Preference-Based Reinforcement Learning for Autonomous Vehicle Control Considering the Benefits of Following Vehicles

Xiao Wen, Xinhu Zheng, Zhiyong Cui, Sisi Jian, *Member, IEEE* and Dengbo He

***Abstract*—Most studies developing car-following controllers for AVs in mixed traffic primarily focus on maximizing the utility of the AVs. However, the utility of the entire mixed traffic flow is largely ignored, which may lead to self-centered AVs. This study aims to develop a leading cruise controller for AVs that can "look behind" the mixed traffic flow. It enables the AV to adapt its car-following behaviors according to the states of both the leading human-driven vehicle (LHDV) and the following human-driven vehicle (FHDV). Car-following trajectories are extracted, processed, and analyzed based on the real-world autonomous driving dataset -- the Waymo Open Dataset. Then car-following behaviors of HDVs are approximated through an inverse reinforcement learning (IRL) approach. After that, this study proposes a preference-based soft actor-critic (PbSAC) algorithm to optimize the speed of AVs in the three-vehicle car-following scenario. In addition to safety, efficiency, and string stability for AV, these metrics for FHDV are also included in the reward function. An adaptive preference-adjusting module is developed to update the weights in the reward function based on expert knowledge. Experimental results indicate that the proposed algorithm can improve safety, efficiency, and string stability for both AV and FHDV. Moreover, if behavioral changes of human drivers when they are following AVs are not accurately captured, the learned AV controller may result in suboptimal performance.**

***Index Terms*—Car-following, inverse reinforcement learning, mixed traffic, reinforcement learning, traffic safety**

I. INTRODUCTION

AUTONOMOUS vehicles (AVs) are becoming a reality with the advancements of deep learning and sensors [1][2]. Although AV technologies are being developed at full speed, transportation researchers, road engineers, and AV manufacturers have realized that a transition period is expected due to the gradual AV deployment [3][4]. The reasons for this are manifold: the need to update the current infrastructure system, the public acceptance and trust towards AVs, and the necessity to enact suitable regulations and legislation [5][6]. During the transition period, there will be a combination of AVs and human-driven vehicles (HDVs) sharing road networks, resulting in a mixed-traffic environment [7]. Hence, the research question that arises is how the control logic of AVs should be designed to improve the safety, efficiency, and string stability of mixed traffic [8].

This study focuses on the car-following model, which depicts the longitudinal interactions between consecutive vehicles. In general, there are three main approaches to designing car-following models for AVs: (1) linear or non-linear AV controllers; (2) model predictive control (MPC) based AV controllers; and (3) deep reinforcement learning (DRL) based AV controllers [9]. The first category is computationally efficient but may fail to handle multiple objectives [10]. Although MPC is adaptable, it is also a time-intensive algorithm, as it requires addressing a constrained finite-horizon optimal control problem at every time step [11]. Therefore, researchers have turned to developing car-following models for AVs using DRL methods. There are two main advantages of DRL-based methods over the aforementioned categories. First, DRL is a learning-based and model-free approach that does not depend on predefined rules or stochastic system modeling [12]. Second, the computational cost of DRL is significantly lower than MPC since a trained DRL model can be implemented in rea time [13].

Although previous studies have concluded that the DRL-based longitudinal controllers for AVs could enhance the traffic flow performance, existing DRL-based models have three major drawbacks. First, these studies often focused on the optimization of only AVs but ignored the interactions between AVs and surrounding HDVs, especially the HDVs that are following AVs. For example, some literature simply considered a two-vehicle car-following scenario where the following vehicle is an AV controlled by a DRL algorithm, which means that the impacts of developed AV controllers on the following HDV (FHDV) are not explored. Although other studies concentrated on the mixed platoon with three or more vehicles, the reward functions of developed DRL algorithms

This work was supported by the National Natural Science Foundation of China under Grant 52202425, Guangzhou Municipal Science and Technology Project under Grant 2023A03J0011, and Project of Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone under Grant HZQB-KCZYB-2020083. *(Corresponding authors: Dengbo He).*

Xiao Wen is with the Intelligent Transportation, Division of Emerging Interdisciplinary Areas (EMIA) under Interdisciplinary Programs Office (IPO), The Hong Kong University of Science and Technology (HKUST), Kowloon, Hong Kong, SAR (e-mail: xwenan@connect.ust.hk).

Xinhu Zheng and Dengbo He are with the Thrust of Intelligent Transportation, HKUST(Guangzhou), Guangdong 511400, China, and also with the Department of Civil and Environmental Engineering, HKUST, Kowloon, Hong Kong, SAR. Dengbo He is also with HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Futian, Shenzhen (e-mail: xinhuzheng@hkust-gz.edu.cn; dengbohe@hkust-gz.edu.cn).

Zhiyong Cui is with the School of Transportation Science and Engineering, Beihang University, Beijing 100191, China (e-mail: zhiyongc@buaa.edu.cn).

Sisi Jian is with the Department of Civil and Environmental Engineering, HKUST, Kowloon, Hong Kong, SAR (e-mail: cesjian@ust.hk).

only included the benefits of AVs. Hence, the learned AV controllers could potentially result in self-centered and aggressive driving behaviors, which may cause significant traffic issues, such as congestion and even crashes [14]. In addition, many studies assumed that the AV could communicate with surrounding vehicles in the mixed-traffic environment. However, this may not be realistic at the current stage since AVs that have been deployed on public roads (e.g., Waymo) cannot communicate with surrounding vehicles.

Second, former studies considered the car-following control as a multi-objective reinforcement learning (MORL) problem by including multiple terms in the reward function. Specifically, they converted the multi-objective reward vector into a scalar weighted-sum reward and then identified the optimal weight combination by massive experiment trials [15]. However, manually determining the weight for each objective may lead to two critical deficiencies: (1) choosing the appropriate scalarization is non-trivial and time-consuming because objectives are often defined in different units and scales and there might be trade-offs among objectives, and (2) optimizing the DRL model upon the fixed weight combination can only infer an optimal policy for specific preferences. As a result, the learned optimal policy has limitations in its adaptability to various user preferences and different driving scenarios. Facing these issues, this paper will propose an innovative algorithm -- preference-based soft actor-critic (PbSAC) for AV longitudinal control. Given any user-specified preference, PbSAC can adaptively adjust the weight combination and then produce the optimal policy.

Third, most studies assumed that HDVs would implement the same car-following behaviors regardless of whether they were following AVs or HDVs [16]. However, some studies suggest that HDVs may adapt their behaviors when following AVs since AVs are fundamentally different from HDVs [3]. For example, by conducting field experiments, Rahmati et al. [17] revealed that HDV-following-AV would make HDVs feel more comfortable compared to HDV-following-HDV based on Kahneman and Tversky's Prospect Theory. Besides, traffic simulation results highlighted the importance of incorporating human behavior adaptations when analyzing the mixed traffic stream. Mahdinia et al. [18] found that AVs in mixed traffic could induce significant behavioral benefits to stability, safety, and the environment. Based on the real-world autonomous driving dataset collected by Waymo (Waymo Open Dataset), Wen et al. [3] compared HDV-following-AV and HDV-following-HDV events and concluded that HDV-following-AV exhibited significantly lower driving volatility, time headway and time-to-collision (TTC). Hence, this study will investigate how the performance of the proposed AV controller is influenced if such behavioral changes are not captured.

AVs have advanced sensors that can collect real-time and accurate information on the surrounding environment. In this light, we argue that AVs should exhibit more selfless behaviors to enhance the traffic flow performance, given their superior ability to perceive the surrounding environment compared to HDVs. This study is among the few studies to develop a DRL-based leading cruise control algorithm for AVs in the three-vehicle car-following scenario, which enables AVs to "look behind" mixed traffic. As a result, AVs make car-following decisions adapting to the states of both the leading human-driven vehicle (LHDV) and FHDV, and improve safety, efficiency, and string stability of mixed traffic. PbSAC is developed as the AV car-following control model to handle the multi-objective optimization problem and a preference generator is proposed to dynamically determine the weight combination in the reward function based on expert evaluation. FHDV car-following behaviors are acquired through inverse reinforcement learning (IRL) using the Waymo Open Dataset. Two types of car-following models are obtained to mimic FHDV driving behaviors in HDV-following-AV and HDV-following-HDV scenarios. A comparison is made to investigate whether implementing different types of car-following models for FHDV has significant impacts on the PbSAC performance.

