Modeling Human-Like Car-Following Behavior: A GRPO Approach for Region-Specific Adaptation

Yang Liu[a], Dengbo He[a, b]*
[a] Intelligent Transportation Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China; [b] Robotics and Autonomous Systems Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China

ABSTRACT

Among various driving scenarios, car-following (CF) is the most fundamental and frequent behavior in traffic and is crucial for maintaining traffic stability. Thus, designing a human-like car-following model for automated driving is essential, as it enables other road users to better understand, predict, and accept the behavior of such vehicles in mixed traffic environments. In this study, we developed a Generative Reinforcement Proximal Optimization (GRPO) model to imitate human drivers' decision-making processes. Experimental results demonstrate that the GRPO model more effectively captured human-like following behavior compared to traditional physical and reinforcement learning (RL) models in two datasets from two countries (NGSIM and CN-truck dataset). This research is the first to apply the novel GRPO framework to CF modeling. The results demonstrate that the superior performance of the GRPO as compared to baseline models and indicate its feasibility to generate human-like CF behaviors in mixed traffic environments.

Keywords: mixed traffic, car-following, GRPO, human-like driving behavior

1. INTRODUCTION

With the rapid advancement of intelligent transportation technologies, the commercialization of autonomous vehicles (AVs) is around the corner. Many countries have begun pilot operations or enacted regulatory frameworks to facilitate the deployment of AVs. Despite these advancements, AVs remain relatively unfamiliar to most drivers [1], and drivers may take different strategies to interact with AVs on the road [2], which may affect the safety and efficiency of the mixed traffic consisting of AVs and human-driven vehicles (HDVs). Such a difference in HDV's strategies may stem from the different strategies AVs take [3]. Therefore, enhancing the behavioral similarity between AVs and human drivers is important for improving the safety and efficiency of mixed traffic.

Among various driving behaviors, car-following (CF) behavior is one of the most fundamental and frequent driving tasks in traffic [4]. Thus, various CF models have been designed to replicate human driving behaviors. Among these, the Imitation Learning (IL) has been widely used, as it can imitate expert trajectories to recover complex car-following strategies with minimal manual modeling. However, previous IL methods can suffer from demonstration bias and distribution shift when encountering unseen states. Thus, in this study, we adopted a new approach of IL, the Group Relative Policy Optimization (GRPO)[5], to capture nuanced human-like decision-making patterns in CF events. Specifically, the GRPO optimizes the

* dengbohe@hkust-gz.edu.cn

car-following policy by comparing each sample to a group baseline, which helps reduce bias and improve generalization to unseen situations.

To comprehensively evaluate model performance, we compare the GRPO-based CF models against several widely used baselines, including Intelligent Driver Model (IDM) [6], Proximal Policy Optimization (PPO) [7], Soft Actor-Critic (SAC) [8], and Generative Adversarial Imitation Learning (GAIL) [9]. To more comprehensively evaluate the model performance, two datasets—NGSIM from the United States [10] and a self-collected highway traffic dataset from China (CN-Truck Dataset)—were used for training and evaluation.

2. METHODS

Problem Formulation

The CF process can be recognized as an MDP. [11] In MDP, the object we aim to control is called an agent, which is influenced by an environment. The interaction between the agent and the environment is called the action. Five elements can describe an MDP: i) state, S, which is a description of the environment. ii) action, A, which is a collection of all possible movements (a) that an agent can perform in each state. iii) transition probability, $P(s'|s,a)$, which describes the probability of moving to the next state s', given the current state s and the action a performed. iv) Reward, R, measures the benefit of an action taken by the agent in the environment. v) and discount factor, γ, which is between 0 and 1, describing the importance of future rewards.

