location x and time t, respectively. The function $Q(\rho)$ denotes the traffic flow fundamental diagram. In this study, a triangular fundamental diagram (Newell 1993) is adopted to characterize the relationship between traffic density and flow. The corresponding equilibrium speed is given by,

$$v(x, t) = V(\rho(x, t)) = \frac{Q(\rho(x, t))}{\rho(x, t)}. \tag{5}$$

For computational implementation, the continuous traffic-flow model is discretized over the cell partition $\mathcal{J}$ using the cell transmission model (CTM) (Daganzo 1994). For each cell $j \in \mathcal{J}$, the density evolution is given by,

$$\rho_j(t + 1) = \rho_j(t) + \frac{\Delta t}{\Delta x}\left(q_{j-1}(t) - q_j(t) + r_j^{\text{on}}(t) - r_j^{\text{off}}(t)\right), \tag{6}$$

where $q_{j-1}(t)$ and $q_j(t)$ denote the upstream inflow and downstream outflow of cell j, respectively. The terms $r_j^{\text{on}}(t)$ and $r_j^{\text{off}}(t)$ represent the on-ramp inflow and off-ramp outflow. The boundary flow between two adjacent cells is determined by the demand and supply functions (Jin 2012),

$$q_j(t) = \min\{D_j(t), S_{j+1}(t)\}, \tag{7}$$

$$D_j(t) = \min\{V_f \rho_j(t), q_{\max}\}, \tag{8}$$

$$S_{j+1}(t) = \min\{w(\rho_{\max} - \rho_{j+1}(t)), q_{\max}\}, \tag{9}$$

where, V_f denotes the free-flow speed, w denotes the backward wave speed, $\rho_{\max}$ denotes the jam density, and $q_{\max}$ denotes the maximum flow capacity.

Accordingly, the cell physical speed prior is given by

$$\bar{v}_j(t+1) = V(\rho_j(t+1)). \tag{10}$$

At the microscopic level, the motion of each CAV is described by a discrete-time kinematic model,

$$v_i(t+1) = \text{clip}\left(v_i(t) + a_i(t)\Delta t, 0, V_{\max}\right), \tag{11}$$

$$x_i(t+1) = x_i(t) + v_i(t)\Delta t, \tag{12}$$

where $x_i(t)$, $v_i(t)$, and $a_i(t)$ denote the position, velocity, and acceleration command of CAV i, respectively. The operator $\text{clip}(\cdot)$ ensures that the updated speed remains within the admissible speed range.

To capture the impact of CAV control on local traffic flow dynamics, CAVs are modeled as controllable moving disturbances embedded in the macroscopic traffic stream. Their longitudinal control actions affect local car-following behavior, gap formation, and traffic flow propagation. Accordingly, a CAV-induced flow correction term is introduced into the CTM boundary flow,

$$q_j(t) = \min\{D_j(t), S_{j+1}(t)\} + \Delta q_j^{\text{CAV}}(t), \tag{13}$$

where $\Delta q_j^{\text{CAV}}(t)$ represents the aggregate impact of nearby CAVs. A positive value of $\Delta q_j^{\text{CAV}}(t)$ indicates an improvement in local throughput induced by cooperative CAV control, whereas a negative value indicates reduced local traffic throughput caused by traffic disturbances. The flow correction term is computed according to the moving-bottleneck Riemann formulation in the literature (Čičić and Johansson 2018).

By combining the macroscopic traffic evolution and microscopic CAV dynamics, the physics-based state transition can be compactly written as

$$\tilde{s}(t+1) = F_{\text{phy}}(s(t), a(t)). \tag{14}$$

This physics-based transition serves as a structural prior rather than a complete replacement for the learned world model. It provides physical guidance for global state inference and future traffic evolution prediction, ensuring that the learned model remains consistent with traffic conservation principles.

METHODOLOGY

Framework Overview

In this section, we propose a physics-informed world model-based offline multi-agent reinforcement learning framework for cooperative control of multiple CAVs in mixed traffic. As illustrated in Fig. 2, the proposed framework consists of two main stages.

In the first stage, a traffic world model is trained using offline interaction data. By integrating traffic flow physics priors and epistemic uncertainty, the learned world model recovers physically meaningful global traffic states from local observations and predicts both traffic state evolution and system-level rewards. In the second stage, world model-based imagination rollout and policy learning are conducted. Starting from historical states sampled from the offline dataset, the policy performs multi-step virtual interactions within the learned world model to generate additional imagined trajectories. These trajectories are then used to optimize cooperative multi-agent control policies under the centralized training with decentralized execution (CTDE) paradigm.

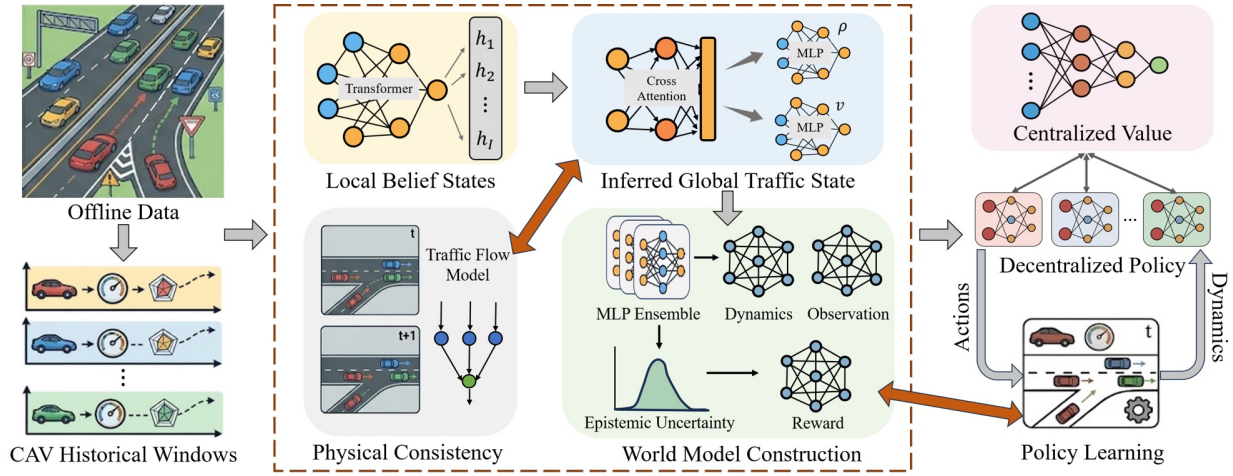


Figure 2 A traffic world model learning framework integrating traffic flow physics priors and epistemic uncertainty

