Electronic Research Archive

ERA, 32(4): 2424–2446.
DOI: 10.3934/era.2024111
Received: 05 November 2023
Revised: 12 March 2024
Accepted: 19 March 2024
Published: 26 March 2024

http://www.aimspress.com/journal/ERA

Research article

Research on reinforcement learning based on PPO algorithm for human-machine intervention in autonomous driving

Gaosong Shi[1], Qinghai Zhao[1,2,*], Jirong Wang[1] and Xin Dong[1]

[1] College of Mechanical and Electrical Engineering, Qingdao University, Qingdao 266071, China
[2] National Engineering Research Center for Intelligent Electrical Vehicle Power System, Qingdao 266071, China

*** Correspondence:** Email: zhaoqh@qdu.edu.cn; Tel: +8615092010969.

Abstract: Given the current limitations in intelligence and processing capabilities, machine learning systems are yet unable to fully tackle diverse scenarios, thereby restricting their potential to completely substitute for human roles in practical applications. Recognizing the robustness and adaptability demonstrated by human drivers in complex environments, autonomous driving training has incorporated driving intervention mechanisms. By integrating these interventions into Proximal Policy Optimization (PPO) algorithms, it becomes possible for drivers to intervene and rectify vehicles' irrational behaviors when necessary, during the training process, thereby significantly accelerating the enhancement of model performance. A human-centric experiential replay mechanism has been developed to increase the efficiency of utilizing driving intervention data. To evaluate the impact of driving intervention on the performance of intelligent agents, experiments were conducted across four distinct intervention frequencies within scenarios involving lane changes and navigation through congested roads. The results demonstrate that the bespoke intervention mechanism markedly improves the model's performance in the initial stages of training, enabling it to overcome local optima through timely driving interventions. Although an increase in intervention frequency typically results in improved model performance, an excessively high intervention rate can detrimentally affect the model's efficiency. To assess the practical applicability of the algorithm, a comprehensive testing scenario that includes lane changes, traffic signals, and congested road sections was devised. The performance of the trained model was evaluated under various traffic conditions. The outcomes reveal that the model can adapt to different traffic flows, successfully and safely navigate the testing segment, and maintain speeds close to the target. These findings highlight the model's robustness and its potential for real-world application, emphasizing the critical role of human intervention in enhancing

the safety and reliability of autonomous driving systems.

Keywords: autonomous driving; driving intervention; experience replay; lane-changing; lane-following

1. Introduction

The field of transportation has experienced significant transformations in recent years, primarily due to the rapid advancement of autonomous driving technology [1]. Deep Reinforcement Learning (DRL) stands as a cornerstone technology in the autonomous driving decision-making process. It enables vehicles to make informed decisions within complex and dynamically changing traffic environments through the use of both simulation and real-world training [2,3].

Imitation Learning (IL) and DRL emerge as key subfields within the realm of machine learning methodologies, especially in the arena of end-to-end autonomous driving, where sensor raw data is used as input and the network model directly outputs the final control command of the vehicle [4–6]. IL aims to mimic human driver behavior by replicating observed control actions under specific scenarios [7]. The behavior-planning runtime assurance method, as delineated by Peng et al. [8] is based on imitation learning and a responsibility-sensitive safety model. This proposed framework effectively ensures the safety of behavior planning for autonomous vehicles in complex situations and demonstrates notable real-time efficacy. However, the training of deep neural networks requires a large volume of data. Eraqi et al. [9] recently introduced a novel IL approach named Dynamic Conditional Imitation Learning (DCIL), assessing its effectiveness through experiments on the CARLA simulator with promising outcomes. Nonetheless, the system's performance may significantly diminish when deployed in novel or uncharted environments. To tackle the intricacies of end-to-end autonomous driving in dense urban settings, Teng et al. [10] proposed a new strategy termed Hierarchical Interpretable Imitation Learning (HIIL). This approach seeks to enhance the vehicle's proficiency in navigating through complex and adverse conditions, necessitating the segmentation of tasks into sub-tasks and employing hierarchical solutions to address the resulting complexity. Within semi-structured driving scenarios, Ahn et al. [11] introduced a technique called "Imitation Learning for Autonomous Driving using Foresight Points", which has improved autonomous driving performance. Yet, it still relies on the use of foresight points for determining the vehicle's trajectory. Despite IL's advantageous attributes, such as high sample efficiency and swift convergence, it faces two primary challenges. It is critical to acknowledge that intelligent automation systems can replicate driver errors or hazardous maneuvers, thus augmenting the risk of accidents. Moreover, it is important to recognize that while IL models are adept at reproducing known behaviors, they struggle in unfamiliar situations, thereby hindering their decision-making capabilities [12].

DRL is a methodology that advances driving strategies through interactive learning between an agent and its environment [13,14]. The fundamental goal of this approach is to optimize the total rewards garnered from the agent's experiences [15]. Khalil and Mouftah [16] developed an innovative method that integrates multimodal fusion with latent deep reinforcement learning to enhance the perception abilities of autonomous vehicles in urban settings. Zhang et al. [17] presented a unique technique for urban autonomous driving that utilizes a dictionary ranking-based Actor-Critic deep reinforcement learning strategy. This method allows for the consideration of multiple objectives and facilitates achieving a balanced performance among them. The application of the dictionary ranking

approach effectively addresses the challenges present in continuous action domains. Furthermore, Du et al. [18] introduced a trajectory planner leveraging deep reinforcement learning to devise an automated parking system. The performance of the parking agent's trajectory planning was examined through simulation experiments. Despite significant progress in DRL, challenges persist, primarily concerning computational or learning efficiencies [19,20]. Often, the interaction between the agent and the environment is inefficient, leading to substantial computational and time investments required for model training [21]. Additionally, the construction of the reward function is of paramount importance. An inadequately designed reward function can negatively impact the algorithm's convergence rate [22].

To overcome these hurdles, a synthesis of IL and DRL has been adopted, featuring a driver intervention mechanism to enhance the capabilities of autonomous driving systems. During the training of DRL models, the integration of driver-driving experiences dynamically enriches the model learning process. Success during this learning phase is continually assessed using driver experiences, with interventions applied as needed for model adjustments. A mechanism for driver-guided experience replay is instituted, allowing the model to iteratively refine towards an optimal configuration. This methodology not only boosts the interaction efficiency between the model and the environment but also maintains the exploratory nature of DRL. Consequently, the model's learning is not solely reliant on the insights derived from driving experiences.

2. Human-machine interactive learning

Within the framework of the Proximal Policy Optimization (PPO) algorithm, this paper introduces a human-computer interaction mechanism to establish the algorithmic structure presented herein, as depicted in Figure 1.

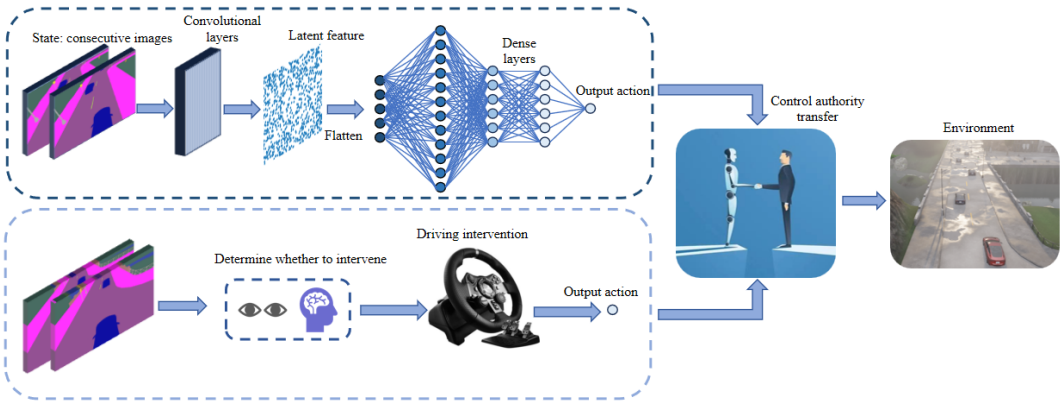


Figure 1. Human-computer interaction deep reinforcement learning.