The main contributions of this study are:

1) A DRL-based AV controller that can make car-following decisions based on the states of both LHDV and FHDV is proposed. The controller can improve the mixed traffic flow performance by incorporating the safety, efficiency, and string stability of both AV and FHDV in the reward function.
2) To address the multi-objective car-following control problem, we extend SAC to PbSAC where a preference learning module is proposed based on the Bayesian learning framework. This module integrates prior knowledge to automatically adjust the weights in the reward function towards expert preferences.
3) FHDV driving behaviors are modeled using the HDV-following-AV or HDV-following-HDV car-following model. Hence, this study can also quantify the impacts of FHDV car-following model types on the performance of the proposed AV controller.

The remainder of this paper is organized as follows. Section II gives a summary of studies on AV control models in mixed traffic. Section III explains the proposed AV controller's methodologies and study scenarios. Section IV discusses the experimental settings, evaluates the model performance, and discusses the results. Lastly, Section V concludes the paper and suggests future research.

II. Related Works

There has been a great deal of literature on the development of longitudinal controllers for AVs in past decades. As mentioned in Section I, these studies can be generally categorized into three types: linear or non-linear-based, MPC-based, and DRL-based AV controllers.

Linear and non-linear controllers have parameterized formulations and are easy to implement. For instance, in Naus et al. [19], a linear cooperative adaptive cruise control (CACC) model was developed and the required criteria for

achieving string stability were analyzed. Morbidi et al. [20] proposed an infinite-time linear-quadratic regulator with measurable disturbances for platoon control. Zhang and Orosz [21] proposed a nonlinear controller for AVs such that they can mimic the behaviors of HDVs in the platoon. There are some other non-linear longitudinal control frameworks, e.g., the intelligent driver model (IDM) [22]. However, it is often difficult to design a multi-objective linear or non-linear controller to include safety constraints within reasonable vehicle acceleration ranges [9].

At each time step, MPC computes a set of acceleration values by solving an optimal control problem over a prediction horizon. However, only the initial acceleration value in the sequence is utilized [12][13]. This computation process will be iterated until the terminal requirements are met. For example, Gong et al. [23] considered the CACC platoon as an interconnected dynamic system that has acceleration, speed, and safety constraints. They solved a one-horizon MPC problem to optimize the objectives that stem from the transient and asymptotic dynamics. Wang et al. [24] developed a decentralized and distributed algorithm for CACC under a receding horizon control framework. In [11], the authors suggested a stochastic optimal control method with a rolling horizon, utilizing the constant time gap policy to regulate ACC and CACC in the presence of uncertainty. However, the problem to be convex for MPC to work effectively. The computational demands may vary depending on the complexity of the formulation.

DRL-based approaches usually take the vehicle speed, gap, and relative speed as inputs and output the acceleration of AV. For instance, Qu et al. [25] developed a DRL-based car-following model to mitigate stop-and-go disturbances and improve the energy economy of electric vehicles. The vehicle speed and time gap were incorporated into the reward function. Zhu et al. [13] proposed a safe, efficient, and comfortable DRL-based controller in the car-following scenario where TTC, time gap, and jerk values were considered in the reward function. Simulation results revealed that the DRL-based controller outperformed the MPC-based controller. Then a cooperative longitudinal control strategy for the CACC platoon based on DRL was developed in [10]. Simulation results showed that compared to human driving, it could dampen the speed fluctuations of LHDV and improve car-following efficiency and energy economy under different CACC market penetration rates (MPRs). In [26], the authors designed an AV control algorithm using the automating entropy adjustment on Tsallis actor-critic (ATAC) and considered the time margin, time gap, and jerk. There are other applications of DRL-based AV controllers, such as [27][28][29].

III. Methodology

This section first describes the formulation of the car-following problem, introduces the design of the reward function and the proposed PbSAC model after that, and finally provides the study scenarios for comparison.

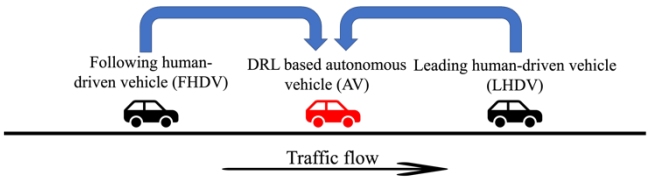


Fig. 1. The proposed "look behind" control system has two HDVs and one AV in the platoon. The black vehicles are HDVs and the red ones are AVs. The blue arrows show the information collected by the AV.

A. Problem Formulation

In Fig. 1, this study concentrates on the three-vehicle car-following scenario, representing a typical segment of mixed traffic. More specifically, the lead vehicle is LHDV which serves as an external input to achieve the speed profile derived from the Waymo Open Dataset. The second vehicle is an AV controlled by PbSAC. The AV can collect real-time information on the states of LHDV and FHDV. Then the HDV-following-AV model is implemented to approximate FHDV car-following behaviors.

During a car-following scenario, the following vehicle modifies its acceleration to follow the leading vehicle. Therefore, the car-following problem is converted to a Markov Decision Process (MDP) defined by a tuple: (S, A, R, P, γ), where S is the state space, A is the action space, $R(s_t, a_t)$ is the reward function representing the reward resulting from the interaction with the environment, the transition function $P(s_{t+1}|s_t, a_t)$ determines the next state given the current state and action, and γ is the discount factor. At each time step t, the vehicle adopts an action a_t according to the policy $\pi(a_t|s_t)$. After that, the state s_t is updated to s_{t+1} based on the transition function, and a scalar reward $R(s_t, a_t)$ is returned. The objective of an MDP is to find the optimal policy to maximize the expected return: $\pi^* = arg\max_{\pi} \sum_{t=1}^{\infty} E_{(s_t,a_t)\sim\rho_\pi}[R(s_t, a_t)]$.

In the following, basic DRL concepts in the scenarios that are under investigation are provided. Note that "AV" and "agent" will be used interchangeably in the remainder of the paper.

1) State: This study assumes that AVs collect real-time and accurate movement information on LHDV and FHDV, e.g., speed and position, via onboard sensors. Let s_t denote the state of AV at the time step t. The state vector s_t involves two aspects of information: (1) the speed of AV $V_1(t)$, the gap between AV and FHDV $S_{1,2}(t)$, the gap between LHDV and AV $S_{0,1}(t)$, the relative speed between AV and FHDV $\Delta V_{1,2}(t)$ and the relative speed between LHDV and AV $\Delta V_{0,1}(t)$; and (2) the weight parameters in the reward function.

2) Action: The action that an AV needs to take at the time step t is the longitudinal acceleration $a_1(t)$ which is bounded between -3 and $3m/s^2$ [30].

3) Transition function: After taking the action at the time step t, the three-vehicle platoon state is updated using the kinematic point-mass function:

$$V_2(t+1) = V_2(t) + \Delta T * a_2(t) \quad (1)$$

$$V_1(t+1) = V_1(t) + \Delta T * a_1(t) \quad (2)$$

$$\Delta V_{1,2}(t+1) = V_2(t+1) - V_1(t+1) \quad (3)$$

$$\Delta V_{0,1}(t+1) = V_1(t+1) - V_0(t+1) \quad (4)$$

$$S_{1,2}(t+1) = S_{1,2}(t) - \Delta T * \frac{\Delta V_{1,2}(t)+\Delta V_{1,2}(t+1)}{2} \quad (5)$$

$$S_{0,1}(t+1) = S_{0,1}(t) - \Delta T * \frac{\Delta V_{0,1}(t)+\Delta V_{0,1}(t+1)}{2} \quad (6)$$

where ΔT is the simulation time interval (which is $0.1s$ in this study); $V_0(t)$ and $V_2(t)$ are the speed of LHDV and FHDV at the time step t; $a_2(t)$ is the acceleration of FHDV at the time step t.

B. Reward Function

This section will describe the development of the reward function, which is designed to address safety, efficiency, and string stability in a typical car-following situation. It should be noted that there are no universally accepted methods for formulating the reward function, and thus in this study, it is defined through a combination of prior research and trial-and-error.