Physics-Informed Global Traffic State Inference

In this subsection, we develop a physically meaningful global traffic state inference module that recovers global traffic states from multi-vehicle historical observations and actions, thereby providing a state foundation for world model learning and centralized policy training.

For each CAV i located within the cooperative control region at time t, the local interaction history $\tau_i(t)$ is used to characterize its recent motion, surrounding traffic evolution, and control response information. Then, a Transformer-based history encoder Enc_ϕ is employed to extract the local belief representation of each CAV,

$$h_i(t) = \mathrm{Enc}_\phi\left(\tau_i(t)\right), \tag{15}$$

where ϕ denotes the learnable parameters. In this paper, all CAVs share the same encoder parameters, which enables the model to accommodate dynamic traffic scenarios where vehicles may enter or leave the control region and improves its generalization capability under different traffic demands and CAV penetration rates.

By aggregating the local beliefs of all CAVs, the proposed module further recovers a global traffic representation according to the spatial structure of the road. Specifically, each road cell is regarded as a spatial query unit. For each cell $j \in \mathcal{J}$, a cross-attention mechanism is used to

adaptively aggregate the information relevant to that cell from the local beliefs of all CAVs,

$$\alpha_{i,j}(t) = \frac{\exp\left(\frac{(\eta_j W_Q)(h_i(t)W_K)^\top}{\sqrt{d}}\right)}{\sum_{i'=1}^{N}\exp\left(\frac{(\eta_j W_Q)(h_{i'}(t)W_K)^\top}{\sqrt{d}}\right)}, \tag{16}$$

$$z_j(t) = \sum_{i=1}^{N}\alpha_{i,j}(t)\,(h_i(t)W_V)\,, \tag{17}$$

where η_j is the learnable spatial query vector associated with cell j, and W_Q, W_K, and W_V are learnable projection matrices. The attention weight $\alpha_{i,j}(t)$ quantifies the degree to which cell j extracts information from CAV i, while $z_j(t)$ represents the inferred state of cell j. The global traffic representation is denoted as $Z(t) = [z_1(t), \ldots, z_M(t)]$.

To enable the world model to learn traffic evolution dynamics in a state space that is consistent with traffic flow mechanisms, we further employ multilayer perceptrons (MLPs) to map the traffic representations into macroscopic traffic variables with clear physical meanings. These variables include cell density and average speed, which are essential for characterizing congestion propagation and flow variations,

$$\hat{\rho}_j(t) = \mathrm{MLP}_\rho\left(z_j(t)\right), \quad \hat{v}_j(t) = \mathrm{MLP}_v\left(z_j(t)\right). \tag{18}$$

The recovered density and average speed over the entire road segment are represented as $\hat{\rho}(t) = (\hat{\rho}_1(t), \ldots, \hat{\rho}_M(t))$ and $\hat{v}(t) = (\hat{v}_1(t), \ldots, \hat{v}_M(t))$, respectively. The density prediction head uses a Softplus activation function to ensure non-negativity, while the speed prediction head uses a Sigmoid mapping scaled to the free-flow speed range $[0, V]$. In addition to macroscopic traffic states, we extract the individual state representation $e_i(t)$ of each CAV from its local observation $o_i(t)$, and denote the individual states of all CAVs as $e(t) = (e_1(t), \ldots, e_N(t))$. The recovered global traffic state is then constructed as $\hat{s}(t) = [\hat{\rho}(t), \hat{v}(t), e(t)]$.

To ensure the validity and interpretability of the inference module, an observation reconstruction constraint is introduced. Specifically, an observation decoder O_ψ is constructed to map the recovered global state back to the corresponding local observation of each CAV i,

$$\hat{o}_i(t) = O_\psi\left(\hat{s}(t), i\right), \tag{19}$$

where ψ denotes the parameters of the observation decoder. The observation reconstruction loss is defined as,

$$\mathcal{L}_{\mathrm{obs}} = \sum_{i=1}^{N}\left\|\hat{o}_i(t) - o_i(t)\right\|^2, \tag{20}$$

where $o_i(t)$ is the truth observation from the offline dataset. This loss encourages the recovered global state to explain the local observations of all CAVs. Moreover, after training, the observation decoder is used in the subsequent imagination rollout process, providing inputs for decentralized policy execution.

Relying solely on observation reconstruction may allow the recovered state to numerically explain local observations, but it does not necessarily guarantee consistency with real traffic flow evolution. Therefore, a traffic flow physics model is introduced as a soft constraint. Given the recovered state $\hat{s}(t)$ and the joint control action $a(t)$, the traffic flow physics model $F_{\mathrm{phy}}(\cdot)$ provides

a physics-based reference state for the next time step,

$$s^{\text{phy}}(t+1) = F_{\text{phy}}\left(\hat{s}(t), a(t)\right). \tag{21}$$

The physical consistency loss is defined as,

$$\mathcal{L}_{\text{phy}} = \left\|\hat{s}(t+1) - s^{\text{phy}}(t+1)\right\|^2. \tag{22}$$

It should be noted that the physics model is not treated as the state transition function in this paper. Instead, it is used as a traffic flow prior to regularize the state inference process. This design constrains the learned state within a physically reasonable range defined by traffic conservation laws and fundamental diagram relationships, thereby improving the interpretability of state recovery and the stability of the constructed world model.