The agent monitors the data gathered from the environment to ascertain the necessity for intervention. Upon the occurrence of intervention, actions directed by human guidance supersede those dictated by the policy. This process of action selection is encapsulated by the following representation:

$$a_t = \chi a_t^{Human} + (1-\chi) a_t^{DRL} \tag{1}$$

where a_t^{Human} is the guidance action given by a human, a_t^{DRL} is the action given by the policy network,

χ is the human intervention flag, which equals 0 in the absence of human intervention and equals 1 when human intervention is present.

The primary aim of Proximal Policy Optimization (PPO) is to enhance the expected reward by amplifying the extent of updates to the policy parameters [23,24]. To assess the differential performance of the current policy against a baseline policy, an advantage function $A(\pi)$ is employed. PPO seeks to optimize the policy parameters θ by maximizing the following objective function:

$$A_t = \widehat{G}_t - V_{\theta_v} \tag{2}$$

$$G_t = \sum_{k=0}^{\infty} \gamma^k \times r_{t+k+1} \times \left(1 - d_{t+k}\right) \tag{3}$$

$$L(\theta) = \mathrm{E}\left[\min\left(\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{old}}(a_t \mid s_t)} A_t\left(\pi_{\theta_{old}}(a_t \mid s_t)\right), clip(\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{old}}(a_t \mid s_t)}, 1-\varepsilon, 1+\varepsilon) A_t\left(\pi_{\theta_{old}}(a_t \mid s_t)\right) \right) \right] \tag{4}$$

where G_t represents the discounted return, V_θ is the estimated value of the current policy network at state s_t, γ is the discount factor, r is the reward value, d_t is the termination flag (equal to 1 if it's a terminal state, otherwise 0), $\pi_\theta(a_t \mid s_t)$ represents the probability of the current policy taking action a in state s, θ_{old} denotes the parameters of the previous policy, ε is a hyperparameter used to control the magnitude of updates.

Employing a minimized objective function as specified in Eq (4) can lead to considerable variability in gradient updates. During the initial phases of training, the presence of unstable gradient updates can induce oscillations within the optimization process. This instability complicates the convergence of policy parameters to optimal values. To address this issue, the objective function is modified to achieve the subsequent formulation:

$$L_{new}(\theta) = -L(\theta) + \lambda_1 \times M\left(V_\theta(s_t), G_t\right) - \lambda_2 \times H\left(\pi_\theta(a_t \mid s_t)\right) \tag{5}$$

where λ_1 and λ_2 are weight coefficients, M is the mean square error loss term used to optimize the prediction of the state value function, thus estimating the dominance function more accurately, and H denotes the entropy regularization term, which is included to encourage the policy to maintain a certain level of randomness in situations characterized by high uncertainty. The expression for M is as follows:

$$M\left(V_\theta(s_t), G_t\right) = \frac{1}{2} \times \frac{1}{n} \sum_{i=1}^{n} \left(V_\theta\left(s_t\right)_i - \widehat{G}_t\right)_i^2 \tag{6}$$

During the initial training stages of autonomous driving, and when the model encounters local optima, a driving intervention mechanism is activated to influence the system's decision-making. This process utilizes the expertise of a human driver to reduce instability throughout the training phase. Should the system display irrational driving behavior, the human driver steps in to steer the system towards more sensible choices, thereby enhancing the safety and dependability of the driving actions [25]. Given that the model has yet to master effective strategies in the early stages of training, the insights from the human driver are deemed significantly more vital than the model's exploratory efforts. The exploratory experiences of the model and the driving experiences of the human are cataloged

separately in buffers A and B, respectively [26]. Throughout the learning phase of the model, samples are extracted from both buffers in accordance with the designated sampling rules:

$$N = N_{agent} + N_{human} \tag{7}$$

$$N_{agent} = \alpha N \tag{8}$$

$$N_{huamn} = (1-\alpha)N \tag{9}$$

$$\alpha = \max\left(\frac{R_{agent}}{R_{human} + R_{agent}}, 0.2\right) \tag{10}$$

where N represents the total number of samples taken during one iteration of model learning, R_{human} stands for the average reward value from the human driver's driving experiences, and R_{agent} represents the average reward over the last 10 episodes obtained by the model.

When α surpasses the predetermined proximity threshold, it signifies the reliance on a human driver's driving experience, and buffer B is consequently disregarded. At this juncture, the model has not yet achieved a state of satisfactory convergence, and its driving performance remains subpar compared to that of the human driver. This transition signifies the commencement of the second phase of training, distinguished by sporadic driver interventions. During this phase, while the model is adept at executing straightforward tasks, it is prone to committing errors and incurring negative rewards within certain complex situations. To mitigate this, prompt driver intervention is deployed when the model verges on entering a detrimental state [27]. The driving data procured through this intervention are conserved in a buffer, with the training data stored therein outlined as follows:

$$D = \{s_t, a_t, s_{t+1}, r_t, I\} \tag{11}$$

where s_t represents the current state, a_t represents the current action, s_{t+1} represents the next state, and I serve as a marker to distinguish between model-generated and human driver-generated driving data.

To guarantee that autonomous vehicles adequately respond to human driver interventions in critical scenarios, a driver intervention term is incorporated into the loss function of the PPO algorithm. The conventional loss function of the PPO algorithm is designed to maximize cumulative rewards. However, in certain perilous situations, this objective could cause the agent to depend excessively on environmental rewards, thus neglecting the crucial interventions made by human driving experts. To rectify this imbalance, the loss function has been revised to assign greater significance to driver intervention signals throughout the training phase. The revised loss function is presented as follows:

$$L_{new}(\theta) = -L(\theta) + \lambda_1 \times M\left(V_\theta(s_t), G_t\right) - \lambda_2 \times H\left(\pi_\theta(a_t \mid s_t)\right) - \lambda_3 \left\| a_t - \pi_\theta(a_t \mid s_t) \right\|^2 \tag{12}$$

Following the outlined methodology, the algorithmic framework depicted in Figure 2 forms the central theme of this manuscript. The term "State" pertains to the input derived from the camera, which is an RGB image with dimensions of 160 × 80 × 3. "Action" corresponds to the output action, confined within the range of [-1, 1].

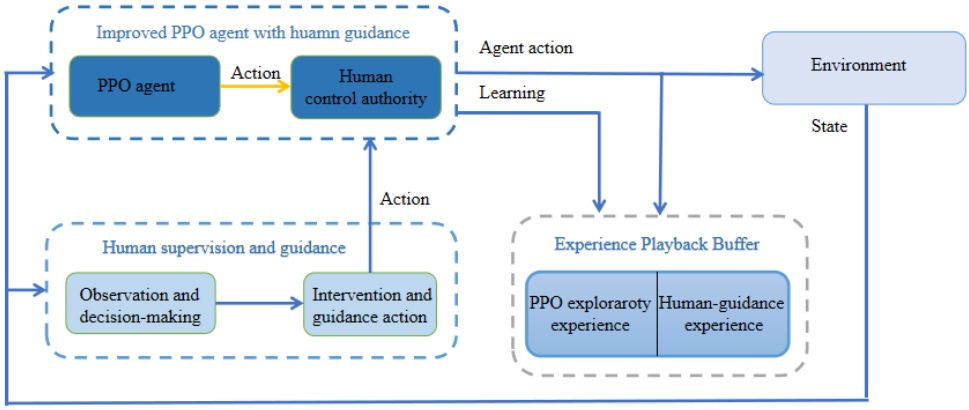


Figure 2. Human-computer interaction deep reinforcement learning algorithm framework.

3. Reward function design

3.1. Design of lane change reward function

To safeguard driving safety, the reward function incorporates the distance to the following vehicle [28]. Should the distance to the vehicle ahead become perilously short, negative rewards escalate swiftly. The computation of the reward concerning the following distance is delineated as follows:

$$r_{fro} = -\left(\frac{1}{\max\left\{d_f, d'\right\}}\right)^2 \tag{13}$$

where d_f is the distance between the front of the vehicle and the rear of the preceding vehicle, d' is a small positive constant to ensure a positive denominator.

The reward terms are designed to motivate the agent to execute smooth lane-changing maneuvers. The smoothness of these maneuvers can be quantified by observing the deviation in the steering angle. The formula for calculating the reward based on steering angle deviation is as follows:

$$r_{smo} = -e^{|\delta_t - \delta_{t-1}|} \tag{14}$$

where δ_t and $\delta_{t\text{-}1}$ represent the current time-step steering angle and the previous time-step steering angle, respectively.