In our study, the agent of the reinforcement learning (RL) model is the ego-vehicle, and the agent's action included acceleration and deceleration. The states (S) can be defined as a vector $[\Delta y_{le}, v_e, v_l, \Delta v_{le}]$, composed of relative distance Δy_{le} between lead vehicle (LV) and ego-vehicle, ego-vehicle velocity v_e, LV velocity v_l, and the relative speed Δv_{le} between LV and ego-vehicle. According to Newton's kinematic laws, the physical relationships among these parameters follow the equations:

$$v_e(t+1) = v_e(t) + a(t) \times \Delta T \tag{1}$$

$$y_e(t+1) = y_e(t) + \frac{1}{2}[v_e(t) + v_e(t+1)] \times \Delta T \tag{2}$$

$$\Delta v_{le}(t+1) = v_l(t+1) - v_e(t+1) \tag{3}$$

$$\Delta y_{le}(t+1) = y_l(t+1) - y_e(t+1) \tag{4}$$

GRPO

Building upon the Proximal Policy Optimization (PPO) [7] framework, the GRPO [5] algorithm extends policy learning to multi-agent or grouped environments by introducing a group-relative advantage term. The objective function of GRPO can be expressed as:

$$argmax_\theta E_{s\sim v^{\pi_\theta}} E_{a\sim\pi_\theta(\cdot|s)} \left[\frac{1}{N}\sum_{i=1}^{N} min\left(r_{i,t}(\theta)\tilde{A}_i^{grp}(s,a_i), clip\big(r_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon\big)\tilde{A}_i^{grp}(s,a_i)\right)\right] \tag{5}$$

* dengbohe@hkust-gz.edu.cn

Where $r_{i,t}(\theta)$ is $\frac{\pi_\theta(a|s)}{\pi_k(a|s)}$, denoting the probability ratio between the updated and old policies, and $\tilde{A}_i^{grp}(s,a_i)$ is $A_i(s,a_i) - \frac{1}{N}\sum_{j=1}^{n} A_j(s,a_j)$, which represents the group-relative advantage.

Instead of optimizing each agent's advantage (A_i) independently, GRPO encourages the policies to be better than their peers, rather than merely improving upon past versions. This relative optimization enhances inter-agent coordination, robustness, and learning stability in both cooperative and competitive settings. As shown in Figure 1, GRPO can be adapted to CF scenarios by treating different driving samples as a group. In this context, the group-relative advantage measures how well the ego-vehicle's policy performs relative to the average behavior across comparable situations. Consequently, GRPO enables the ego-vehicle to learn driving strategies that are not only individually effective but also robust and relatively superior across various CF contexts. During training, the stability of the learning process was evaluated by measuring the discrepancy between generated and real-world trajectories, which was treated as the reward in GRPO:

$$r = log\left(\frac{V_{et}-V_{em}}{V_{et}}\right) - 1000 \times c \tag{6}$$

Where V_{et} was the true driving speed of ego-vehicle in the CF segment, V_{em} was the velocity of the ego-vehicle generated by the model, and c was 1 if a collision occurred (otherwise 0). This equation was also the reward function for all the reinforcement learning in this paper.

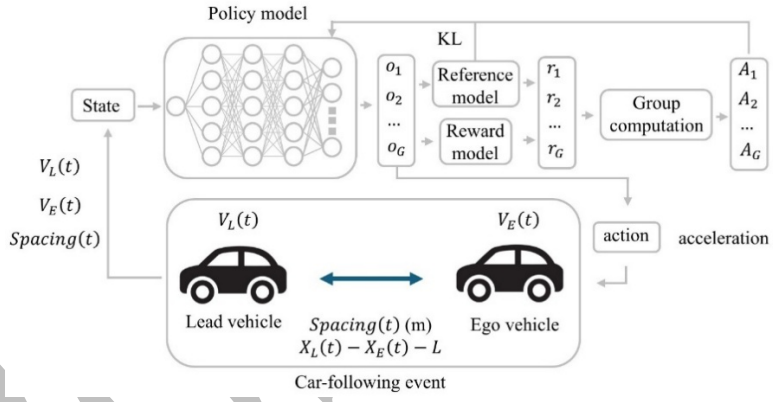


Fig. 1. The structure of GRPO in the CF scenarios. In the figure, the reference model is the old policy from the previous round of training, the reward model is described in Eq. 6, and KL is the Kullback–Leibler divergence.