Overall, the training objective of the global traffic state inference module is formulated as,

$$\mathcal{L}_{\text{state}} = \lambda_1 \mathcal{L}_{\text{obs}} + \lambda_2 \mathcal{L}_{\text{phy}}, \tag{23}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are the weighting coefficients.

Probabilistic Ensemble Traffic World Model

This subsection focuses on learning the evolution dynamics of the traffic system in the state–action space. In mixed traffic, future traffic states are uncertain due to the stochastic behavior of human-driven vehicles (HDVs) and the control actions of CAVs. Meanwhile, offline datasets cannot cover all possible state–action combinations, and a single predictive model may suffer from large errors in sparsely sampled regions. To address these issues, we develop a probabilistic ensemble traffic world model to jointly predict state transitions, estimate system-level rewards, and quantify epistemic uncertainty.

Specifically, we initialize K MLPs with the same network architecture but independent parameters, denoted as $\mathcal{M}\theta_k k = 1^K$. Each MLP k takes the recovered global traffic state $\hat{s}(t)$ and the joint action $a(t)$ as inputs, and predicts the distribution of the next global traffic state,

$$\hat{s}(t+1) \sim \mathcal{N}\left(\mu_{\theta_k}\left(\hat{s}(t), a(t)\right), \text{diag}\left(\sigma^2_{\theta_k}\left(\hat{s}(t), a(t)\right)\right)\right), \tag{24}$$

where $\mu_{\theta_k}\left(\hat{s}(t), a(t)\right) \in \mathbb{R}^{d_s}$ and $\sigma_{\theta_k}\left(\hat{s}(t), a(t)\right) \in \mathbb{R}^{d_s}_{>0}$ denote the mean vector and standard deviation vector of the predicted state transition distribution, respectively. d_s is the dimension of the global traffic state vector. Meanwhile, the MLP k also outputs the corresponding immediate reward prediction,

$$R_{\theta_k}\left(\hat{s}(t), a(t), \hat{s}(t+1)\right), \tag{25}$$

where R_{θ_k} denotes the reward prediction head.

To enhance the diversity among different MLPs and prevent all models from converging to similar solutions, which may invalidate uncertainty estimation, we introduce a bootstrap-based data perturbation mechanism. At each parameter update, a mini-batch of size B, denoted as $\left\{\left(o^{(b)}(t), a^{(b)}(t), o^{(b)}(t+1)\right)\right\}_{b=1}^{B}$, is randomly sampled from the offline dataset $\mathcal{D}$, and the corresponding global traffic states are inferred using the state inference module, where b denotes the sample index in the mini-batch. For the MLP k and the training sample b, an independent Bernoulli

mask is sampled as,

$$m_k^{(b)} \sim \text{Bern}(p), \quad b = 1, \ldots, B, \quad k = 1, \ldots, K, \tag{26}$$

where p denotes the sample retention probability. When $m_k^{(b)} = 1$, the MLP k uses this sample for parameter updating; otherwise, when $m_k^{(b)} = 0$, this sample is excluded from gradient computation for the MLP k. This mechanism allows different MLPs to observe different subsets of data during each training iteration, thereby inducing diverse generalization behaviors in the learned state transition functions.

For learning the probabilistic state transition distribution of each MLP, we adopt the negative log-likelihood (NLL) loss as the training objective. Let the target state be $y = \text{sg}(\hat{s}(t+1))$, where $\text{sg}(\cdot)$ denotes the stop-gradient operation. The state prediction loss for a single sample is defined as,

$$\mathcal{L}_{\text{dyn}} = \frac{1}{2} \sum_{d=1}^{d_s} \left[\frac{\left(y^d - \mu_{\theta_k}^d\right)^2}{\left(\sigma_{\theta_k}^d\right)^2} + \log\left(\sigma_{\theta_k}^d\right)^2 \right] + \frac{d_s}{2} \log(2\pi), \tag{27}$$

where $\mu_{\theta_k}^d$ and $\sigma_{\theta_k}^d$ denote the d-th elements of the predicted mean and standard deviation vectors, respectively. For reward prediction, we use the mean squared error loss,

$$\mathcal{L}_{\text{rew}} = \left\| \hat{r}_{\theta_k}(t) - r(t) \right\|^2, \tag{28}$$

where $r(t)$ is the truth reward from the offline dataset. The training loss of the MLP k is then defined as

$$\mathcal{L}_k = \mathcal{L}_{\text{dyn}} + \lambda_r \mathcal{L}_{\text{rew}}, \tag{29}$$

where λ_r is the weighting coefficient for the reward prediction loss. For a given state–action pair, the ensemble transition loss is formulated as,

$$\mathcal{L}_{\text{transition}} = \frac{\sum_{k=1}^{K} \sum_{b=1}^{B} \left(m_k^{(b)} \mathcal{L}_k^{(b)} \right)}{\sum_{k=1}^{K} \sum_{b=1}^{B} m_k^{(b)}}. \tag{30}$$

We integrate the global traffic state inference module and the probabilistic dynamics model into a unified traffic world model. The overall training objective is defined as,

$$\mathcal{L}_{\text{WM}} = \gamma_1 \mathcal{L}_{\text{state}} + \gamma_2 \mathcal{L}_{\text{transition}}, \tag{31}$$

where $\gamma_1 > 0$ and $\gamma_2 > 0$ are weighting coefficients.