The reward mechanism is designed to prompt the agent to keep its position centered within the current lane, thus preventing the vehicle from approaching too closely to the lane boundaries and facilitating its ability to remain near the centerline following a successful lane change [29,30]. Recognizing that a vehicle will inevitably stray from the lane's center during the process of changing lanes, this specific reward is conferred only upon the successful completion of the lane change maneuver. In an effort to further encourage the vehicle to cover longer distances, a cooperative driving distance term is integrated into the reward function, articulated as follows:

$$r_{cen} = -\frac{1}{1+e^{-|d_i|}} \tag{15}$$

$$r_{dis} = \left\| d_{cur} - d_{old} \right\|^2 \tag{16}$$

where d_i represents the distance from the lane centerline, while d_{cur} and d_{old} respectively, indicate the distance to the target point at the current time step and the distance to the target point at the previous time step.

Human intervention aims to amend the conduct of the Deep Reinforcement Learning (DRL) agent and avert disastrous outcomes. During sporadic intervention periods, human decision to intervene in certain situations signals that the current state is deemed adverse. To maximize the accumulation of rewards, the DRL agent is encouraged to limit exposure to detrimental states, thereby reducing the necessity for human intervention [31]. Reward penalties are applied at the initial time step of artificial intervention incidents to dissuade the agent from entering such states. To motivate the agent towards the successful execution of lane changes and integration into the target lane, additional positive rewards are allocated. This serves to stimulate the agent's pursuit of its objectives within the given task. To avert collisions, negative rewards are issued if the agent's maneuvers are likely to result in potential collisions, with the intensity of the reward or penalty being proportional to the gravity of the anticipated collision. Incorporating all these elements, the ultimate comprehensive reward function is established as follows:

$$r_{change} = \omega_1 r_{fro} + \omega_2 r_{smo} + \omega_3 r_{cen} + \omega_4 r_{dis} + r_{gui} + r_{col} + r_{fin} \tag{17}$$

where ω_1 and ω_2 are weight coefficients, r_{gui} represents intervention punishment, r_{col} represents the collision penalty, and r_{fin} represents the reward for completing the lane change.

3.2. Designing a reward function in a following-vehicle scenario

In scenarios involving following another vehicle, the reward function should take into account variables such as maintaining a safe following distance, matching the speed of the vehicle ahead, and ensuring smooth following behavior [32,33]. To incentivize the agent vehicle to exhibit smooth following behavior, reward terms can be formulated to penalize behaviors such as abrupt braking, rapid acceleration, sharp steering, or frequent speed changes. The calculation of rewards could be structured as follows:

$$r'_{smo} = -e^{\left|k\Delta a_x + \Delta a_y\right|} \tag{18}$$

where Δa_x represents the change in the vehicle's longitudinal acceleration, and Δa_y represents the change in lateral acceleration, k is a constant greater than 1. Due to the greater impact of lateral acceleration on the smoothness of the vehicle's motion, it receives higher attention.

Additional positive rewards are awarded to the agent upon successfully completing a lane change and entering the target lane, thereby motivating the agent to accomplish objectives within the task. To deter collisions, negative rewards are dispensed if the agent's actions pose a risk of leading to potential collisions, with the penalties calibrated according to the severity of the possible collisions [34]. Taking into account the aforementioned factors, the ultimate comprehensive reward function is formulated as follows:

$$r_{follow} = \omega_3 r'_{smo} + \omega_5 r_{fro} + r_{gui} + r_{col} + r_{fin} \tag{19}$$

The process described encapsulates the entire workflow of this manuscript. Figure 3 illustrates the research methodology employed in this study.

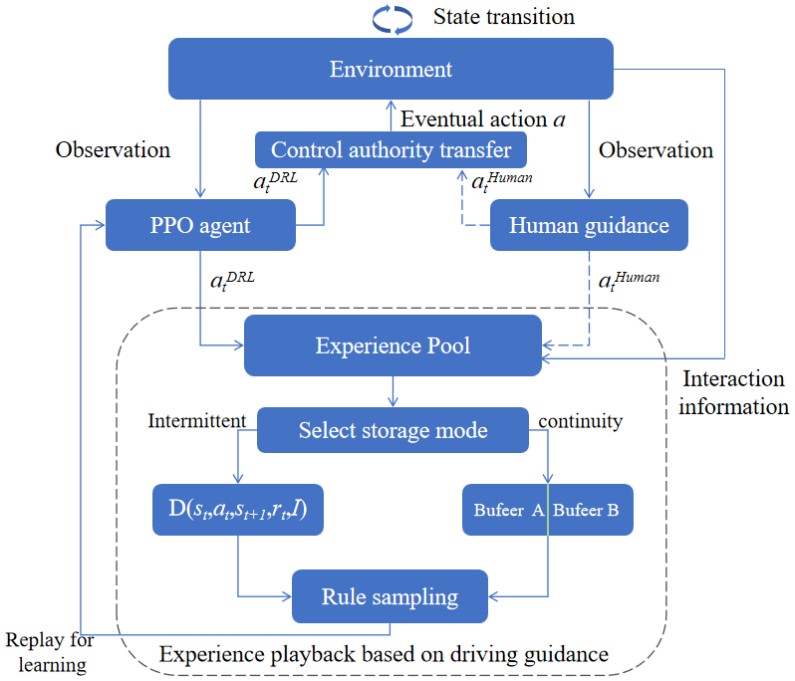


Figure 3. Research process.

4. Experiment

To evaluate the efficacy of the PPO algorithm when augmented with driving intervention, a series of experiments were performed. These experiments encompassed maneuvers such as left lane changes, right lane changes, and car-following scenarios. The trials were executed using the simulation software Carla-Town01, with the experimental scenario showcased in Figure 4(a). The experimental bench is shown in Figure 4(b).

The performance of experimental equipment, especially the computational capabilities of CPUs and GPUs, has a significant impact on the time and efficiency of model training. The detailed specifications of the experimental configuration are presented in Table 1.

Table 1. Experimental configuration.

Assembly	Specifications
CPU	Intel i9-12900H
GPU	NVIDIA RTX3050Ti
Memory	16G
Driving equipment	Kraton 900

(a) Experimental scenario

(b) Experimental bench

Figure 4. Experimental scenario.

The design of the experimental scenarios aims to fully simulate various situations in the real-world driving environment, with the speed of the ego vehicle and the obstacle vehicle, as well as the position of the obstacle vehicle, as shown in Table 2.

Table 2. Scene settings.

Scene	Initial lane	Ego vehicle speed (m/s)	The speed of both obstacle courses (m/s)	Obstacle vehicle location (Forward, m)
Left lane change	left lane	5	3	10
Right lane change	right lane	5	3	10
Following	randomization	-	1–6	5–15

4.1. Intervention analysis

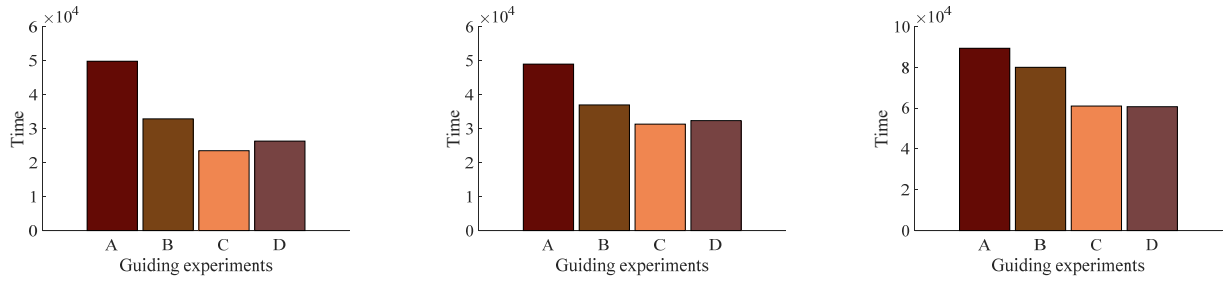
During the experimental phase, the driver's interventions with the agent were not scheduled at specific intervals but were instead dictated by the intervention protocols established for the four experimental scenarios. Interventions were initiated through manual steering adjustments, thereby guiding the agent's actions. The primary aim was to ensure the intelligent agent remained on the road and to reduce the incidence of collisions with road boundaries or nearby vehicles. Interventions ceased once the driver observed that the agent was navigating correctly and demonstrating acceptable behavior. The objective of these experiments is to explore how human-guided interventions can enhance the performance of the agent. Four distinct experimental scenarios were conducted.