Inspired by [9][10], the reward function is designed to account for both control efficiency and string stability. The control efficiency reward $R_{control}$ is expressed as a quadratic function that regulates the vehicle to maintain the predefined state:

$$R_{control} = -(\boldsymbol{x}(t))^T \boldsymbol{Q} \boldsymbol{x}(t) \quad (7)$$

$$\boldsymbol{x}(t) = [\Delta\tilde{S}(t), \Delta\tilde{V}(t)] \quad (8)$$

$$\Delta\tilde{S}(t) = S_{n-1,n}(t) - t^* * V_n(t) \quad (9)$$

$$\Delta\tilde{V}(t) = V_n(t) - V_{n-1}(t) \quad (10)$$

$$\boldsymbol{Q} = \begin{bmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{bmatrix}, \alpha_1, \alpha_2 > 0 \quad (11)$$

where $\boldsymbol{x}(t)$ refers to the difference from the predetermined state. $\Delta\tilde{S}(t)$ indicates the discrepancy between the actual gap and the target gap, and $S_{n-1,n}(t)$ denotes the actual gap between consecutive vehicles. The target gap is defined as $t^* * V_n(t)$ where t^* is the target time gap. In this study, t^* is set to be $1.5s$, according to [31]. $\Delta\tilde{V}(t)$ represents the relative speed between consecutive vehicles while $V_n(t)$ and $V_{n-1}(t)$ are the speed of the vehicle and its leader. α_1 and α_2 are fixed at 0.04 and 0.08 based on massive experiments.

As the control efficiency reward $R_{control}$ is optimized to be maximum, $\Delta\tilde{S}(t)$ and $\Delta\tilde{V}(t)$ will both converge to zero. This indicates that the actual gap is approaching the target gap. Besides, note that the commonly-used safety indicator – TTC is calculated in Eq. (12):

$$TTC(t) = \frac{S_{n-1,n}(t)}{V_n(t)-V_{n-1}(t)} \quad (12)$$

When $\Delta\tilde{V}(t) = V_n(t) - V_{n-1}(t) \to 0$, $TTC(t) \to \infty$. Since the proposed AV controller considers the benefits of AV and FHDV, both $R_{control\_AV}$ and $R_{control\_FHDV}$ are incorporated into the reward function.

The capacity of a vehicle to dampen the speed oscillations of its leader, known as string stability, is a crucial attribute. If the magnitude of the leader's perturbation is greater than the magnitude of the vehicle's perturbation, the vehicle is deemed string stable. By referring to [9][10][30], it is found that string stability is negatively correlated with acceleration behaviors, meaning that sharp and frequent acceleration behaviors result in the string instability issue. Hence, this study uses the negative square of the acceleration as the string stability reward $R_{stability}$ and another term $-\frac{(a_n(t)-a_n(t-1))^2}{(a_{max}-a_{min})^2}$ is added to ensure driving comfort:

$$R_{stability} = -a_n(t)^2 - \frac{(a_n(t)-a_n(t-1))^2}{(a_{max}-a_{min})^2} \quad (13)$$

where a_{max} and a_{min} are the maximum and minimum acceleration (which are $3m/s2$ and $-3m/s2$ in this study), respectively. Similarly, both $R_{stability\_AV}$ and $R_{stability\_FHDV}$ are considered in the reward function.

In summary, Eq. (14) shows the completed reward function:

$$R_1 = \omega_1 R_{control\_AV} + \omega_2 R_{control\_FHDV} + \omega_3 R_{stability\_AV} + \omega_4 R_{stability\_FHDV} \quad (14)$$

$$s.t. \quad \forall i: \omega_i > 0$$

$$\sum_{i=1}^{4} \omega_i = 1$$

where ω_1, ω_2, ω_3, and ω_4 are the weight parameters that need to be tuned.

C. Preference-based Soft Actor-Critic (PbSAC)

The reward function (as presented in Eq. (14)) contains four terms, and manually setting the weight parameters is a challenging and laborious task. However, this study proposes a new preference generator that can automatically optimize the weights for each reward term using expert knowledge. For example, the developers of the AV algorithm can assess AV car-following trajectories by evaluating the average time gap. The weights will then be adjusted based on their preferences. Subsequently, the control policy can be improved by considering the updated weights in the reward function.

According to [15][32], the weight parameters are updated on human preferences using a Bayesian learning framework. Let τ_i denote the ith car-following trajectory of an AV. The total rewards obtained from the trajectory τ_i is:

$$\boldsymbol{r}(\tau_i) = (r_1^i, r_2^i, \dots, r_T^i) \quad (15)$$

where T is the length of the car-following trajectory.

Then any trajectory pair (i.e., τ_i and τ_j) can be compared according to expert evaluation, e.g., a pair of car-following trajectories can be compared based on safety and efficiency indicators. Following [32], a probabilistic model over expert preferences is employed to compare two trajectories:

$$p(\tau_i > \tau_j | w) = \frac{\exp(w^T r(\tau_i))}{\exp(w^T r(\tau_i)) + \exp(w^T r(\tau_j))} \quad (16)$$

where $\tau_i > \tau_j$ means the preference of τ_i over τ_j based on expert evaluation. As a result, trajectories that achieve a higher linearly scalarized reward will be given a higher ranking.

Algorithm 1 Preference-based Soft Actor-Critic (PbSAC)

Initialize memory buffer M.
Initialize system buffer S.
Initialize trajectory buffer T.
Initialize weight parameters w.
Initialize critic Q_θ and actor π_ϕ networks.
for ep in {1,2,3, ... N} **do**
  for t in {1,2,3,…T} **do**
    Sample action a_t based on actor network, given current state s_t.
    Implement a_t and transfer to the next state s_{t+1}.
    Obtain reward vector $[R_{safety_{AV}}, R_{safety_{FHDV}}, R_{stability_{AV}}, R_{stability_{FHDV}}, R_{comfort\_AV}, R_{comfort\_FHDV}]$.
    Save the reward vector into trajectory buffer T.
    Calculate reward r_t based on Eq. (14) and current weights w.
    Save transition (s_t, a_t, r_t, s_{t+1}) into M.
    for k in {1,2,3,…K} **do**
      Sample random batch of B transitions from M
      Update θ and ϕ
    end for
  end for
  Save T into system buffer S.
  if ep mod L ≡ 0 **then**
    Sample L trajectory pairs from S and evaluate them upon expert knowledge.
    Calculate the likelihood of w using Eq. (16).
    Update weight parameters w through MCMC using Eq. (18).
    Clear trajectory buffer T.
  end if
end for

After each policy improvement, a pair of trajectories (τ_i, τ_j) will be randomly sampled and evaluated based on the removal of posterior volume (as shown in Eq. (17)). (τ_i, τ_j) that removes the most volume among the previous pairs since the last posterior update will be saved into the volume buffer and prepared for the next comparison.

$$\max_{(\tau_i, \tau_j)} \min(E_w[1 - p(\tau_i > \tau_j | w)], E_w[1 - p(\tau_j > \tau_i | w)]) \quad (17)$$

Let q_m represent the comparison between two trajectory samples: $q_m = (\tau_i > \tau_j)_m$. After multiple pairwise comparisons $\{q_1, q_2, \dots, q_n\}$, the posterior of the weights can be updated in a Bayesian fashion:

$$p(w | q_1, q_2, \dots, q_n) \propto p(w) \prod_{m=1}^{n} p(q_m | w) \quad (18)$$

where $p(w)$ is the prior which is set to be normally distributed. This Bayesian model can prioritize the reward components in each iteration. Then an expert can regularize the agent's behaviors by providing pairwise preferences. Finally, the weight parameters can be updated using Markov Chain Monte Carlo (MCMC), and the Softmax function is implemented to normalize weight parameters.

Such a weight-adjusting module can be combined with most off-the-shelf DRL methods. In this study, soft actor-critic (SAC) [33] is selected as the backbone DRL algorithm for multi-objective car-following control. The SAC algorithm is a model-free DRL algorithm that employs an actor-critic approach, and is designed to work with continuous action spaces. It considers both the expected rewards and the policy entropy. For a detailed explanation of the SAC algorithm, readers are encouraged to refer to [33]. The implementation specifics of PbSAC are outlined in Algorithm 1.

D. Study Scenarios

This paper proposes a PbSAC-based leading cruise controller for AVs in the three-vehicle car-following condition. To validate the effectiveness of the proposed approach, the four scenarios are carried out for comparison:

- *Scenario 1*: The AV control algorithm based on PbSAC utilizes data from the three vehicles and takes into account the benefits of AV and FHDV. As a result, the agent's state is defined as $[V_1(t), S_{1,2}(t), S_{0,1}(t), \Delta V_{1,2}(t), \Delta V_{0,1}(t), \omega_1, \omega_2, \omega_3, \omega_4]$ and the reward function is specified in Eq. (14). According to [31], there is a trade-off among safety, efficiency, and stability where pursuing less efficiency (e.g., larger time gaps) can lead to improvements in safety and stability. Thus, expert preferences are applied in the following way: given two car-following trajectories (τ_i, τ_j), if the discrepancy between the average time gap and $1.5s$ in τ_i is smaller than that in τ_j, $(\tau_i > \tau_j)$ is returned.
- *Scenario 2*: The AV control algorithm based on PbSAC uses data from both LHDV and AV and only takes into account the benefits of AV. Consequently, the state of the agent is $[V_1(t), S_{0,1}(t), \Delta V_{0,1}(t), \omega_5, \omega_6]$ and the reward function is defined in Eq. (19). Expert preferences are provided like in scenario 1.

$$R_2 = \omega_5 R_{control\_AV} + \omega_6 R_{stability\_AV} \quad (19)$$

- *Scenario 3*: The sole difference between scenario 2 and scenario 3 is that the agent is controlled by SAC, resulting in the manual determination of weight parameters through extensive experiments: [0.5,0.5].