Benchmark models

A. Intelligent Driver Model

Microscopic CF models have long been used to simulate individual driver behavior in the presence of LVs, of which the Intelligent Driver Model (IDM) [6] is one of the most used:

$$a_n(t) = a_{max}^{(n)}\left(1 - \left(\frac{V_n(t)}{\widetilde{V_n}(t)}\right)^{\beta} - \left(\frac{\widetilde{S_n}(t)}{S_n(t)}\right)^2\right) \tag{7}$$

* dengbohe@hkust-gz.edu.cn

$$\widetilde{S_n}(t) = S_{jam}^{(n)} + max\left(0, V_n(t)\widetilde{T_n}(t) + \frac{V_n(t)\Delta V_n(t)}{2\sqrt{a_{max}^{(n)} a_{comfort}^{(n)}}}\right) \quad (8)$$

In Eq. 7 and 8, $a_{max}^{(n)}$ is the maximum acceleration, $\widetilde{V_n}(t)$ is the desired speed, $\widetilde{S_n}(t)$ is the desired gap, β is usually set as 4, $S_{jam}^{(n)}$ is the minimum standstill gap, $V_n(t)$ is the current speed, and $a_n(t)$ is the acceleration.

B. Proximal Policy Optimization

The PPO algorithm [7] improves policy gradient methods by constraining the size of policy updates. Instead of making large, unstable changes to the policy, PPO uses a clipped objective function to keep the new policy $\pi_\theta(a \mid s)$ close to the old policy $\pi_k(a \mid s)$. This balances exploration and stability, ensuring efficient and reliable learning by maximizing the expected advantage, while preventing overly aggressive policy shifts. The objective equation for PPO is shown below.

$$argmax_\theta E_{s\sim v^{\pi_\theta}} E_{a\sim\pi_\theta(\cdot|s)}\left[min\left(r_{i,t}(\theta)A^{\pi_\theta}(s,a), clip\big(r_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon\big)A^{\pi_{\theta_k}}(s,a)\right)\right] \quad (9)$$

In Eq. 9, θ is the trainable policy parameter, $v^{(\pi_\theta)}$ is the state-visitation distribution induced by the current policy, $A^{(\pi_\theta)}(s,a)$ is the advantage under the current policy, $A^{(\pi_{\theta_k})}(s,a)$ is the advantage under the previous policy, and ε is the clip rate, which was set as 0.2 in this paper.

C. Generative Adversarial Imitation Learning

The Generative Adversarial Imitation Learning (GAIL) [9] is based on the Generative Adversarial Network (GAN) [12]. GAN consists of two main components: a generator and a discriminator. These two networks compete with each other through adversarial learning to improve the performance of the generator model. When using GAIL, firstly, a real-world trajectory from the dataset was randomly selected, and the state of the trajectory was initialized. Then, the initial state and LV data were input into the generator, and the PPO algorithm generated a trajectory based on this data. The discriminator optimized itself by discriminating between the generated trajectory and the real-world trajectory. The loss function of the discriminator was as follows:

$$L(\psi) = -E_{\rho\pi}\left[logD_\psi(s,a)\right] + E_{\rho E}\left[log\left(1 - D_\psi(s,a)\right)\right] \quad (10)$$

The pseudo-rewards that the discriminator provides were: $R_{GAIL} = -logD(s,a)$. ψ is the parameter of the discriminator network, $D_\psi(s,a)$ is the discriminator's estimated probability that a state–action pair comes from the expert, ρ_π is the occupancy measure induced by the policy to be learned, and ρ_E is the occupancy measure of the expert demonstrations.