Experiment A: No intervention was applied throughout the training process.

Experiment B: Interventions were made continuously for every 10 episodes, following every 100 episodes.

Experiment C: Interventions were made continuously for every 30 episodes, following every 100 episodes.

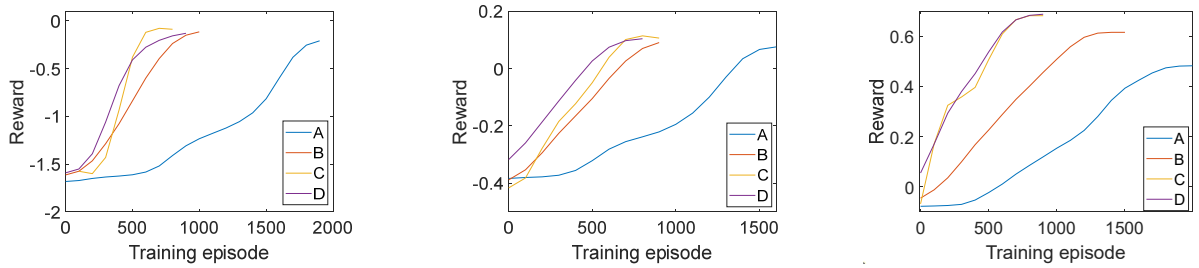
Experiment D: Interventions were made continuously for every 50 episodes, following every 100 episodes.



(a) Left lane change (b) Right lane change (c) Following

Figure 5. Training end time.

Figure 5 depicts the convergence times of four different driving intervention experiments across three scenarios, showcasing the training efficiency of intelligent agents under varying levels of driving intervention. The results indicate that in the absence of driving intervention, the time required for the intelligent agent to converge is the longest. This suggests that without intervention, intelligent agents take a longer duration to adapt and learn how to perform tasks effectively. In the lane change scenarios (Figure 5(a),(b)), Experiment C exhibits the shortest training time, implying that introducing a moderate degree of driving intervention positively influences the training of the intelligent agent. However, it is important to acknowledge that excessive intervention might diminish the exploration capability of the intelligent agent, rendering it less adaptable to future unforeseen circumstances and consequently increasing the time to convergence. In the car-following scenario, Experiment D demonstrates the shortest training duration, which may be ascribed to the complexities associated with speed control in dense traffic conditions, necessitating more precise control measures. The challenge faced by intelligent agents in discovering suitable control strategies underscores the significance of human intervention in enhancing learning and adaptation processes.



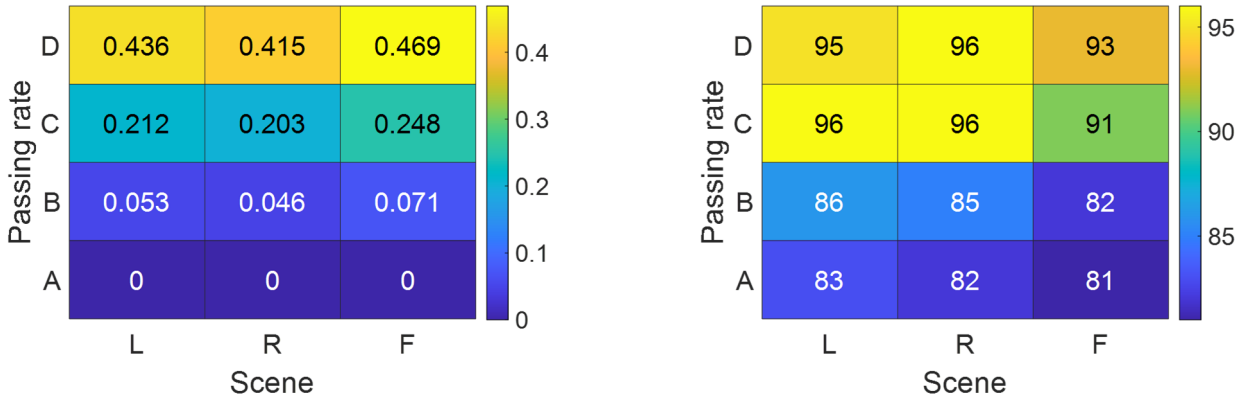
(a) Left lane change (b) Right lane change (c) Following

Figure 6. Reward value change.

Figure 6 illustrates the progression of the average reward value for every 100 episodes across four experiments in three distinct scenarios. Initially, in the absence of driving intervention, the intelligent agent predominantly remains in a negative state during the early training phases, reflected by low reward values. This scenario likely arises because the intelligent agents are required to independently navigate complex driving tasks without any intervention, necessitating extensive exploration and trial-

and-error, which in turn leads to suboptimal performance at the outset. However, the introduction of driving intervention marks a significant improvement in the agent's early performance. This improvement suggests that driving intervention offers essential guidance and feedback during the initial training stages, thereby facilitating the intelligent agent's quicker adaptation to the tasks at hand. Such early performance enhancement is particularly crucial for autonomous driving systems in real-world applications, as it has the potential to mitigate risks and safety concerns.

Upon completion of training without driving intervention, the agent's reward value is observed to be lower compared to the scenarios where driving intervention is applied. This outcome underscores the considerable positive impact of driving intervention on further enhancing model performance. By expediting the learning process, driving intervention not only accelerates the pace at which intelligent agents acquire new skills but also elevates their performance levels in executing given tasks.



(a) Intervention rate

(b) Scenario pass rate

Figure 7. Scenario testing.

Figure 7 illustrates the dynamics between the average intervention rate during training across three scenarios and the corresponding test pass rates. As depicted in Figure 7(a), the intervention rates vary, while Figure 7(b) showcases that pass rates are notably higher in instances where driving intervention is employed, compared to scenarios devoid of such interventions. A notable observation is that when the intervention rate is minimal, the improvement in model performance is similarly limited. Conversely, as the intervention rate escalates, there's a discernible uptick in the model's pass rate, indicating a positive correlation between the rate of intervention and model efficacy.

However, it is observed that when the intervention rate falls within the 0.2 to 0.5 range, the marginal gains in model performance begin to diminish with further increases in the intervention rate. This pattern suggests that, for the methodology proposed in this manuscript, maintaining the intervention rate within the 0.2–0.5 spectrum is optimal. Moreover, to mitigate the risk of excessive fatigue for the individual conducting interventions, it's advisable to keep the intervention rate between 0.2 and 0.3 for lane-changing scenarios. For scenarios involving following another vehicle, adjusting the intervention rate to fall between 0.3 and 0.5 is recommended. This strategic modulation of intervention rates aims to balance the dual objectives of optimizing model performance and minimizing the intervener's strain.

Table 3 comprehensively presents the data for various metrics measured during the experiment.

The calculation of the reward values is based on the average reward over the last 100 training cycles. The data from the table reveals that appropriate driving interventions can significantly enhance performance metrics throughout the training process. Such improvements exhibit notable variations depending on the frequency of interventions. Moreover, experimental data also indicate that the response to intervention measures varies across different driving scenarios.

Table 3. Experimental data.

Value type	Intervention plan	Left lane change	Right lane change	Following
Reward	A	-0.20	0.07	0.48
	B	-0.11	0.09	0.61
	C	-0.08	0.11	0.68
	D	-0.13	0.10	0.69
Episode	A	1892	1586	1926
	B	978	893	1455
	C	791	852	886
	D	889	796	874
Training time (s)	A	49,807	48,932	89,320
	B	32,832	36,932	80,019
	C	23,465	31,275	61,023
	D	26,321	32,321	60,693
Passing rate (%)	A	83	82	81
	B	86	85	82
	C	96	96	91
	D	95	96	93

Table 4 selects the intervention plans that show the most significant improvement in performance across various experimental scenarios and compares them with data from experiments without interventions. Interventions not only significantly enhance the model's performance on specific tasks but also accelerate the model's adaptation to complex driving environments.