In addition, we characterize model epistemic uncertainty using the dispersion among the predicted means of different ensemble members,

$$\chi\left(\hat{s}(t), a(t)\right) = \sqrt{\frac{1}{K} \sum_{k=1}^{K} \left\| \mu_{\theta_k}\left(\hat{s}(t), a(t)\right) - \overline{\mu}\left(\hat{s}(t), a(t)\right) \right\|^2}, \tag{32}$$

$$\overline{\mu}\left(\hat{s}(t), a(t)\right) = \frac{1}{K} \sum_{k=1}^{K} \mu_{\theta_k}\left(\hat{s}(t), a(t)\right). \tag{33}$$

When a region is sufficiently covered by the offline dataset, the predictions of different ensemble

members are generally consistent, resulting in low uncertainty. In contrast, when the policy visits state–action regions that are rarely covered by the training data, the predictions of different members may diverge significantly, leading to increased uncertainty. Therefore, $\chi\left(\hat{s}(t), a(t)\right)$ can serve as an important indicator of model extrapolation risk and is used in subsequent policy learning to restrict unreliable imagined interactions.

World Model-Based Offline Multi-Agent Policy Learning

After training, the world model is fixed and used to perform multi-step imagination rollouts to generate virtual trajectories for policy optimization. We adopt the CTDE paradigm and use multi-agent proximal policy optimization (MAPPO) (Yu et al. 2022) to optimize the cooperative control policy for multiple CAVs. Specifically, during training, the recovered global traffic state and joint actions are used to evaluate the system-level return, thereby capturing the coupling effects among different CAV control behaviors. During execution, each CAV makes decisions independently based only on its local observations, satisfying the requirement for decentralized control in practical traffic scenarios.

However, policy optimization in world model-based offline learning is prone to model exploitation, where the policy exploits poorly covered state–action regions with overestimated returns, leading to spurious improvement. To mitigate this issue, we construct a pessimistic reward based on the epistemic uncertainty estimated by the ensemble model,

$$\tilde{r}(\hat{s}(t), a(t)) = \hat{r}(\hat{s}(t), a(t)) - \beta\chi(\hat{s}(t), a(t)), \tag{34}$$

where β is the pessimistic penalty coefficient. When the ensemble models exhibit large prediction disagreement for a given state–action pair, the pessimistic reward automatically reduces the estimated return in that region.

During imagination rollouts, we start from the recovered state $\hat{s}(t)$ sampled from the offline dataset and first use the observation decoder O_ψ to obtain the local observation $\hat{o}_i(t)$ for each CAV. Each CAV then selects its control action according to its decentralized execution policy $\pi_{\nu_i}(a_i \mid \tau_i(t))$, and the individual actions are combined into the joint action $a(t)$. subsequently, the world model predicts the next state $\hat{s}(t+1)$ and outputs the pessimistic reward $\tilde{r}(\hat{s}(t), a(t))$ and epistemic uncertainty $\chi(\hat{s}(t), a(t))$. This procedure is repeated recursively to construct multi-step virtual trajectories.

To prevent error accumulation in long-horizon imagination rollouts and avoid policy learning from highly uncertain virtual experiences, we further introduce an epistemic uncertainty-based rollout truncation mechanism. Let $\chi_{\max}$ denote the threshold. If $\chi(\hat{s}(t), a(t)) \geq \chi_{\max}$ during rollouts, the current virtual trajectory is immediately terminated.

Solution Algorithm

Algorithm 1 summarizes the overall training procedure of the proposed framework. The world model is parameterized by $\Theta = \{\phi, \psi, \xi, \theta_{1:K}\}$, where ϕ, ψ, ξ, and $\theta_{1:K}$ denote the learnable parameters of the history encoder, observation decoder, cell-wise Cross-Attention module, and K ensemble transition-and-reward models, respectively.

NUMERICAL RESULTS

Experimental Setup

To evaluate the effectiveness of the proposed method, we conduct numerical experiments using the simulation of urban mobility (SUMO) platform in an on-ramp merging bottleneck scenario based on a section of the Hujin expressway, as illustrated in Fig. 3. The CAV cooperative control zones on the mainline and on-ramp are 3000 m and 600 m long, respectively. Both are discretized with a spatial step of $\Delta x = 300$ m, resulting in 10 traffic cells on the mainline and 2 cells on the on-ramp. The traffic demands on the mainline and on-ramp are set to 2500 veh/h and 1800 veh/h, respectively. The baseline CAV penetration rate is set to 15%, and the free-flow speed of all vehicles is 30 m/s. The longitudinal acceleration of each CAV is constrained within $a_i \in [-3\, m/s^2, 3\, m/s^2]$. Each simulation lasts for 1200 time steps, with the first 200 time steps used as a warm-up period.

Algorithm 1 Physics-Informed World Model-Based Offline MARL

Require: Offline dataset $\mathcal{D}$; ensemble size K; uncertainty threshold $\chi_{\max}$; pessimism coefficient β.

Ensure: World model Θ^* and decentralized policies Π^*.