D. Soft Actor-Critic

Soft Actor-Critic (SAC) [8] is a policy gradient algorithm based on Maximum Entropy Reinforcement Learning (Maximum Entropy RL) [13], aiming to enhance the cumulative return while maintaining the randomness of the policy, thereby achieving better exploration and stability. The objective equation for SAC is shown below.

$$J(\pi) = E\left[\sum_t r(s_t,a_t) + \alpha H\big(\pi(\cdot\,|s_t)\big)\right] \quad (11)$$

* dengbohe@hkust-gz.edu.cn

In the equation, $r(s_t, a_t)$ is the reward obtained by taking action a_t in state s_t, $H(\pi(\cdot|\, s_t))$ is the entropy of the policy at state s_t, and α is the temperature coefficient that balances reward maximization and exploration.

3. DATASETS FOR MODEL EVALUATION

NGSIM Dataset and CN-Truck Dataset

NGSIM is a dataset on public transportation in the United States, collected by the U.S. Department of Transportation Intelligent Transportation Systems Joint Program Office (JPO) in 2005-2006 [14]. It is the largest and most widely used highway traffic trajectory dataset globally and was recorded by installing seven cameras on a 30-story building near a highway. The data used in this paper were collected at Interstate 80 (I-80) in Emeryville, CA, which covers a 500m-long high-occupancy vehicle (HOV) lane.

The CN-Truck Dataset was collected by installing advanced sensing equipment (LiDAR and video cameras) on trucks running between Wuxi, Jiangsu, and Qingdao, Shandong, with an average one-way distance of 830 km and a duration of 10 hours. The dataset was collected between November 21, 2024, and November 27, 2024, and the perception range of the sensors was 300 meters at a frequency of 10 Hz.

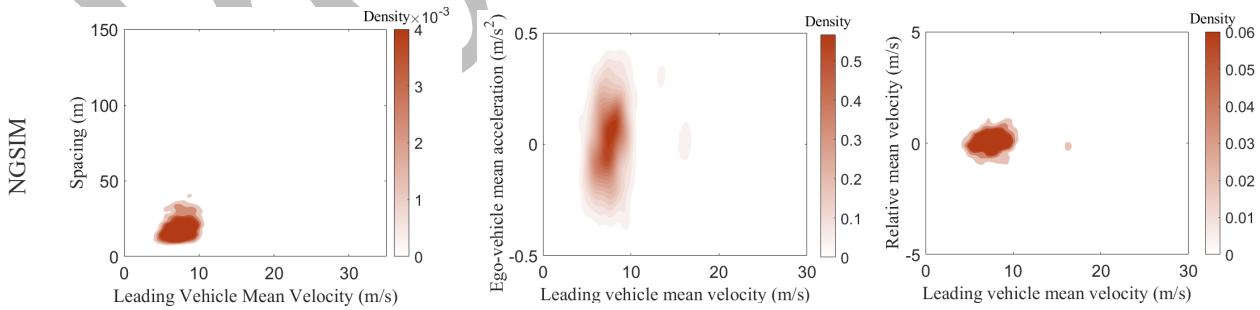
Extraction of the Car Following Segments

Following previous research [15,16], we adopted the following criteria to extract the CF segments: 1) lead and ego vehicles need to be in the same lane; 2) lead and ego vehicles should be within 100m; 3) the speed of both the lead and ego vehicles should be greater than 10km/h to avoid traffic jams; and 4) the duration of the CF segments should be over 15s.

According to the above criteria, we extracted a total of 999 CF segments from NGSIM and 354 CF segments from CN-Truck. When training the data, 70% of each dataset was randomly selected as the training set, while the remaining 30% was used as the test set.

Heterogeneity of the Datasets

To evaluate whether the models can consistently perform well and mimic diverse CF behaviors, we need to ensure that the CF behaviors differ between the two datasets. Thus, we compared CF characteristics in CN-Truck and NGSIM. As illustrated in Figure 2, substantial differences in driving styles can be observed. This indicates that driving behavior is region-dependent, underscoring the feasibility of using these two datasets for model evaluation.