- *Scenario 4*: The agent is an AV with an MPC-based adaptive cruise control (ACC) module. It also does not use FHDV's information and ignores the benefits of FHDV. The constrained linear-quadratic MPC model developed in [12][13] is adopted:

$$\min_{a} \sum_{t=0}^{N-1} [\beta_1 \left(\frac{\Delta V_{0,1}(t+1)}{\Delta V_{max}}\right)^2 + \beta_2 \left(\frac{S_{0,1}(t) - \tilde{S}_{0,1}(t)}{S_{max}}\right)^2 + \beta_3 (\frac{a_1(t)}{a_{max}})^2 + \beta_4 (\frac{a_1(t) - a_1(t-1)}{a_{max} - a_{min}})^2] \quad (20)$$

$$s.t. \quad \boldsymbol{f}(t+1) = \boldsymbol{A}\boldsymbol{f}(t) + \boldsymbol{B}\boldsymbol{u}(t)$$

$$S_{0,1} > 0$$

$$V_1 > 0$$

$$-3 \le a \le 3$$

where N is the prediction horizon which is set to be 10; β_1, β_2, β_3 and β_4 are fixed at 20, 20, 1 and 1; $\tilde{S}_{0,1}(t) = 1.5 * V_1(t)$ represents the constant time gap policy; ΔV_{max}, and S_{max} are constants to normalize various trajectory errors ($\Delta V_{max} = 15, S_{max} = 30$); $\boldsymbol{A}$ and $\boldsymbol{B}$ are matrices used to update the state of AV ($\boldsymbol{A} = [[1, \Delta T, 0], [0,1,0], [0,0,1]]^T, \boldsymbol{B} = [[-0.5\Delta T^2], [-\Delta T], [\Delta T]]^T$); $\boldsymbol{f}(t) = [S_{0,1}(t), \Delta V_{0,1}(t), V_1(t)]$; $\boldsymbol{u}(t) = [a(0), a(1), \dots, a(N-1)]$. Note that these hyper-parameters are defined based on [12][13] and our experiments.

IV. Experimental Results

A. Waymo Open Dataset

In this study, trajectory data is extracted from the Waymo Open Dataset [34][35], which contains the real-world autonomous driving scenarios. Waymo AVs are equipped with 5 LiDARs and 5 cameras, enabling high-precision data collection on AVs and surrounding vehicles at a frequency of 10 Hz on public roads in the U.S. The Waymo Open Dataset is divided into two parts: perception and motion. The perception part includes 1,000 $20s$ segments with well-synchronized and calibrated high-resolution LiDAR and camera data recorded in urban and suburban areas [34]. The motion part has 103,354 $20s$ segments, representing 574 hours of driving data over 1,750 km of roadways [35]. Each segment denotes each road agent as a 3D ground truth bounding box and high-resolution maps that correspond to the recorded data are provided. Both the perception and motion parts offer continuous and high-quality records of road agents' type, size, position, and trajectory.

To extract qualified car-following events, this study adopts the following rules based on previous research in car-following event extraction [3][5]:

1) The lead and following vehicles were traveling in the same lane on a straight segment of the highway.
2) Throughout the event, neither the lead nor the following vehicle changed lanes.
3) The gap between the lead and following vehicles was less than $60m$.
4) The speed of the lead and following vehicles was higher than $3m/s$.
5) The duration of a car-following event was longer than $15s$.

253 HDV-following-AV and 1,301 HDV-following-HDV events are extracted.

B. Data Processing

This section describes the data processing framework for preparing raw data for car-following research. The framework involves three steps. First, the raw data is denoised using a two-step trajectory reconstruction method. Second, each car-following event is calibrated using the intelligent driver model (IDM) and the calibrated parameter distributions are compared. Third, an inverse reinforcement learning (IRL) based model is used to fit HDVs' trajectories in HDV-following-AV and HDV-following-HDV scenarios.

1) Data denoising: Following the guidelines in [4], a two-step data denoising approach is implemented to enhance the quality of car-following trajectories. In the first step, this study involves developing an optimization-based outlier removal approach. Trajectory points with anomalous acceleration are identified as outliers and a linear optimization model is utilized to replace the original trajectories with outlier-free and smooth trajectories. Smoothness in this context refers to the variation between the maximum and minimum acceleration within the optimization horizon. In the second step, a second-order Savitzky-Golay filter is applied to denoise the car-following trajectories to remove unusual fluctuations in the position and speed data.

TABLE I
Calibrated IDM Parameters and Mann-Whitney U Test Results

	HDV-following-AV		HDV-following-HDV		Mann-Whitney U test
	Mean	Std	Mean	Std	
a_{max}	1.61	0.86	1.46	0.83	**0.01**
$a_{comfort}$	3.01	1.67	3.04	1.73	0.74
$\tilde{V}$	22.53	6.97	21.95	6.98	0.18
$\tilde{T}$	0.91	0.67	0.99	0.63	**0.02**
S_{jam}	4.62	3	4.87	3.09	0.26

2) Car-following model comparison: To quantify the effects of AVs on FHDV, IDM is fitted for each extracted car-following pair in HDV-following-AV and HDV-following-HDV scenarios. The study specifically employs the genetic algorithm (GA), using gap as the measure of performance and root mean square error (RMSE) as the goodness-of-fit, as recommended by [36]. The mean values and standard deviation (Std) of calibrated IDM parameters in two scenarios and the results of the Mann-Whitney U test are shown in Table I. One can observe that when comparing calibrated parameter sets in two scenarios, the differences in maximum acceleration a_{max} ($p - value = .02$) and desired time gap $\tilde{T}$ ($p - value < .01$) are significant at the 95% confidence level. According to [37], the most significant contribution to RMSE of gap comes from $\tilde{T}$ which accounts for the most share of the variance of $RMSE$, followed by a_{max},

whereas the other parameters are negligible. In summary, it is revealed that AVs have significant effects on the car-following behaviors of FHDV.

3) Car-following model fitting: According to the results of the previous steps, an IRL method -- Inverse soft-Q Learning [38] is conducted to recover human driver reward functions in HDV-following-AV and HDV-following-HDV scenarios. By approximating the Q-function, which represents both the reward and policy, this method turns the complex min-max game in traditional IRL methods into a simple minimization problem. With the recovered reward functions, a SAC model is trained to mimic human driving policies when they are following AVs or HDVs. The model training and selection procedures are similar to our previous studies [5][7]. Finally, two types of car-following models, HDV-following-AV and HDV-following-HDV, are retained and will be adopted to control FHDV in this study.

C. Simulation Setup and PbSAC Training

As mentioned earlier, 1,301 HDV-following-HDV events have been identified in the Waymo Open Dataset. Out of these car-following events, 459 three-HDV platoons are extracted. These three-HDV platoons are randomly divided into training and testing sets, where 80% (367) of them are used for training while the rest of them (92) are used for testing.

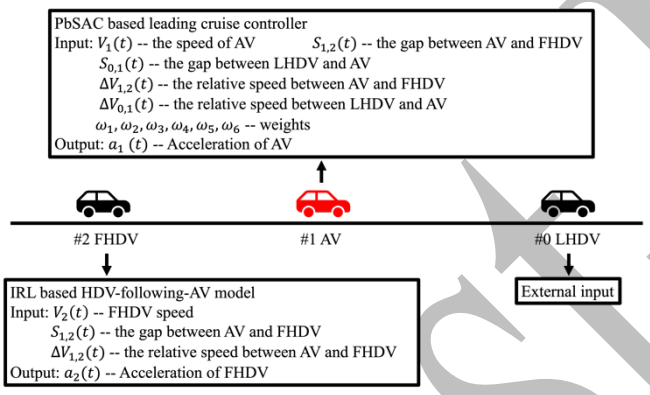


Fig. 2. PbSAC-based AV control framework in scenario 1.