Initialize world model $\Theta = \{\phi, \psi, \xi, \theta_{1:K}\}$
Initialize actors $\{\pi_{v_i}\}_{i\in\mathcal{N}}$ and critic Q_ς
for $iter = 1, \ldots, N_{\text{WM}}$ **do**
    Sample mini-batch from $\mathcal{D}$
    Build histories $\{\tau_i(t), \tau_i(t+1)\}_{i\in\mathcal{N}}$
    Infer states $\hat{s}(t)$ and $\hat{s}(t+1)$
    Compute physical prior $s^{\text{phy}}(t+1) = F_{\text{phy}}(\hat{s}(t), a(t))$
    Predict $(\hat{s}_{\theta_k}(t+1), \hat{r}_{\theta_k}(t)) = MLP_{\theta_k}(\hat{s}(t), a(t)), \quad k = 1, \ldots, K$
    Estimate uncertainty $\chi(t)$
    Update Θ by minimizing L_{WM}
end for
Set $\Theta^* \leftarrow \Theta$
for $iter = 1, \ldots, N_{\text{RL}}$ **do**
    Sample history from $\mathcal{D}$ and infer initial state $\hat{s}(t_0)$
    for $h = 0, \ldots, H_{\text{img}} - 1$ **do**
        Decode observations and build $\{\hat{\tau}_i(t_0+h)\}_{i\in\mathcal{N}}$
        Sample actions $a_i(t_0+h) \sim \pi_{v_i}(\cdot \mid \hat{\tau}_i(t_0+h))$
        Form joint action $a(t_0+h) = \{a_i(t_0+h)\}_{i\in\mathcal{N}}$
        Predict $(\hat{s}(t_0+h+1), \hat{r}(t_0+h), \chi(t_0+h))$ using Θ^*
        if $\chi(t_0+h) \geq \chi_{\max}$ **then**
            break
        end if
        Store transition with $\tilde{r}(t_0+h)$
    end for
    Update Q_ς and $\{\pi_{v_i}\}_{i\in\mathcal{N}}$
end for
return Θ^* and $\Pi^* = \{\pi^*_{v_i}\}_{i\in\mathcal{N}}$

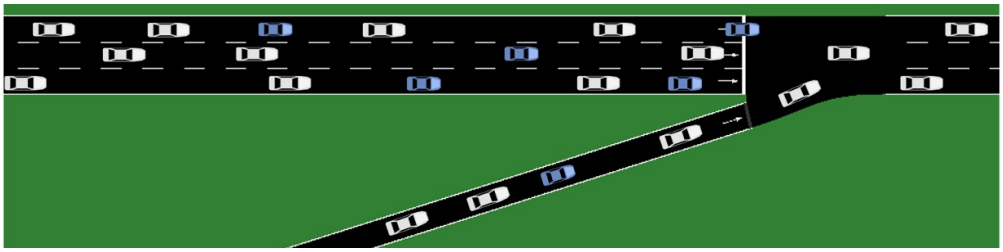


Figure 3 Simulation scenario of the on-ramp merging bottleneck

The offline dataset is generated using predefined behavior policies with different exploration levels under randomized traffic conditions. A total of 1000 independent traffic trajectories are collected, resulting in approximately 1×10^6 state-transition samples. To avoid information leakage caused by temporal correlations between consecutive samples, the dataset is split at the trajectory level into training, validation, and test sets with a ratio of 70%/15%/15%. All offline learning methods use the same training data, local observations, and action spaces to ensure a fair comparison.

The world model consists of 7 independently initialized ensemble members. The historical observation window is set to $H = 10$, and the imagination horizon is set to $H_{\text{img}} = 5$. The learning rates of the world model and policy networks are set to 1×10^{-3} and 1×10^{-4}, respectively. The pessimistic reward coefficient is set to $\beta = 0.5$, while the uncertainty truncation threshold $\chi_{\max}$ is determined as the 95th percentile of the epistemic uncertainty measured on the validation set. All control experiments are independently conducted using 5 random seeds, and the mean and standard deviation are reported.

Baselines and Evaluation Metrics

In addition to the without control case and the online multi-agent proximal policy optimization (MAPPO) (Yu et al. 2022) baseline, we consider eight methods to systematically investigate the effects of physics supervision, global state representation, and world model-based learning on offline multi-agent cooperative control.

- M1: Proposed. The proposed framework.
- M2: WM-Physical. This variant adopts the same physically interpretable state representation as M1 but removes the physics supervision from the state reconstruction process. Both the state reconstruction module and the world model are therefore trained in a fully data-driven manner.
- M3: WM-Latent. This variant learns a global latent state without explicit physical interpretation from the local historical information of multiple CAVs and trains a purely data-driven world model in the resulting latent space.
- M4: WM-Concat. This variant directly concatenates the local observations of all CAVs to form the global information representation and trains a purely data-driven world model based on this representation.
- M5-M8: Model-Free Offline Multi-agent Reinforcement Learning. These methods directly learn control policies from the offline dataset without using a world model to generate imagined trajectories. To ensure a fair comparison, all four methods adopt the same multi-agent implicit q-learning (MAIQL) (Tampuu et al. 2015) policy learning

framework and differ only in their global state representations. Specifically, M5 uses the physics-supervised global physical state, M6 uses the global physical state reconstructed without physics supervision, M7 uses the learned latent state, and M8 uses the concatenation of all CAV local observations.

These baselines enable the contribution of each component to be examined from complementary perspectives. The comparison between M1 and M2 evaluates the effect of physics supervision. The paired comparisons M1/M5, M2/M6, M3/M7, and M4/M8 assess the benefit of introducing a world model under comparable global state representations. In addition, comparisons among M1-M4 and among M5-M8 are used to examine the influence of different global state representations.

The methods are evaluated from two aspects: world model prediction accuracy and traffic control performance. For the world model, the mean absolute error (MAE) of state prediction and reward prediction is used to evaluate predictive accuracy. For traffic control, average travel time (ATT), average waiting time (AWT), and average time loss (ATL) are adopted to evaluate the system-level traffic control performance of different methods.

World-Model Prediction Accuracy

We first evaluate the ability of different world models to predict traffic-state evolution and system rewards. As shown in Table 1, M1-M4 achieve good next-state prediction accuracy within their respective representation spaces. In particular, M4 uses the concatenation of all CAV local observations as the state representation, without requiring additional state reconstruction or latent feature mapping. Its prediction task is therefore relatively direct within its own representation space, resulting in a comparatively low state prediction error.