* dengbohe@hkust-gz.edu.cn

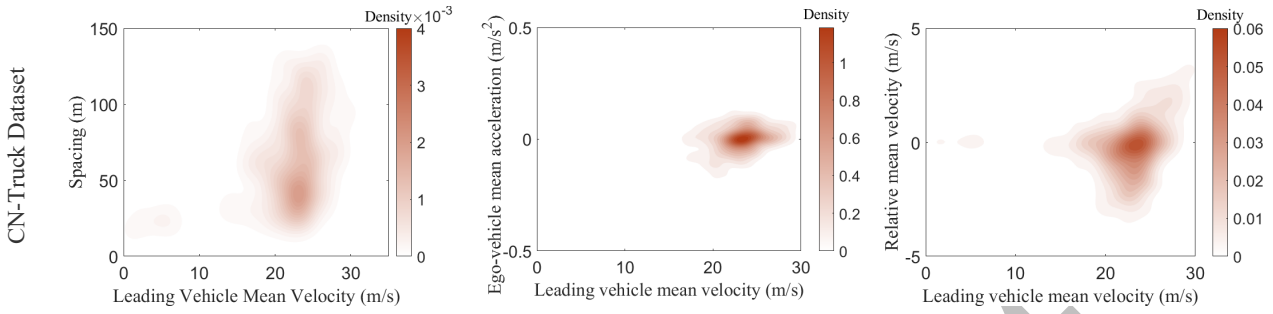
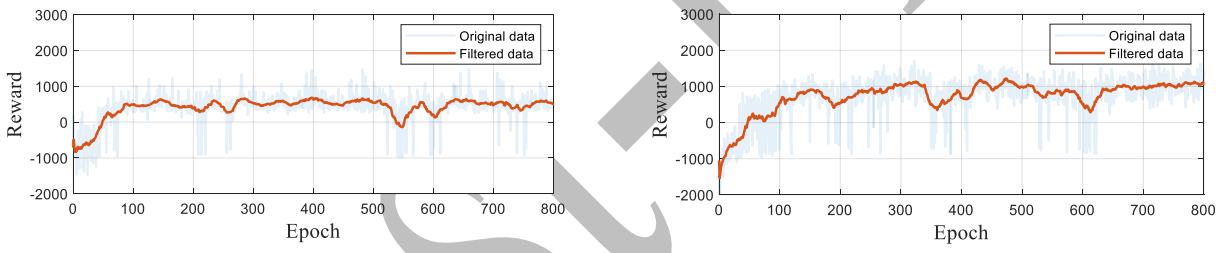


Fig. 2. Driver behavior density heatmap for CF segments. The color intensity represents sample density, with darker colors indicating a higher density.

4. RESULTS

As shown in Figure 3, the GRPO algorithm converged on both datasets after approximately 100 training epochs, indicating overall training stability. It is also observed that the average reward on the CN-Truck dataset was slightly higher, primarily due to the longer duration of its CF episodes, which leads to higher cumulative rewards.

Then, the performance of different models in imitating the ego-vehicle's speed and spacing strategies is illustrated in Figures 4 (NGSIM) and 5 (CN-Truck). It can be observed that the GRPO model achieved superior performance in replicating both spacing and velocity profiles, demonstrating a strong capability for human-like behavior imitation.



Ngsim CN-Truck Dataset

Fig. 3. GRPO model training return change trend

Tables 1 provide quantitative comparisons of the models on the test sets. In the table, MSE spacing and MSE velocity represent the mean squared errors between the generated and real-world trajectories in spacing and velocity, respectively. The jerk metric measures the change rate of acceleration, reflecting the smoothness of the driving process and the comfort. TTC (Time-To-Collision) is a widely used surrogate safety measure, originally introduced by Vogel [17], and is defined as:

$$TTC(t) = \frac{X_L(t) - X_F(t) - L}{V_F(t) - V_L(t)} \tag{13}$$

where $X_L(t)$ and $X_F(t)$ denote the positions of the centers of mass of the LV and ego-vehicles; L is the vehicle length; and $V_L(t)$ and $V_F(t)$ are their respective velocities. The results show that GRPO consistently achieved the lowest MSE in spacing across all datasets, highlighting its effectiveness in reproducing human-like following distances. Further, most RL-based models exhibited relatively high jerk values, as reflected by the noticeable oscillations in the velocity profiles. In contrast, the IDM model performed better in terms of smoothness. No difference was observed among the models for the TTC metric, which aligns with the focus on behavior imitation rather than safety optimization.