Table 4. Performance improvement.

Scene	Intervention plan	Enhancement ratio (%)		
		Episode	Training time	Passing rate
Left lane change	C	57.89	52.89	14.46
Right lane change	C	43.75	36.08	17.07
Following	D	55	31.68	14.81

4.2. Impact of human proficiency and driving qualifications

The study investigates the influence of human proficiency and driving qualifications on the enhancement of algorithm performance. Six participants were enlisted for this experiment, categorized based on their driving qualifications and proficiency. Two participants possessing a driver's license

were deemed qualified participants. Two individuals without a driver's license were considered unqualified participants. Meanwhile, two licensed participants who received 10 minutes of preparatory training to familiarize themselves with the environment and equipment prior to the experiment were labeled as proficient participants. The intervention protocol was set to occur every 30 episodes after every set of 100 episodes, with the experimental scenario focused on lane changing.

Driver proficiency emerges as a potential factor impacting experimental outcomes, and Table 5 delineates the results from drivers of varying proficiency levels. Notably, as driving proficiency increases, the model's required training duration diminishes, and its pass rate escalates. Despite these trends, the improvements in training duration and pass rate are modest. Remarkably, even training with unqualified drivers successfully accomplishes lane-changing tasks. This outcome suggests that while the model benefits from the driver's experience, it predominantly relies on its inherent exploratory learning capabilities. Consequently, driving proficiency exerts a relatively minor influence on the model's performance.

Table 5. Training results for different drivers.

Participants	Age	Proficiency	Training time	Passing rate (%)
A	22	Unqualified driver	8.5 h	92
B	22	Unqualified driver	8.6 h	93
C	24	Qualified driver	8.3 h	93
D	25	Qualified driver	8.1 h	95
E	24	Skilled driver	7.8 h	95
F	25	Skilled driver	7.7 h	96

4.3. Lane changing experiment

In this section, we elaborate on the training specifics for the left lane change scenario in Experiment C. A notable observation is that due to the driver's inability to initiate intervention immediately at each training cycle's beginning, the intervention rate did not achieve 1 even under continuous intervention, as illustrated in Figure 8. During the stages of intermittent intervention, as the model's performance enhances, the frequency of interventions gradually diminishes. Consequently, the model progressively reduces its reliance on driving interventions, primarily depending on its capability for autonomous exploration.

Figure 9 displays the evolution of reward values throughout the training process. A significant increase in reward value is observed during the intervention period. Following the cessation of intervention, there is a noticeable decrease in reward value, yet it remains substantially higher than pre-intervention levels. With intermittent intervention, even after the discontinuation of intervention measures, the reward value stabilizes at a high level. This outcome underscores the beneficial impact of driver intervention measures on augmenting model performance, highlighting the effectiveness of integrating human expertise with autonomous learning processes to achieve optimal outcomes in complex driving scenarios.

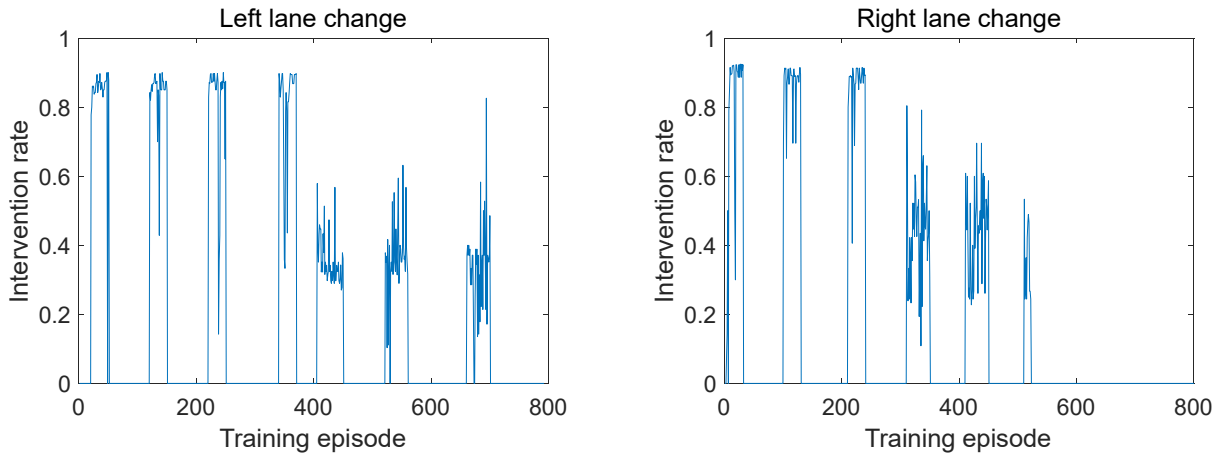


Figure 8. Change in intervention rate.

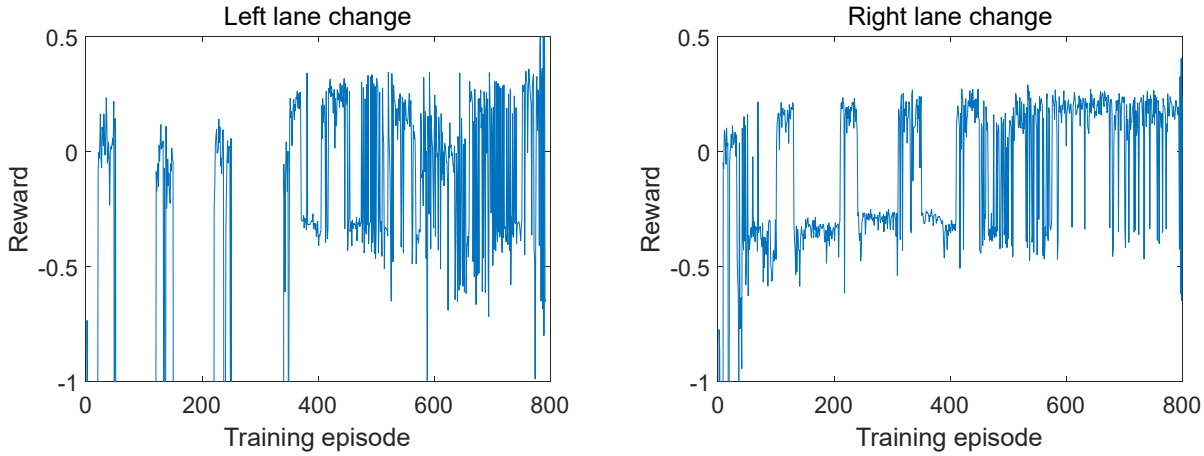


Figure 9. Reward value change.

Figure 10 illustrates the progression in the number of survival steps per episode following driver intervention. There is a marked increase in the survival steps per episode for the model post-intervention, underscoring the importance of intervention in enhancing the model's ability to navigate complex driving tasks. Driver intervention, both direct and controlled, equips the model with improved mechanisms to handle uncertainties and potential hazards, thereby prolonging its operational duration within scenarios.

Notably, after ceasing the intervention, there's a decrease in the number of survival steps per episode, yet the performance remains significantly enhanced relative to scenarios devoid of any intervention. This observation affirms the effectiveness of driver intervention in boosting model performance. Following the discontinuation of intervention, the survival steps exhibit a stepwise improvement across every subsequent set of 100 episodes, indicating the model's gradual adjustment to previous interventions and its ongoing enhancement in survival capabilities throughout this adaptation period. The survival capacity of the model in the intermittent intervention stage demonstrates a trend toward stabilization, highlighting the critical role of driving intervention in bolstering system stability and safety.