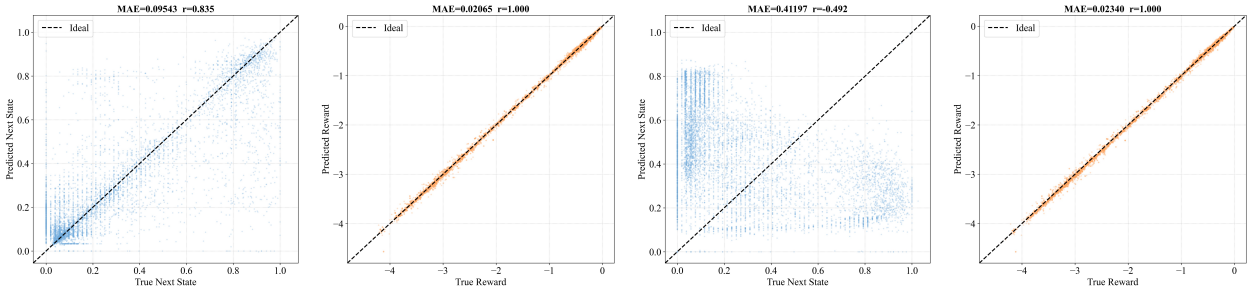
TABLE 1 World-model prediction accuracy under different state representations

Method	Recovered-State MAE	Recovered-Reward MAE	True-State MAE	True-Reward MAE
M1: Proposed	0.01633	0.02031	0.09543	0.02065
M2: WM-Physical	0.01912	0.02377	0.41197	0.02340
M3: WM-Latent	0.06158	0.03221	–	–
M4: WM-Concat	0.00704	0.03427	–	–

For M1 and M2, whose state representations have explicit physical interpretations, we further compare their predictions with the ground-truth traffic states obtained from SUMO, as illustrated in Fig. 4. The true-state MAE of M1 is 0.09543, compared with 0.41197 for M2, corresponding to a reduction of 76.84%. Similarly, the true-reward MAE decreases from 0.02340 for M2 to 0.02065 for M1, representing an improvement of 11.75%. These results indicate that traffic-flow physics guidance not only improves the predictive consistency within the reconstructed state space, but, more importantly, strengthens the correspondence between the reconstructed states and the actual traffic evolution.

Cooperative Control Performance

Table 2 reports the traffic control performance of different methods in the on-ramp merging scenario shown in Fig. 1. The proposed M1 achieves the best overall control performance among all offline learning methods, with the lowest ATT and ATL. Compared with the without control



(a) M1: State (b) M1: Reward (c) M2: State (d) M2: Reward

Figure 4 World-model prediction results of M1 and M2 against the SUMO ground truth

case, M1 reduces ATT, AWT, and ATL by 19.79%, 61.62%, and 36.10%, respectively. Moreover, its performance approaches that of Online MAPPO, indicating that the proposed method can learn effective cooperative control policies for multiple CAVs without requiring online interaction with the environment.

TABLE 2 Traffic control performance of different methods

Metric	Without Control	Online MAPPO	M1	M2	M3	M4	M5	M6	M7	M8
ATT (s)	145.81	107.93	**116.96**	122.92	127.59	132.14	120.69	127.40	129.23	135.90
AWT (s)	0.99	0.14	0.38	**0.32**	0.37	0.34	0.38	0.53	0.51	0.49
ATL (s)	85.44	47.51	**54.60**	65.43	68.33	76.32	57.10	70.37	72.80	81.21

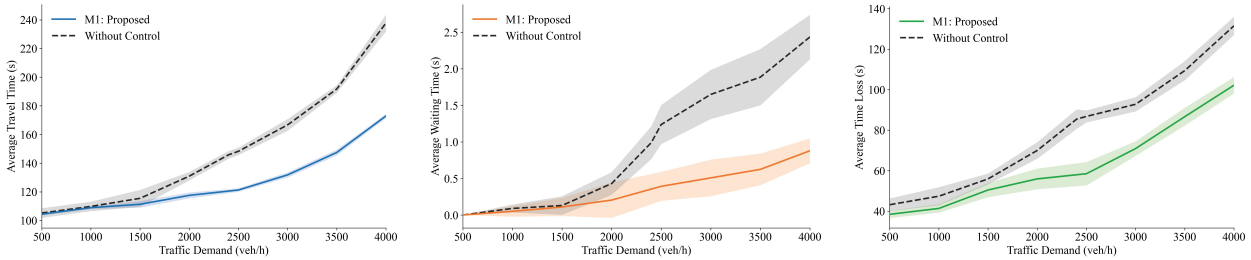
By incorporating physics supervision, M1 reduces ATT and ATL by 4.85% and 16.55%, respectively, compared with M2, although its AWT increases slightly. Combined with the world-model prediction results in Section 4.3, these results further suggest that physics supervision improves the reliability of global state reconstruction and world-model prediction, which in turn leads to better overall system-level traffic control performance.

In addition, the paired comparisons between M1-M4 and their corresponding model-free counterparts M5-M8 show that world model-based policy learning generally achieves better traffic control performance under comparable state representations. This indicates that imagined rollouts generated by the learned world model can effectively expand the coverage of the fixed offline dataset and provide additional training experience for offline policy optimization, thereby improving the learned cooperative control policies.

Sensitivity Analysis

To further evaluate the sensitivity of the proposed method under different traffic conditions, we examine the effects of traffic demand and CAV penetration rate on control performance.