Fig. 2 shows the framework of the PbSAC-based AV controller in scenario 1. As it is shown, LHDV is treated as the external input. The AV takes the states of itself and FHDV as inputs, outputs the longitudinal acceleration, and considers the benefits of itself and FHDV. In contrast, the control models of AVs in scenarios 2, 3, and 4 only focus on AVs without considering FHDVs. In scenarios 1, 2, 3, and 4, FHDVs are controlled by the HDV-following-AV model acquired through IRL.

The training process of a three-vehicle car-following event is defined as an episode. At the beginning of each episode, a three-HDV platoon from the training set is randomly drawn. Then the speed profile of the three-HDV platoon leader is assigned to LHDV in Fig. 2 and the initial states of AV and FHDV in Fig. 2 are set to match the two followers in the three-HDV platoon. AV and FHDV will take actions based on corresponding controllers. Afterwards, the states of AV and FHDV are updated according to Eqs. (1)-(6). When a traffic crash happens, the training will be terminated, and return a reward -10. Otherwise, the AV control algorithm ends the current episode at the maximum time step of the LHDV speed profile. Then another car-following event will be randomly selected from the training set, and the states will be re-initialized with the new empirical data.

The hyperparameters of the PbSAC model in scenario 1 have been listed in Table II. The critic and actor networks both contain two fully connected layers (64 neurons in each layer) with the rectified linear unit (ReLU) activation function. Fig. 3 presents the training process of the PbSAC model in scenario 1. The study employs the rolling mean episode reward as a metric to determine model convergence. To calculate the mean episode reward, the average reward of one episode is recorded since the training step for each episode is not consistent. The rolling mean episode reward is then obtained by averaging the mean episode rewards using a window size of 50. It is shown that the PbSAC-based algorithm has been trained with 1,200 episodes. A sharp increasing trend can be identified in the first 200 episodes. As the training process proceeds, the rolling mean episode reward begins to converge, indicating the stableness and effectiveness of the reward function and the preference generator. As a result, the PbSAC model with the highest rolling mean episode reward is selected as the AV controller in scenario 1. The same model training and selection strategy is also applied to scenarios 2 and 3.

TABLE II
HYPERPARAMETER SETTINGS OF PBSAC IN SCENARIO 1

PARAMETERS	VALUE
Optimizer	Adam
Learning rate	0.0001
Replay buffer size	10,000
Discount factor	0.99
Minibatch size	64
Soft update factor	0.005

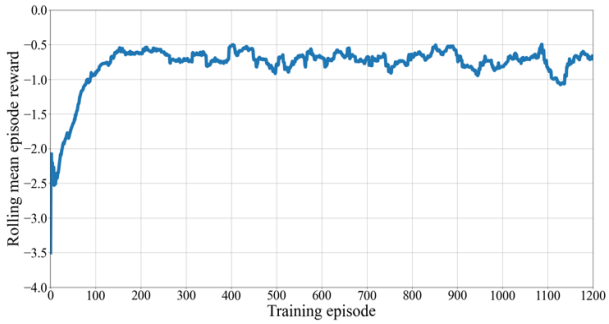


Fig. 3. Training process of the proposed PbSAC-based model in scenario 1.

D. Evaluation of the Proposed Approach

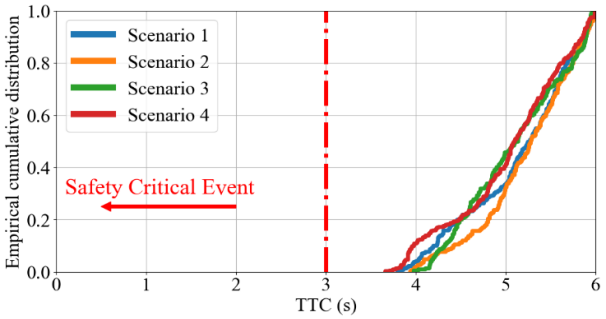
To demonstrate the superior capability of the proposed leading cruise controller to follow LHDV safely, efficiently, and stably, the three-vehicle trajectories generated in the four scenarios are compared. All comparison results are obtained based on the testing set. The testing process is similar to the training process which externally inputs the speed profile of LHDV. It starts by extracting car-following events from the testing set sequentially and performing the action selection for the agent with the learned model. The state of the three-vehicle platoon will be updated using Eqs. (1)-(6) till the

TABLE III
TRAFFIC FLOW PERFORMANCE ON THE WAYMO OPEN DATASET

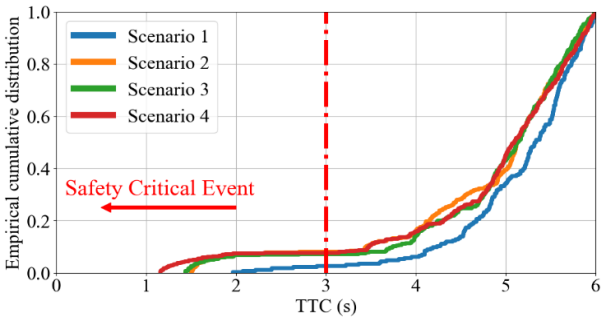
Scenarios	% of $TTC < 1s$		% of $TTC < 2s$		% of $TTC < 3s$		Average speed (m/s)		Average cumulative dampening ratios		Average Std of speed (m/s)	
	AV-LHDV	FHDV-AV	AV-LHDV	FHDV-AV	AV-LHDV	FHDV-AV	AV	FHDV	AV	FHDV	AV	FHDV
1	**0**	**0**	**0**	**0.45**	**0**	**2.69**	9.80	9.78	**0.93**	**0.58**	**1.26**	**1.16**
2	0	0	0	7.17	0	8.19	**9.83**	**9.81**	0.95	0.62	1.30	1.17
3	0	0	0	6.77	0	7.32	9.82	9.79	0.96	0.64	1.34	1.20
4	0	0	0	7.53	0	7.80	9.80	9.77	0.95	0.66	1.38	1.23

maximum time step of the LHDV speed profile.

1). Car-following safety: The study assesses car-following safety by examining critical TTC values in the testing set. Fig. 4 illustrates the empirical cumulative distributions of TTCs in the four scenarios. Upon visual inspection, it is apparent that scenarios 1, 2, 3, and 4 have no critical TTCs for AV-LHDV. Also, it is clear that scenario 1 is the safest for FHDV-AV out of all the scenarios as the blue line consistently remains below the other lines.



(a) AV-LHDV TTC



(b) FHDV-AV TTC

Fig. 4. TTC empirical cumulative distributions.

To quantitatively assess car-following safety under different scenarios, critical TTCs in each scenario are determined based on commonly-used thresholds including $1s$, $2s$, and $3s$, and presented in Table III. The findings indicate that scenario 1 consistently has the lowest proportion of critical TTCs for AV-LHDV and FHDV-AV, regardless of the threshold used. Scenarios 2, 3, and 4, on the other hand, have a comparable number of critical AV-LHDV TTCs but significantly more critical FHDV-AV TTCs than scenario 1. It is confirmed that considering FHDV safety in the AV control algorithm will not compromise AV safety.

To strengthen the statistical validity of the findings, the Kolmogorov-Smirnov test is conducted to compare the AV-LHDV and FHDV-AV TTC distributions in scenario 1 with those in the other scenarios. The results indicate that, except for AV-LHDV TTCs between scenarios 1 and 2 and scenarios 1 and 3, other comparison results are all statistically significant with $p - values$ lower than 0.05. This means that the proposed AV controller, which takes into account the benefits of AV and FHDV, can reduce the risks of rear-end crashes for both AV and FHDV, compared to AV controllers that do not consider FHDV.

2). Traffic efficiency: This study evaluates traffic efficiency by examining the average speeds. Table III shows the average speeds for AV and FHDV. The results indicate that scenarios 1-4 have similar average speeds whereas the Mann-Whitney U test also supports this finding. This is because scenarios 1-4 use DRL or MPC controllers to regulate the AV, which helps it follow LHDV closely.

3). String stability: This study uses the l_2-norm acceleration cumulative dampening ratio $d_{p,i}$ to quantify the string stability as below [39]:

$$d_{p,i} = \frac{\|a_i^t\|_2}{\|a_0^t\|_2} = \frac{(\Sigma_{t=0}^{N}|a_i^t|^2)^{\frac{1}{2}}}{(\Sigma_{t=0}^{N}|a_0^t|^2)^{\frac{1}{2}}} \tag{21}$$

where a_i^t is the acceleration of the i^{th} vehicle at the time step t; a_0^t is the acceleration of LHDV at the time step t; N is the car-following event duration. A lower dampening ratio $d_{p,i}$ means the vehicle has better capability to mitigate traffic perturbation, making it more string stable.