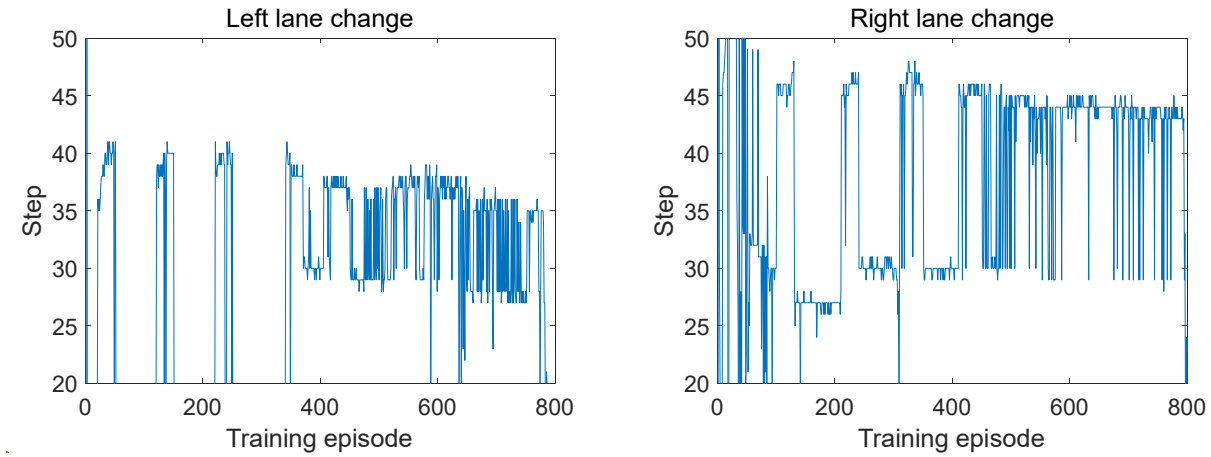


Figure 10. Steps run per episode.

Figure 11 displays the trajectory data garnered from 100 tests conducted on the trained model within lane-changing scenarios. The vehicle adeptly executed a complex lane change, showcasing a smooth trajectory throughout the operation. In the context of autonomous driving systems, the ability to successfully navigate lane changes is paramount, necessitating that intelligent agents change lanes safely and efficiently in response to the surrounding traffic conditions. The smooth trajectory depicted in Figure 11 signifies that the trained model possesses high control precision and stability for this particular task. Through the application of driving intervention, the model gains a deeper comprehension of the appropriate timing and methodology for executing lane changes, thereby facilitating smooth and effective maneuvers. This outcome underscores the significance of integrating human insights with autonomous learning to refine the operational capabilities of autonomous driving models, ensuring they can adeptly handle the complexities of real-world driving scenarios.

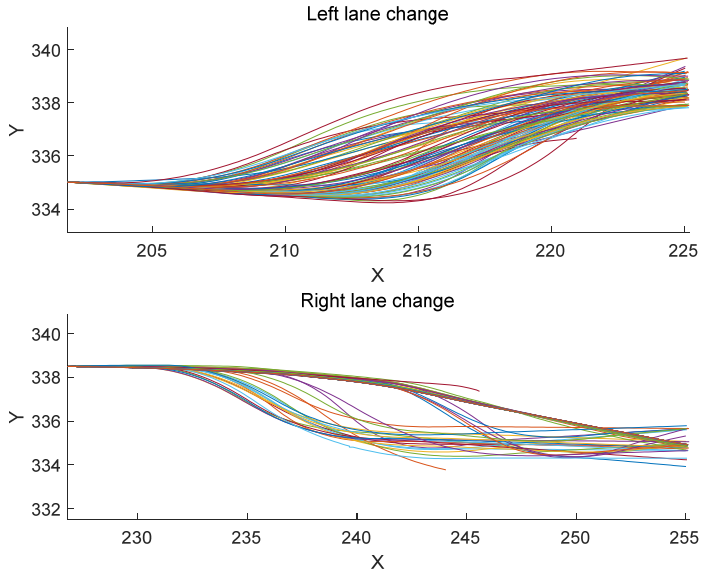


Figure 11. Test trajectory.

Figure 12 illustrates the variation in reward values throughout the training process. During the initial phase of training, without the implementation of driving intervention, the fluctuation in reward values remains relatively minor. However, upon introducing driving intervention, a significant positive shift in the reward values is observed, leading to an increased range of fluctuation. As the process transitions into the intermittent intervention stage, the variability of the reward values gradually begins to stabilize around 0.3. This pattern suggests that driving intervention not only enhances the performance of the model but also contributes to its stability. The ability of driving intervention to mitigate extreme fluctuations in reward values indicates its effectiveness in guiding the model towards more consistent and reliable performance outcomes.

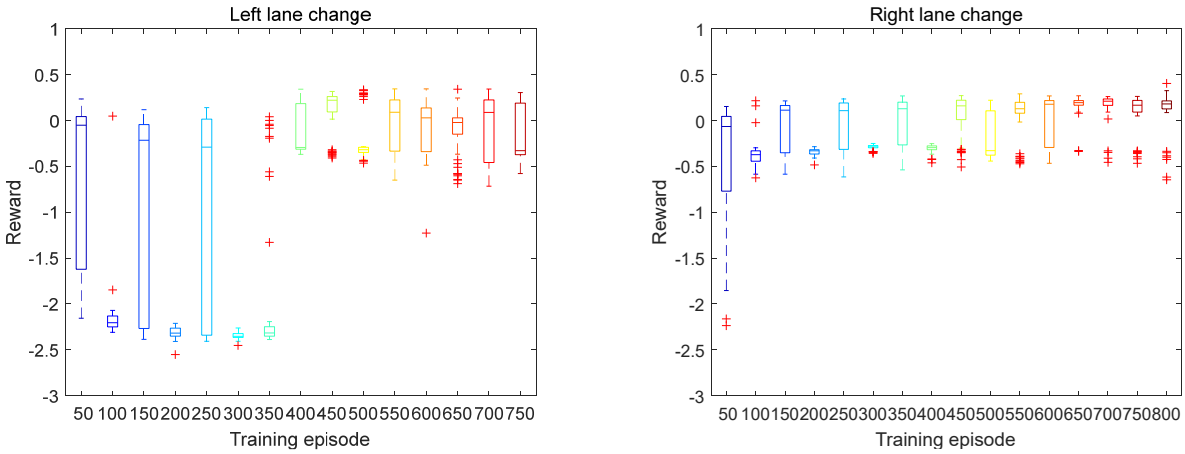


Figure 12. Reward value fluctuation.

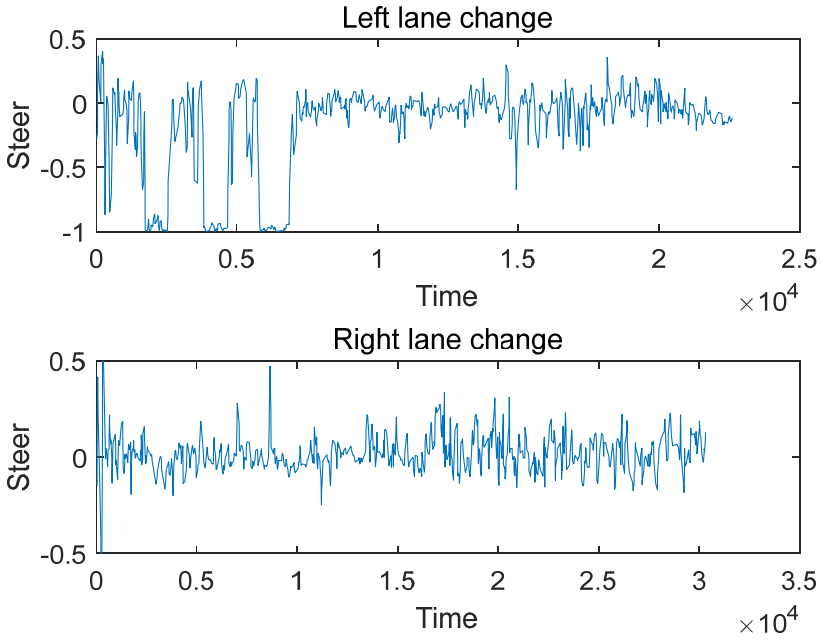
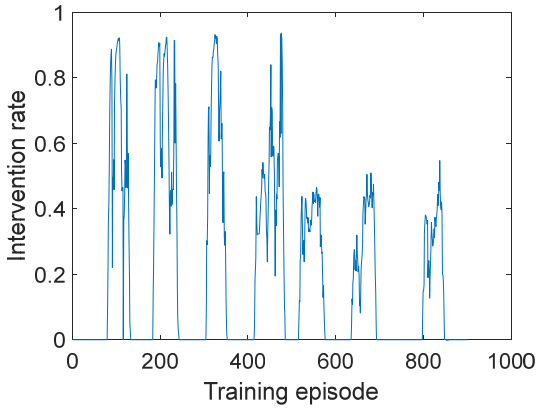


Figure 13. Steering angle.