* dengbohe@hkust-gz.edu.cn

Overall, the findings demonstrate that the GRPO model excelled at capturing driver-like CF behavior. Nevertheless, the relatively high jerk values suggest that future work should incorporate jerk and TTC into the reward design to achieve smoother and safer human-like CF models that go beyond merely imitating driver behavior.

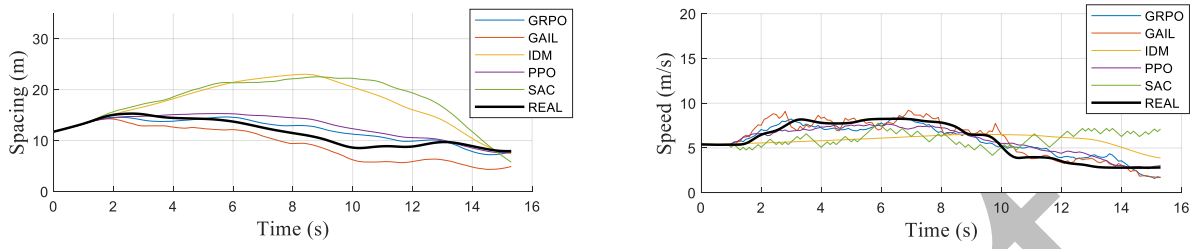


Fig. 4. CF model performance for NGSIM dataset.

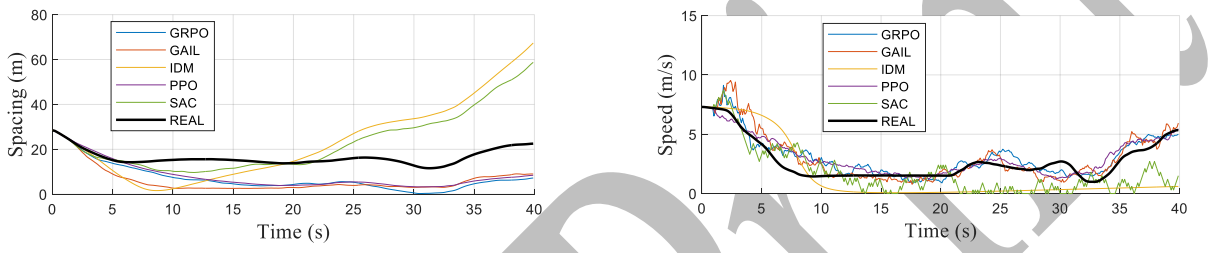


Fig. 5. CF model performance for CN-Truck Dataset.

Table 1 Comparison of model performance on NGSIM and CN-Truck datasets. The bolded texts mean optimal result.

Model Test / standard	NGSIM					CN-Truck Dataset				
	GRPO	GAIL	IDM	PPO	SAC	GRPO	GAIL	IDM	PPO	SAC
MSE (spacing)	**25.73**	26.61	30.03	30.88	36.43	**33.88**	64.14	47.95	43.73	42.57
MSE (velocity)	**1.02**	2.50	1.98	2.88	3.31	2.56	6.00	**1.32**	3.35	5.23
Jerk	16.07	18.54	**0.06**	8.01	15.65	22.22	50.9	**0.02**	20.69	21.59
TTC	4.70	4.61	4.79	**4.99**	4.32	6.22	4.54	4.63	**7.08**	3.05

Note: In the table, the bolded text indicates the best model for the corresponding metric.