Table III shows the string stability for AV and FHDV, which is measured by the average cumulative dampening ratios. The results show that the average cumulative dampening ratios for AV and FHDV are the lowest in scenario 1, indicating that the proposed AV controller can also optimize the dampening performance of FHDV without compromising the string stability for AV.

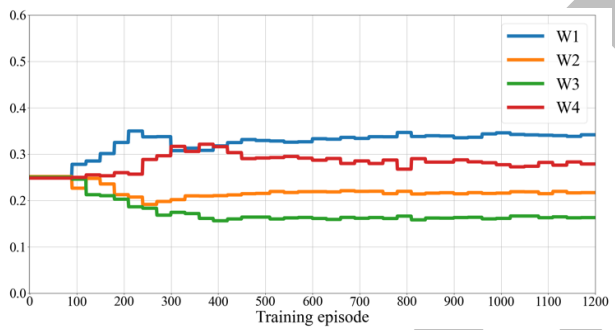
Apart from cumulative dampening ratios, average Std of speeds for AV and FHDV are also calculated and shown in Table III. One can observe that the proposed control strategy has lower average Std of speeds for AV and FHDV than the three baseline methods. This means that the proposed AV control model serves as a virtual regulator that refines FHDV driving decisions, ultimately reducing the traffic oscillations for the entire mixed traffic stream. Besides, compared to scenario 3, scenario 2 can lower the average Std of speed for AV by 3%, suggesting that the preference adjusting module can enhance the DRL-based AV controller. By adjusting the weights towards the human-preferred time gap, the time gap

TABLE IV
TRAFFIC FLOW PERFORMANCE ON THE NGSIM DATASET

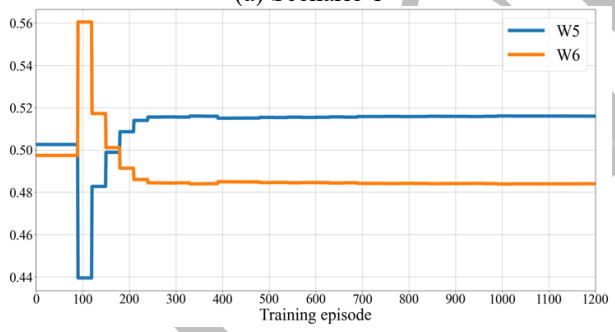
Scenarios	% of $TTC < 1s$		% of $TTC < 2s$		% of $TTC < 3s$		Average speed (m/s)		Average cumulative dampening ratios		Average Std of speed (m/s)	
	AV-LHDV	FHDV-AV	AV-LHDV	FHDV-AV	AV-LHDV	FHDV-AV	AV	FHDV	AV	FHDV	AV	FHDV
1	**0**	**0**	**0**	**0**	1.79	**1.67**	7.83	7.96	0.59	**0.33**	**1.45**	**1.41**
2	0	0	0	1.09	**0.78**	7.23	**7.85**	**7.97**	0.60	0.35	1.47	1.41
3	0	0	0	0.78	1.04	7.98	7.83	7.96	0.58	0.36	1.50	1.43
4	0	0	0	2.77	0.98	9.96	7.83	7.95	**0.55**	0.36	1.52	1.45

distribution becomes more centralized, resulting in the dampening of speed perturbations. This demonstrates the capability of incorporating expert knowledge to handle multiple objectives in car-following control.

In summary, the main findings include: (1) The AV controller proposed in scenario 1 enhances safety and string stability while maintaining comparable efficiency to other scenarios; (2) Scenario 2, which involves PbSAC with a preference adjusting module, is more efficient than scenario 3 (SAC) and scenario 4 (MPC) in improving the string stability of the AV; and (3) SAC demonstrates superior string stability compared to MPC.



(a) Scenario 1



(b) Scenario 2

Fig. 5. Convergence of the weights in the reward function.

E. Weight Learning Curves

Fig. 5 shows the convergence of the weight parameters during the model training. It is revealed that the proposed preference generator can efficiently determine the stable weight combination, even if there are four terms in the reward function in scenario 1. In addition, experimental results show that weighting rewards with preference weights does not significantly induce instability for policy improvement.

F. Generalizability Analysis

To evaluate the generalization capability of the proposed AV controller, this subsection randomly selects 200 three-HDV platoons from the reconstructed NGSIM dataset [40], each with a time length of over $15s$. In more detail, each speed profile of the three-HDV platoon leader is assigned to the leader of the simulated platoon. Then the initial states of AV and FHDV are set to match the two followers in the three-HDV platoon and controlled by their respective models, resulting in 200 simulated platoons under each scenario. After that, the states of AV and FHDV are updated using Eqs. (1)-(6).

The performance indicators of the mixed traffic stream for four scenarios are presented in Table IV. With regard to traffic safety, scenario 1 exhibits a slightly higher number of critical AV-LHDV TTCs compared to the other scenarios. However, the critical FHDV-AV TTCs in scenario 1 are significantly lower than those in scenarios 2, 3, and 4, which is supported by the Kolmogorov-Smirnov test. The speeds of AV and FHDV remain comparable across the different scenarios. As for string stability, scenario 1 shows the lowest average cumulative dampening ratios for FHDV and the lowest average Std of speeds for AV and FHDV. These analysis results provide evidence that the proposed AV controller can effectively adapt to different traffic conditions.

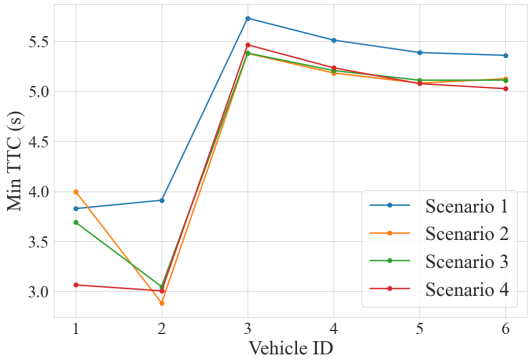
G. Platoon Analysis

This subsection will explore the AV controller's performance in a seven-vehicle platoon, with the second vehicle being the AV and the rest being HDVs. The Waymo Open Data comprises $20s$ segments, meaning that car-following events cannot exceed this duration. However, it is unlikely that speed fluctuations of LHDV can propagate through the entire platoon in $20s$. So, we again extract 100 vehicle trajectories from the NGSIM dataset that last over $30s$, assigning each as the platoon leader's trajectory. The AV and following HDVs will be controlled by corresponding models. The vehicle length and the initial time gap of two consecutive vehicles are fixed as $5m$ and $1.5s$, and all the initial speeds of the followers are equal to that of LHDV. As a result, there will be 100 simulated seven-vehicle platoons.

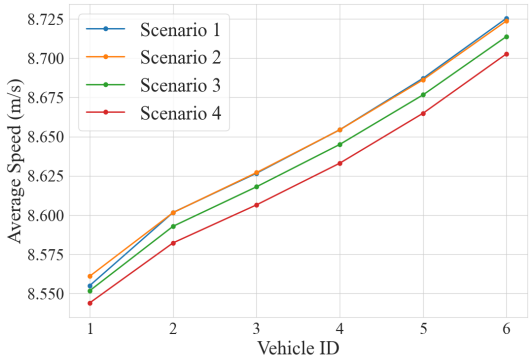
To evaluate the performance of individual vehicles within the platoon, we calculate the performance metrics for each vehicle across 100 platoons generated in various scenarios, as illustrated in Fig. 6. The findings demonstrate that scenario 1 outperforms the other scenarios in terms of safety and string stability. In particular, scenario 1 exhibits the highest minimum TTC for each vehicle, indicating better car-

TABLE V
TRAFFIC FLOW PERFORMANCE UNDER TWO FHDV CONTROL SETTINGS

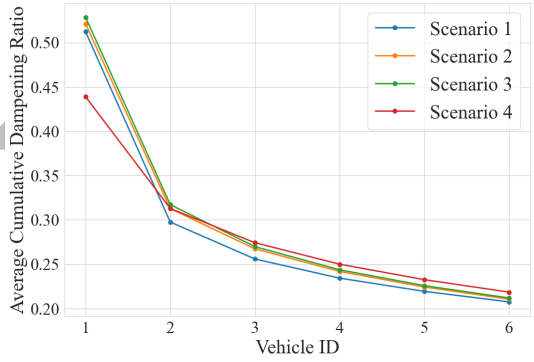
FHDV control setting	% of $TTC < 1s$		% of $TTC < 2s$		% of $TTC < 3s$		Average speed (m/s)		Average cumulative dampening ratios		Average Std of speed (m/s)	
	AV-LHDV	FHDV-AV	AV-LHDV	FHDV-AV	AV-LHDV	FHDV-AV	AV	FHDV	AV	FHDV	AV	FHDV
HDV-following-AV	**0**	**0**	**0**	**0.45**	**0**	**2.69**	**9.80**	**9.78**	**0.93**	**0.58**	**1.26**	**1.16**
HDV-following-HDV	0	0	0	2.44	0	18.29	9.77	9.71	0.97	0.63	1.28	1.16