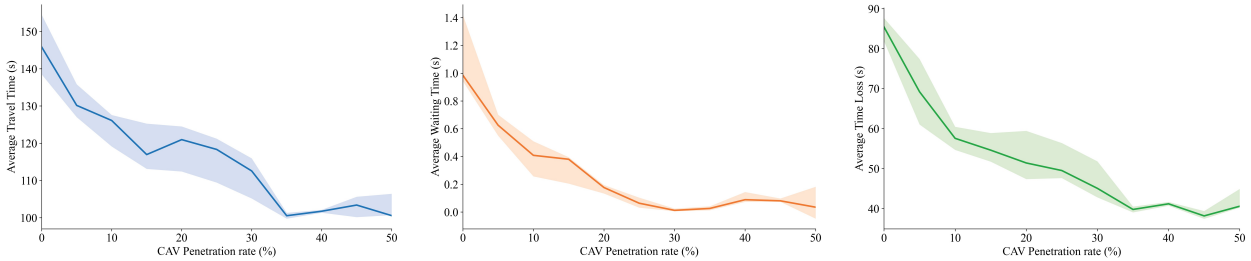
Fig. 5 compares the traffic performance of the proposed method with the without control case under different traffic demand levels. Under low traffic demand, the bottleneck operates under relatively uncongested conditions, resulting in only a small difference in ATT between the two cases. As traffic demand increases and bottleneck congestion intensifies, the performance gap progressively widens. In particular, when traffic demand exceeds approximately 2000 veh/h, ATT, AWT, and ATL increase rapidly in the without control case, whereas the proposed method



(a) Average travel time (b) Average waiting time (c) Average time loss

Figure 5 Sensitivity of traffic control performance to traffic demand

effectively suppresses their growth and maintains effective congestion mitigation even under high-demand conditions.



(a) Average travel time (b) Average waiting time (c) Average time loss

Figure 6 Sensitivity of traffic control performance to CAV penetration rate

Fig. 6 further investigates the effect of CAV penetration rate. Overall, ATT, AWT, and ATL exhibit decreasing trends as the CAV penetration rate increases, with the most pronounced improvements occurring at relatively low penetration levels. When the penetration rate reaches approximately 30%-35%, the marginal benefits of further increasing the number of CAVs gradually diminish. This indicates that once a sufficient number of CAVs are available to effectively regulate upstream traffic flow, additional controllable vehicles provide progressively smaller performance gains.

CONCLUSIONS

This paper proposes a physics-informed world model-based offline multi-agent reinforcement learning framework for cooperative CAV control at partially observable mixed-traffic bottlenecks. The proposed method reconstructs a structured global traffic state with explicit physical meaning from the local observation-action histories of multiple CAVs and enhances its consistency with traffic evolution through a coupled macroscopic-microscopic traffic model. Based on the reconstructed state, a probabilistic ensemble world model is developed to learn traffic state transitions and system rewards. Pessimistic rewards and uncertainty-driven rollout truncation are further incorporated to generate reliable imagined trajectories for offline cooperative policy learning. Simulation experiments based on a real-world on-ramp bottleneck demonstrate that physics supervision improves the reliability of state reconstruction and world-model prediction, leading to better cooperative control performance. Sensitivity analysis further shows that the proposed method remains effective under varying traffic demands and CAV penetration rates and provides consistent congestion mitigation under congested conditions. Overall, the proposed framework

enables effective cooperative CAV control without online trial-and-error interactions or access to complete global traffic states.

Future work will focus on validating the proposed framework using field-collected trajectory and CAV observation data, investigating its robustness to sensing and communication uncertainties as well as traffic-model mismatch, and extending the framework to more complex bottleneck configurations and larger-scale traffic networks. These extensions will further assess the practical transferability and generalization capability of the proposed method under realistic traffic conditions.

ACKNOWLEDGMENTS

The authors used OpenAI ChatGPT to assist with language editing and the identification of potentially relevant literature. All technical content and research results were independently developed by the authors, and all references were independently verified. The authors take full responsibility for the content of this manuscript.

AUTHOR CONTRIBUTIONS

The authors confirm contribution to the paper as follows: study conception and methodology: Lu Liu, Xi Xiong; data collection, experiments, and manuscript preparation: Lu Liu; manuscript review and revision: Lu Liu, Chi Xie, Xi Xiong. All authors reviewed and approved the final version of the manuscript.

DECLARATION OF CONFLICTING INTERESTS

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

FUNDING

The authors disclosed no financial support for the research, authorship, and/or publication of this article.

REFERENCES

Bernstein, Daniel S, Robert Givan, Neil Immerman, and Shlomo Zilberstein. 2002. The Complexity of Decentralized Control of Markov Decision Processes. *Mathematics of operations research*, Vol. 27, No. 4, pp. 819–840.

Bui, The Viet, Thanh Hong Nguyen, and Tien Mai. 2024. ComaDICE: Offline Cooperative Multi-Agent Reinforcement Learning with Stationary Distribution Shift Regularization. *arXiv preprint arXiv:2410.01954*.

Cang, Catherine, Aravind Rajeswaran, Pieter Abbeel, and Michael Laskin. 2021. Behavioral Priors and Dynamics Models: Improving Performance and Domain Transfer in Offline RL. *arXiv preprint arXiv:2106.09119*.

Chung, Koohong, Jittichai Rudjanakanoknad, and Michael J. Cassidy. 2007. Relation Between Traffic Density and Capacity Drop at Three Freeway Bottlenecks. *Transportation Research Part B: Methodological*, Vol. 41, No. 1, pp. 82–95.

Daganzo, Carlos F. 1994. The Cell Transmission Model: A Dynamic Representation of Highway Traffic Consistent with the Hydrodynamic Theory. *Transportation Research Part B: Methodological*, Vol. 28, No. 4, pp. 269–287.

Fang, Xing, Qichao Zhang, Yinfeng Gao, and Dongbin Zhao, Offline Reinforcement Learning for Autonomous Driving with Real World Driving Data. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems*, 2022, pp. 3417–3422.