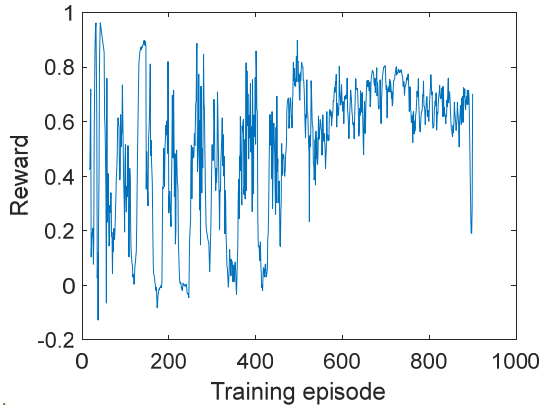
Figure 13 captures the variation in steering angle over the course of training. Initially, the steering angle is subject to significant fluctuations, reflecting the model's struggle to stabilize its driving behavior. However, as intervention measures are introduced, these fluctuations begin to diminish, signaling an enhancement in the smoothness of the vehicle's driving, particularly during lane-changing maneuvers. This trend suggests that the interventions not only aid in refining the model's decision-making processes but also contribute significantly to the improvement of its operational smoothness. The decrease in steering angle variability is indicative of the model's growing proficiency in executing lane changes with greater control accuracy and stability, underscoring the value of targeted intervention in facilitating the development of more refined and reliable autonomous driving behaviors.

4.4. Following experiment

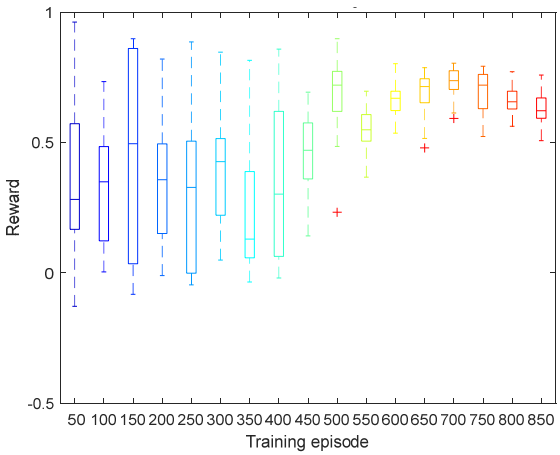
In the training setup, the lane length is limited to 50 m. The lane in which the vehicle is positioned is randomly determined. Within each training batch, obstacle vehicles are strategically placed at distances ranging from 5 to 15 m ahead of the vehicle's lane and the adjacent lanes. The speeds of these randomly generated obstacle vehicles vary between 1 to 6 m/s. The rate of intervention during the training process is depicted in Figure 14(a), while Figure 14(b) illustrates the corresponding reward values.



(a) Change in intervention rate



(b) Reward value change



(c) Reward value fluctuation

Figure 14. The following training.

Given the heightened risk of the vehicle remaining in a potentially hazardous state of imminent collision for extended periods in car-following scenarios, a significant level of driver vigilance is necessitated. These situations are inherently more complex, leading to pronounced fluctuations in reward values during the periods of intermittent intervention. However, as the model's performance incrementally improves, the necessity for intensive driver focus diminishes. Consequently, as demonstrated in Figure 14(c), the fluctuations in reward values begin to stabilize. This progression underscores the critical role of driver intervention in navigating complex scenarios and enhancing the model's capability to predict and avoid collisions, thereby reducing the demand for continuous human oversight and facilitating a transition towards more autonomous operational stability.

4.5. *Full section experiment*

To facilitate autonomous vehicles in decelerating to an appropriate speed as they approach a stop line upon encountering a red or yellow traffic signal, the following acceleration control equation (Eq (20)) can be formulated:

$$a = \frac{v^2}{2\left(d_l - pv\right)} \tag{20}$$

where a is the acceleration, and v represents the current vehicle speed. d_l is the distance of the vehicle from the stop line, and p is a hyperparameter less than 1.

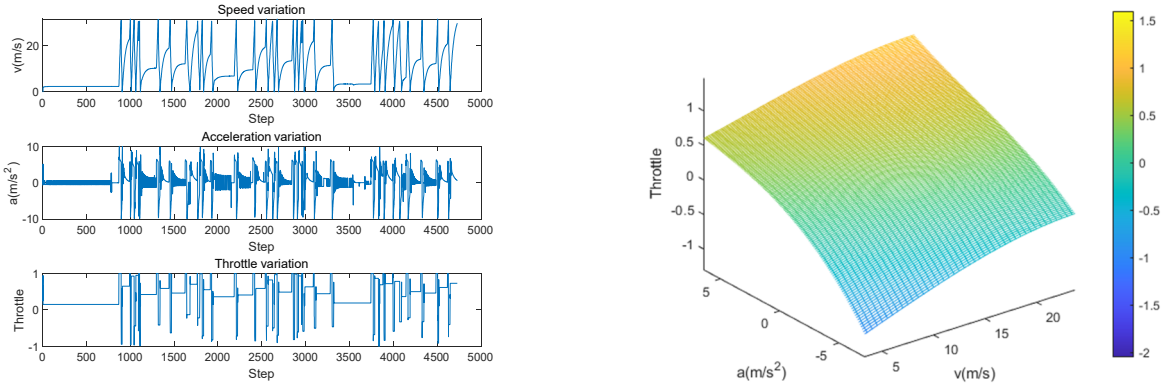
To ensure meticulous management of the vehicle's acceleration dynamics, the vehicle within the Carla simulation environment undergoes a series of continuous acceleration and deceleration maneuvers. Throughout these maneuvers, data pertaining to speed, acceleration, throttle, and brake inputs are meticulously logged, as showcased in Figure 15(a). To accurately describe the interplay among these variables, a third-order polynomial regression model is adopted. This model serves as the cornerstone for executing controlled acceleration strategies, with the ultimate goal of achieving the precise halting of the vehicle at signalized intersections. The mathematical formulation of this model is articulated as follows:

$$h_\psi\left(x\right) = \psi_0 + \psi_1\phi_1\left(x\right) + \psi_2\phi_2\left(x\right) + \psi_3\phi_3\left(x\right) \tag{21}$$

$$J\left(\psi\right) = \frac{1}{2n}\sum_{i=1}^{n}\left(h_\psi\left(x_i\right) - z_i\right)^2 \tag{22}$$

where x represents the information of the vehicle's speed and acceleration, and z_i corresponds to the braking or acceleration input.

The correlations between speed, acceleration, and braking derived from the training of Eqs (22) and (23) are illustrated in Figure 15(b). These relationships facilitate the precise regulation of vehicle acceleration, allowing for more accurate and responsive control over the vehicle's dynamics. Through the application of these trained equations, autonomous vehicles can adjust their acceleration and braking forces with a high degree of accuracy, ensuring smoother transitions and safer driving behaviors, especially in complex driving scenarios like approaching signalized intersections.



(a) Sampled data (b) Speed, acceleration, braking relationship diagram

Figure 15. Speed, acceleration, braking relationship diagram.

In the Carla simulation environment, the ego vehicle is instantiated with a designated target speed of 5 m/s. Obstacle vehicles are spawned ahead of the ego vehicle, with the distances between the leading and the following vehicles progressively decreasing. To evaluate the performance of the test model, different numbers of vehicles were generated within 100 meters in front of the self vehicle-specifically, 5, 10 and 15 vehicles-and set to Carla's self-driving mode. The testing segment extends over a total distance of 200 meters and includes a variety of driving scenarios, such as lane changes, traffic lights, and areas of traffic congestion, as depicted in Figure 16. This comprehensive test setup is designed to rigorously assess the models' capabilities in navigating complex driving environments, thereby providing insights into their effectiveness and potential areas for improvement.

(a) Left lane change (b) Right lane change (c) Signal light section (d) Congestion section

Figure 16. Test scenario.