(a) Minimum TTC



(b) Average speed



(c) Average cumulative dampening ratio

Fig. 6. Performance indicators of each vehicle in the platoon under different scenarios.

following safety. While the average speeds in scenarios 1 and 2 are similar, they are greater than those in scenarios 3 and 4. Furthermore, the average cumulative dampening ratios of each vehicle in scenario 1 are lower than those in scenarios 2 and 3, signifying a better capacity to mitigate traffic disturbances throughout the platoon. Despite vehicle 1 in scenario 4 having a lower average cumulative dampening ratio than vehicle 1 in other scenarios, it also has the lowest minimum TTC and average speeds. More importantly, the average cumulative dampening ratios for the following vehicles are still higher than those in scenario 1.

H. Comparison of FHDV Control Algorithms

In Section IV-B, it is highlighted that HDVs exhibit distinct car-following behaviors when they are following AVs versus following HDVs. In scenario 1, FHDV car-following behaviors are approximated using the HDV-following-AV model. Here we aim to understand how the car-following model settings of FHDV will affect the performance of the proposed AV controller. To achieve this, the HDV-following-AV model is replaced with the HDV-following-HDV model to control FHDV in scenario 1. The other settings, e.g., the AV control algorithm and the reward function, are identical to those in scenario 1.

Table V illustrates the performance of the proposed AV controller under two FHDV control settings based on the testing set. As can be seen, if car-following behaviors of HDVs are imitated through the HDV-following-AV model, the AV controller has better safety and string stability performance. For example, the average cumulative dampening ratio for AV in the HDV-following-AV scenario is 4.30% ($\frac{0.97-0.93}{0.97} * 100\% = 4.30\%$) lower than that in the HDV-following-HDV scenario. In addition, significance tests on the distributions of FHDV-AV TTC values and speeds for FHDV both exhibit significant differences between the two scenarios ($p - value < .01$). Hence, it can be concluded that the proposed approach using the HDV-following-AV model outperforms the counterpart. More importantly, it is revealed that during the development of AV controllers, if the responses of human drivers to the existence of AVs are not accurately modeled, the learned AV control policies may result in suboptimal performance.

V. CONCLUSION

In this study, a leading cruise controller for a mixed platoon of three vehicles is developed using PbSAC. Unlike previous studies that only focused on optimizing the AV, this approach also considers the benefits of FHDV. Real-world car-following trajectories from the Waymo Open Dataset are extracted, processed, and analyzed. An IRL approach is used to model the car-following behaviors of HDVs. To address the multi-objective car-following control problem, a PbSAC-based AV control model is introduced, which integrates SAC

with a preference generator. This preference generator can dynamically optimize multiple car-following objectives based on expert evaluation. Simulation results demonstrate that this approach can enhance the safety, efficiency, and string stability of the mixed traffic stream when compared to other AV controllers that do not consider FHDV. In addition, a comparative analysis is conducted to evaluate the impact of different FHDV car-following model settings on the performance of the proposed approach. The results reveal that using an HDV-following-AV model to mimic HDV car-following behaviors can achieve better performance than an HDV-following-HDV model.

This paper has practical implications for various sectors related to AVs. It emphasizes the significance of considering human driver behavioral adaptations when studying AV-involved traffic flow, which can be valuable for transportation researchers. For road engineers, it may drive them to propose appropriate traffic management policies to take advantage of AVs when AVs are publicly available. For AV manufacturers, it elucidates the potential for designing advanced AV control algorithms that can also consider the surrounding vehicles and thus benefit the mixed traffic flow, especially in situations where AVs on public roads currently lack communication technologies.

This study has three major limitations. First, the proposed controller utilizes the HDV-following-AV model obtained from the Waymo Open Dataset. It is appealing to assess the controller if other autonomous driving datasets can be leveraged to calibrate the HDV-following-AV model. Second, the AV control algorithm has only been tested through simulation. However, it can be further evaluated for its effectiveness through the use of driving simulators, test tracks, or even public roads in the future. Third, DRL-based methods incorporating prior knowledge have shown potential in addressing multi-objective car-following control. Further research is necessary to improve control efficiency through the development of a more informative evaluation.

REFERENCES

[1] L. Chen, *et al.,* "Milestones in autonomous driving and intelligent vehicles: Survey of surveys," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 2, pp. 1046-1056, Nov. 2022.

[2] Y. Huang, Y. Gu, K. Yuan, S. Yang, T. Liu, and H. Chen, "Human Knowledge Enhanced Reinforcement Learning for Mandatory Lane-Change of Autonomous Vehicles in Congested Traffic," *IEEE Transactions on Intelligent Vehicles*, early access.

[3] X. Wen, Z. Cui, and S. Jian, "Characterizing car-following behaviors of human drivers when following automated vehicles using the real-world dataset," *Accident Analysis & Prevention*, vol. 172, Jul. 2022, Art. no. 106689.

[4] X. Hu, Z. Zheng, D. Chen, X. Zhang, and J. Sun, "Processing, assessing, and enhancing the Waymo autonomous vehicle open dataset for driving behavior research," *Transportation Research Part C: Emerging Technologies*, vol. 134, Jan. 2022, Art. no. 103490.

[5] X. Wen, S. Jian, and D. He, "Modeling Human Driver Behaviors When Following Autonomous Vehicles: An Inverse Reinforcement Learning Approach," in *Proc. 25th International Conference on Intelligent Transportation Systems (ITSC)*, 2022, pp. 1375-1380.

[6] J. Zhu, I. Tasic, and X. Qu, "Flow-level coordination of connected and autonomous vehicles in multilane freeway ramp merging areas. Multimodal transportation," *Multimodal Transportation*, vol. 1, no. 1, Mar. 2022.

[7] X. Wen, S. Jian, and D. He, "Modeling the Effects of Autonomous Vehicles on Human Driver Car-Following Behaviors using Inverse Reinforcement Learning," *IEEE Transactions on Intelligent Transportation Systems*, Aug. 2023.

[8] X. Wen, C. Huang, S. Jian, and D. He, "Analysis of discretionary lane-changing behaviours of autonomous vehicles based on real-world data," *Transportmetrica A: Transport Science*, Nov. 2023.

[9] H. Shi, Y. Zhou, X. Wang, S. Fu, S. Gong, and B. Ran, "A deep reinforcement learning-based distributed connected automated vehicle control under communication failure," *Computer-Aided Civil and Infrastructure Engineering*, vol. 37, pp. 2033–2051, Dec. 2022.

[10] H. Shi, Y. Zhou, K. Wu, X. Wang, Y. Lin, and B. Ran, "Connected automated vehicle cooperative control with a deep reinforcement learning approach in a mixed traffic environment," *Transportation Research Part C: Emerging Technologies*, vol. 133, Dec. 2021, Art. no. 103421.

[11] Y. Zhou, S. Ahn, M. Chitturi, and D. Noyce, "Rolling horizon stochastic optimal control strategy for ACC and CACC under uncertainty," *Transportation Research Part C: Emerging Technologies*, vol. 83, pp. 61-76, Oct. 2017.

[12] Y. Du, J. Chen, C. Zhao, C. Liu, F. Liao, and C.-Y. Chan, "Comfortable and energy-efficient speed control of autonomous vehicles on rough pavements using deep reinforcement learning," *Transportation Research Part C: Emerging Technologies*, vol. 134, Jan. 2022, Art. no. 103489.

[13] M. Zhu, Y. Wang, Z. Pu, J. Hu, X. Wang, and R. Ke, "Safe, efficient, and comfortable velocity control based on reinforcement learning for autonomous driving," *Transportation Research Part C: Emerging Technologies*, vol. 117, Aug. 2020, Art. no. 102662.

[14] J. Zhang, C. Chang, X. Zeng, and L. Li, "Multi-Agent DRL-Based Lane Change With Right-of-Way Collaboration Awareness," *IEEE Transactions on Intelligent Transportation Systems*, Oct. 2022.