5. CONCLUSIONS

In this study, we revealed the superior performance of GRPO by modeling CF strategies in the well-known NGSIM dataset and naturalistic driving data collected on Chinese highways. First, by comparing heatmaps of driving behaviors across different datasets, the differences in CF styles across two countries are revealed. Second, CF models were developed using the advanced GRPO algorithm and commonly used CF models, aiming to mimic human drivers' CF behaviors. Experimental results demonstrated that the GRPO model performed the best in replicating driver-like CF trajectories and capturing CF dynamics. However, similar to other RL-based approaches (i.e., GAIL, PPO and SAC), the GRPO model exhibited relatively poor performance in terms of smoothness of speed control (i.e., Jerk). The GRPO still failed to yield the safest CF patterns in terms of the TTC. Future work should incorporate jerk and TTC metrics into the reward function to develop more comfortable and safer CF algorithms.

* dengbohe@hkust-gz.edu.cn

ACKNOWLEDGEMENTS

This work was supported by the Scientific Research Projects for the Higher-educational Institutions of Guangzhou (No. 2024312135). The authors would also like to express their sincere gratitude to Ms. Didan Hu (hudidan@163.com) for her valuable assistance in data processing and partial model generation during the preparation of this paper.

REFERENCES

[1] Omeiza, Daniel, et al. "Explanations in autonomous driving: A survey." IEEE Transactions on Intelligent Transportation Systems 23.8 (2021): 10142-10162.

[2] Huang, Chunxi, et al. "Sharing the road: how human drivers interact with autonomous vehicles on highways." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 66. No. 1. Sage CA: Los Angeles, CA: SAGE Publications, 2022.

[3] He, Yaqin, Dingshan Xiang, and Daobin Wang. "Traffic safety evaluation of emerging mixed traffic flow at freeway merging area considering driving behavior." *Scientific Reports* 15.1 (2025): 10686.

[4] Zhou, Anye, et al. "Car-following behavior of human-driven vehicles in mixed-flow traffic: A driving simulator study." IEEE Transactions on Intelligent Vehicles 8.4 (2023): 2661-2673.

[5] Shao, Zhihong, et al. "Deepseekmath: Pushing the limits of mathematical reasoning in open language models." arXiv preprint arXiv:2402.03300 (2024).

[6] Treiber, Martin, Ansgar Hennecke, and Dirk Helbing. "Congested traffic states in empirical observations and microscopic simulations." Physical review E 62.2 (2000): 1805.

[7] Schulman, John, et al. "Proximal policy optimization algorithms." arXiv preprint arXiv:1707.06347 (2017).

[8] Haarnoja, Tuomas, et al. "Soft actor-critic algorithms and applications." arXiv preprint arXiv:1812.05905 (2018).

[9] Ho, Jonathan, and Stefano Ermon. "Generative adversarial imitation learning." Advances in neural information processing systems 29 (2016).

[10] Coifman, Benjamin, and Lizhe Li. "A critical evaluation of the Next Generation Simulation (NGSIM) vehicle trajectory dataset." Transportation Research Part B: Methodological 105 (2017): 362-377.

[11] Puterman, Martin L. "Markov decision processes." Handbooks in operations research and management science 2 (1990): 331-434.

[12] Creswell, Antonia, et al. "Generative adversarial networks: An overview." IEEE signal processing magazine 35.1 (2018): 53-65.

[13] Ziebart, Brian D., et al. "Maximum entropy inverse reinforcement learning." Aaai. Vol. 8. 2008.

[14] DOT, US. "Intelligent Transportation Systems Joint Program Office." *The Estimated Average Cost to Install Chargers and Outlets for Level* 2 (2014).

[15] Wen, Xiao, Zhiyong Cui, and Sisi Jian. "Characterizing car-following behaviors of human drivers when following automated vehicles using the real-world dataset." Accident Analysis & Prevention 172 (2022): 106689.

[16] Wang, Keyin, et al. "Research on Car‐Following Model considering Driving Style." Mathematical Problems in Engineering 2022.1 (2022): 7215697.

[17] Vogel, Katja. "A comparison of headway and time to collision as safety indicators." Accident analysis & prevention 35.3 (2003): 427-433.

* dengbohe@hkust-gz.edu.cn