Feng, Shuo, Ziyou Song, Zhaojian Li, Yi Zhang, and Li Li. 2021. Robust Platoon Control in Mixed Traffic Flow Based on Tube Model Predictive Control. *IEEE Transactions on Intelligent Vehicles*, Vol. 6, No. 4, pp. 711–722.

Hafner, Danijar, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. 2019. Dream to Control: Learning Behaviors by Latent Imagination. *arXiv preprint arXiv:1912.01603*.

Huang, Xiaohan, Xuesong Wang, and Yuhu Cheng. 2025. Uncertainty-Based Alternative Diffusion Policy for Safe Autonomous Driving. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 26, pp. 18854–18863.

Jin, Wen-Long. 2012. A Kinematic Wave Theory of Multi-Commodity Network Traffic Flow. *Transportation Research Part B: Methodological*, Vol. 46, No. 8, pp. 1000–1022.

Kang, Sehyeok, Yongsik Lee, Gahee Kim, Song Chong, and Se-Young Yun, MA^{2}E: Addressing Partial Observability in Multi-Agent Reinforcement Learning with Masked Auto-Encoder. In *International Conference on Learning Representations*, 2025, Vol. 2025, pp. 20145–20165.

Kidambi, Rahul, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims, MOReL: Model-Based Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 2020, Vol. 33, pp. 21810–21823.

Li, Li, Rui Jiang, Zhengbing He, Xiqun Chen, and Xuesong Zhou. 2020. Trajectory Data-Based Traffic Flow Studies: A Revisit. *Transportation Research Part C: Emerging Technologies*, Vol. 114, pp. 225–240.

Li, Ye, Bin Pan, Zhibin Chen, and Lu Xing. 2023. Developing a Dynamic Speed Control System for Mixed Traffic Flow to Reduce Collision Risks near Freeway Bottlenecks. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 24, No. 11, pp. 12560–12581.

Liao, Xishun, Ziran Wang, Xuanpeng Zhao, Kyungtae Han, Prashant Tiwari, Matthew J. Barth, and Guoyuan Wu. 2022. Cooperative Ramp Merging Design and Field Implementation: A

Digital Twin Approach Based on Vehicle-to-Cloud Communication. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 23, No. 5, pp. 4490–4500.

Lu, Gongyuan, Zili Shen, Xiaobo Liu, Yu Nie, and Zhiqiang Xiong. 2022. Are Autonomous Vehicles Better Off Without Signals at Intersections? A Comparative Computational Study. *Transportation Research Part B: Methodological*, Vol. 155, pp. 26–46.

Newell, Gordon F. 1993. A Simplified Theory of Kinematic Waves in Highway Traffic, Part II: Queueing at Freeway Bottlenecks. *Transportation Research Part B: Methodological*, Vol. 27, No. 4, pp. 289–303.

Qiu, Jiahua and Lili Du. 2023. Cooperative Trajectory Control for Synchronizing the Movement of Two Connected and Autonomous Vehicles Separated in a Mixed Traffic Flow. *Transportation Research Part B: Methodological*, Vol. 174, p. 102769.

Ran, Yuhang, Yi-Chen Li, Fuxiang Zhang, Zongzhang Zhang, and Yang Yu, Policy Regularization with Dataset Constraint for Offline Reinforcement Learning. In *Proceedings of the 40th International Conference on Machine Learning*, 2023, Vol. 202, pp. 28701–28717.

Rauch, Robert, Zdenek Becvar, Pavel Mach, and Juraj Gazda. 2024. Cooperative Multi-Agent Deep Reinforcement Learning for Dynamic Task Execution and Resource Allocation in Vehicular Edge Computing. *IEEE Transactions on Vehicular Technology*, Vol. 74, No. 4, pp. 5741–5756.

Richards, Paul I. 1956. Shock Waves on the Highway. *Operations Research*, Vol. 4, No. 1, pp. 42–51.

Ross, Stephane and J. Andrew Bagnell, Agnostic System Identification for Model-Based Reinforcement Learning. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, 2012, Vol. 2.

Tampuu, Ardi, Tambet Matiisen, Dorian Kodelja, Ilya Kuzovkin, Kristjan Korjus, Juhan Aru, Jaan Aru, and Raul Vicente. 2015. MultiAgent Cooperation and Competition with Deep Reinforcement Learning. *arXiv preprint arXiv:1511.08779*.

Wang, Jiao, Mingrui Yuan, Yun Li, and Zihui Zhao. 2023. Hierarchical Attention Master–Slave for Heterogeneous Multi-Agent Reinforcement Learning. *Neural Networks*, Vol. 162, pp. 359–368.

Wiesemann, Wolfram, Daniel Kuhn, and Berç Rustem. 2013. Robust Markov Decision Processes. *Mathematics of Operations Research*, Vol. 38, No. 1, pp. 153–183.

Xiong, Xi, Maonan Wang, Dengfeng Sun, and Li Jin. 2024. An Approximate Dynamic Programming Approach to Vehicle Platooning Coordination in Networks. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 25, No. 11, pp. 16536–16547.

Yu, Chao, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. 2022. The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games. *Advances in neural information processing systems*, Vol. 35, pp. 24611–24624.

Zhou, Yihe, Yuxuan Zheng, Yue Hu, Kaixuan Chen, Tongya Zheng, Jie Song, Mingli Song, and Shunyu Liu. 2025. Cooperative Policy Agreement: Learning Diverse Policy for Offline MARL. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39, pp. 23018–23026.

Čičić, Mladen and Karl Henrik Johansson, Traffic Regulation via Individually Controlled Automated Vehicles: A Cell Transmission Model Approach. In *2018 21st International Conference on Intelligent Transportation Systems*, 2018, pp. 766–771.