[15] J. Wang, and L. Sun, "Multi-objective multi-agent deep reinforcement learning to reduce bus bunching for multiline services with a shared corridor," *Transportation Research Part C: Emerging Technologies*, vol. 155, Oct. 2023, Art. no. 104309.

[16] X. Di and R. Shi, "A survey on autonomous vehicle control in the era of mixed-autonomy: From physics-based to AI-guided driving policy learning," *Transportation Research Part C: Emerging Technologies*, vol. 125, Apr. 2021, Art. no. 103008.

[17] Y. Rahmati, M. Khajeh Hosseini, A. Talebpour, B. Swain, and C. Nelson, "Influence of Autonomous Vehicles on Car-Following Behavior of Human Drivers," *Transportation Research Record*, vol. 2673, no. 12, pp. 367–379, Dec. 2019.

[18] I. Mahdinia, A. Mohammadnazar, R. Arvin, and A. J. Khattak, "Integration of automated vehicles in mixed traffic: Evaluating changes in performance of following human-driven vehicles," *Accident Analysis & Prevention*, vol. 152, Mar. 2021, Art. no. 106006.

[19] G. Naus, R. Vugts, J. Ploeg, M. Molengraft, and M. Steinbuch, "String-stable CACC design and experimental validation: A frequencydomain approach," *IEEE Transactions on vehicular technology*, vol. 59, pp. 4268–4279, Nov. 2010.

[20] F. Morbidi, P. Colaneri, and T. Stanger, "Decentralized optimal control of a car platoon with guaranteed string stability," in *Proc. 2013 European Control Conference (ECC)*, 2013, pp. 3494–3499.

[21] L. Zhang and G. Orosz, "Consensus and disturbance attenuation in multi-agent chains with nonlinear control and time delays," *International Journal of Robust and Nonlinear Control*, vol. 27, no. 5, pp. 781–803, Mar. 2017.

[22] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Physical review E*, vol. 62, Aug. 2000, Art. no. 1805.

[23] S. Gong, J. Shen, and L. Du, "Constrained optimization and distributed computation based car following control of a connected and autonomous vehicle platoon," *Transportation Research Part B: Methodological*, vol. 94, pp. 314-334, Dec. 2016.

[24] M. Wang, W. Daamen, S.P. Hoogendoorn, and B. van Arem, "Cooperative car-following control: Distributed algorithm and impact on moving jam features," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 5, pp. 1459-1471, May 2016.

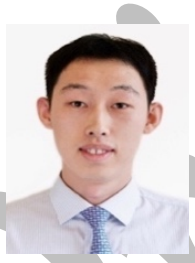
[25] X. Qu, Y. Yu, M. Zhou, C.-T. Lin, and X. Wang, "Jointly dampening traffic oscillations and improving energy consumption with electric, connected and automated vehicles: A reinforcement learning based approach," *Applied Energy*, vol. 257, Jan. 2020, Art. no. 114030.

[26] T.-Q. Tang, Y. Gui, and J. Zhang, "ATAC-Based Car-Following Model for Level 3 Autonomous Driving Considering Driver's Acceptance,"

IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 8, pp. 10309–10321, Aug. 2022.
[27] Y. Yan, L. Peng, T. Shen, J. Wang, D. Pi, D., Cao, and G. Yin, "A Multi-Vehicle Game-Theoretic Framework for Decision Making and Planning of Autonomous Vehicles in Mixed Traffic," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 11, pp. 4572–4587, Nov. 2023.
[28] X. He, H. Yang, Z. Hu, and C. Lv, "Robust lane change decision making for autonomous vehicles: An observation adversarial reinforcement learning approach," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp.184-193, Apr. 2022.
[29] Y. Lin, J. McPhee, and N.L. Azad, "Comparison of deep reinforcement learning and model predictive control for adaptive cruise control," *IEEE Transactions on Intelligent Vehicles*, vol. 6, no. 2, pp. 221-231, Jul. 2020.
[30] L. Jiang, Y. Xie, N. G. Evans, X. Wen, T. Li, and D. Chen, "Reinforcement Learning based cooperative longitudinal control for reducing traffic oscillations and improving platoon stability," *Transportation Research Part C: Emerging Technologies*, vol. 141, Aug. 2022, Art. no. 103744.
[31] X. Shi and X. Li, "Empirical study on car-following characteristics of commercial automated vehicles with different headway settings," *Transportation Research Part C: Emerging Technologies*, vol. 128, Jul. 2021, Art. no. 103134.
[32] M. Peschl, A. Zgonnikov, F.A. Oliehoek, and L.C. Siebert. "MORAL: Aligning AI with human norms through multi-objective reinforced active learning," in *Proc. 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2022, pp. 1038–1046.
[33] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. International conference on machine learning (ICML)*, 2018, pp. 1861–1870.
[34] P. Sun, K. Henrik, D. Xerxes, C. Aurelien, P. Vijaysai, T. Paul, G. James et al., "Scalability in perception for autonomous driving: Waymo open dataset," in *Proc. IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2020, pp. 2446-2454.
[35] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. Qi, Y. Zhou et al., "Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset," in *Proc. IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2021, pp. 9710-9719.
[36] V. Punzo, Z. Zheng, and M. Montanino, "About calibration of car-following dynamics of automated and human-driven vehicles: Methodology, guidelines and codes," *Transportation Research Part C: Emerging Technologies*, vol. 128, Jul. 2021, Art. no. 103165.
[37] V. Punzo, M. Montanino, and B. Ciuffo, "Do we really need to calibrate all the parameters? Variance-based sensitivity analysis to simplify microscopic traffic flow models," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, Jul. 2014, Art. no. 1.
[38] D. Garg, S. Chakraborty, C. Cundy, J. Song, and S. Ermon, "IQ-Learn: Inverse soft-Q Learning for Imitation," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2021, pp. 4028-4039.
[39] J. Ploeg, N. Van De Wouw, and H. Nijmeijer, "Lp string stability of cascaded systems: Application to vehicle platooning," *IEEE Transactions on Control Systems Technology*, vol. 22, no. 2, pp.786-793, May. 2013.
[40] M. Montanino, and V. Punzo, "Trajectory data reconstruction and simulation-based validation against macroscopic traffic patterns," *Transportation Research Part B: Methodological*, vol. 80, pp. 82-106, Oct. 2015.

Xiao Wen received a B.S. degree in Transportation Engineering from Beijing Jiaotong University and a M.S. degree in Civil Engineering from University of Washington. He is currently pursuing his Ph.D. degree at The Hong Kong University of Science and Technology. His research interests include transportation data analysis, machine learning, causal inference, autonomous driving, and traffic safety.

Xinhu Zheng is currently an Assistant Professor in the Intelligent Transportation Thrust of the Systems Hub, at Hong Kong University of Science and Technology (GZ). He received his Ph.D. degree in Electrical and Computer Engineering from the University of Minnesota, Minneapolis, in 2022. He is currently an Associated Editor for IEEE Transactions on Intelligent Vehicles. His current research interests including multi-agent information fusion, multi-modal data fusion and data analysis in intelligent transportation system and ITS related intelligent systems, by exploiting different modality of data, leveraging optimization and machine learning techniques.



Zhiyong Cui is a Professor in the School of Transportation Science and Engineering at Beihang University. He was selected into the National Natural Science Foundation of China Outstanding Youth (Overseas) Fund Project. He was a Data Science Postdoctoral Fellow at the University of Washington. He received Ph.D. degree in Civil Engineering in 2021 at UW. He received M.S. and B.S. degrees both in software engineering from Peking University in 2015 and Beihang Univerity in 2012, respectively. His research mainly focusses on transportation data science, artificial intelligence, traffic prediction and control.

Sisi Jian received the Ph.D. degree in transport engineering from the University of New South Wales. She is currently an Assistant Professor with the Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology. Her major research interests are transportation network modeling, game theory and bi-level optimization.

Dengbo He received his bachelor's degree from Hunan University in 2012, M.S. degree from the Shanghai Jiao Tong University in 2016 and Ph.D. degree from the University of Toronto in 2020. He is currently an assistant professor from the Intelligent Transpiration Thrust and Robotics and Autonomous Systems Thrust, the Hong Kong University of Science and Technology (Guangzhou). He is also affiliate with the Department of Civil and Environmental Engineering, the Hong Kong University of Science and Technology and HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Futian, Shenzhen. His research interests include human factors, driver behavior analysis and driver state estimation.