Figure 17 showcases the trajectory variations of vehicles across three experimental setups. In varying traffic conditions, the vehicles demonstrate an effective ability to adapt to the surrounding traffic environment, maintaining stable trajectories throughout. Figure 18 captures the speed variations observed in these experiments. The red and green dots within the figure symbolize the states of red and green traffic signals encountered by the vehicle. Speeds tend to oscillate around the 5 m/s mark, with the vehicles undertaking lane changes 1, 2, and 1 times in each of the three experiments, respectively. The approach to red traffic lights is marked by a smooth deceleration to a halt, whereas

green lights permit uninterrupted passage. In sections marked by traffic congestion, notable speed fluctuations are observed, particularly as traffic density increases.

To adhere to the target speed while averting collisions, vehicles engage in a continuous cycle of acceleration and deceleration, thereby adjusting their proximity to the vehicle ahead. The impact of speed fluctuations is more evident under conditions of higher traffic density, especially when dealing with traffic volumes of 10 and 15 vehicles, which notably extend the duration required to traverse the test segment. The most prolonged passage time is recorded at a traffic density of 10, attributed to prolonged halts at red lights compounded by increased vehicle density. Conversely, at a traffic density of 5, despite encountering two red lights, the passage time is minimized due to the lower volume of traffic. In each scenario, vehicles efficiently navigate through intersections while prioritizing safety, underscoring the model's robust adaptability to diverse traffic situations and its capability to balance efficiency with safety considerations.

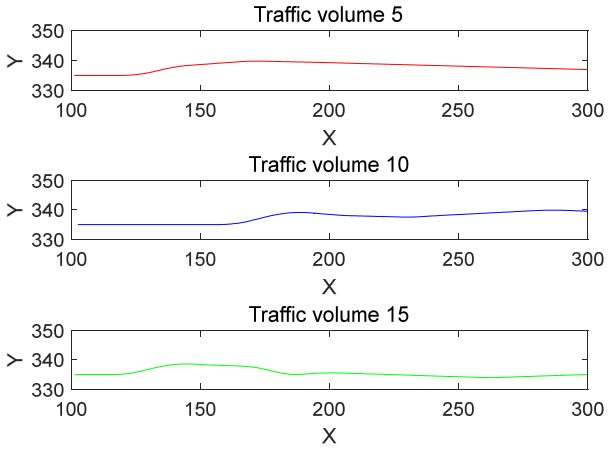


Figure 17. Trajectory change.

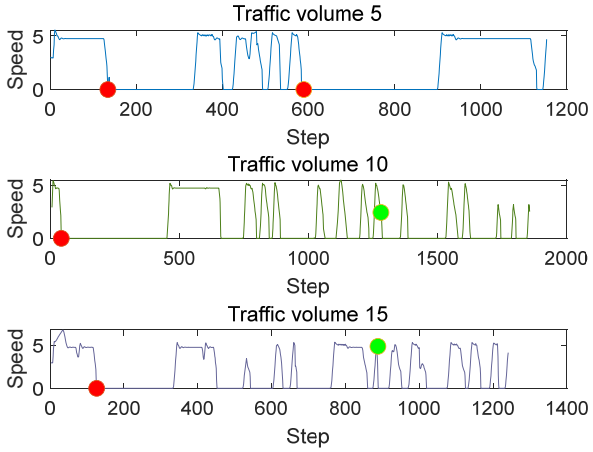


Figure 18. Vehicle speed change.

5. Conclusions

Given the current limitations in intelligence and capabilities, machine learning cannot yet fully supplant human involvement in practical applications due to its inability to navigate various situations autonomously. This article introduces an innovative approach that marries human driving expertise with the exploratory learning process of reinforcement learning, culminating in a human-guided reinforcement learning framework. This framework integrates driving intervention within the PPO algorithm and introduces a human-guided experience playback mechanism. The utility of driving intervention in hastening the enhancement of model performance was validated by examining the effects of varying intervention frequencies on agent performance, underscoring the importance of fine-tuning intervention rates to optimize benefits. The model's adept handling of diverse driving scenarios further underscores the practical viability of the proposed method, nudging us closer to realizing a safer and more dependable autonomous driving system.

However, the training regimen necessitates continuous driver vigilance and intervention, presenting a challenge to the driver's endurance and focus, as extended periods of intervention could induce fatigue, potentially impairing model performance. Moreover, the efficacy of the model can be affected by the driver's skill level and subjective decision-making regarding the timing of interventions. Future research will concentrate on how to maintain effective driving intervention mechanisms while minimizing reliance on drivers and alleviating the burdens associated with driving interventions, thereby ensuring the sustainable development of autonomous driving technologies.

Use of AI tools declaration

The authors declare they have not used artificial intelligence (AI) tools in the creation of this article.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant 52175236).

Conflict of interest

The authors declare there are no conflicts of interest.

References

1. I. Yaqoob, L. U. Khan, S. M. A. Kazmi, M. Imran, N. Guizani, C. S. Hong, Autonomous driving cars in smart cities: Recent advances, requirements, and challenges, *IEEE Network*, **34** (2020), 174–181. https://doi.org/10.1109/MNET.2019.1900120
2. B. R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. A. Sallab, S. Yogamani, et al., Deep reinforcement learning for autonomous driving: a survey, *IEEE Trans. Intell. Transp. Syst.*, **23** (2022), 4909–4926. https://doi.org/10.1109/TITS.2021.3054625
3. L. Anzalone, P. Barra, S. Barra, A. Castiglione, M. Nappi, An end-to-end curriculum learning approach for autonomous driving scenarios, *IEEE Trans. Intell. Transp. Syst.*, **23** (2022), 19817–19826. https://doi.org/10.1109/TITS.2022.3160673

4. J. Hua, L. Zeng, G. Li, Z. Ju, Learning for a Robot: Deep reinforcement learning, imitation learning, transfer learning, *Sensors*, **21** (2021), 1278. https://doi.org/10.3390/s21041278
5. K. Makantasis, M. Kontorinaki, I. Nikolos, Deep reinforcement-learning-based driving policy for autonomous road vehicles, *IET Intell. Transp. Syst.*, **14** (2019), 13–24. https://doi.org/10.1049/iet-its.2019.0249
6. L. L. Mero, D. Yi, M. Dianati, A. Mouzakitis, A survey on imitation learning tech-niques for end-to-end autonomous vehicles, *IEEE Trans. Intell. Transp. Syst.*, **23** (2022), 14128–14147. https://doi.org/10.1109/TITS.2022.3144867
7. A. Hussein, M. M. Gaber, E. Elyan, C. Jayne, Imitation learning: A survey of learning methods, *ACM Comput. Surv.*, **50** (2017), 1–35. https://doi.org/10.1145/3054912
8. Y. Peng, G. Tan, H. Si, RTA-IR: A runtime assurance framework for behavior planning based on imitation learning and responsibility-sensitive safety model, *Expert Syst. Appl.*, **232** (2023). https://doi.org/10.1016/j.eswa.2023.120824
9. H. M. Eraqi, M. N. Moustafa, J. Honer, Dynamic conditional imitation learning for autonomous driving, *IEEE Trans. Intell. Transp. Syst.*, **23** (2022), 22988–23001. https://doi.org/10.1109/TITS.2022.3214079
10. S. Teng, L. Chen, Y. Ai, Y. Zhou, Z. Xuanyuan, X. Hu, Hierarchical interpretable imitation learning for end-to-end autonomous driving, *IEEE Trans. Intell. Transp. Syst.*, **8** (2023), 673–683. https://doi.org/10.1109/TIV.2022.3225340
11. J. Ahn, M. Kim, J. Park, Autonomous driving using imitation learning with a look ahead point for semi-structured environments, *Sci. Rep.*, **12** (2022), 21285. https://doi.org/10.1038/s41598-022-23546-6
12. B. Zheng, S. Verma, J. Zhou, I. W. Tsang, F. Chen, Imitation learning: Progress, taxonomies and challenges, *IEEE Trans. Neural Networks Learn. Syst.*, (2022), 1–16. https://doi.org/10.1109/TNNLS.2022.3213246
13. Z. Wu, K. Qiu, H. Gao, Driving policies of V2X autonomous vehicles based on reinforcement learning methods, *IET Intell. Transp. Syst.*, **14** (2020), 331–337. https://doi.org/10.1049/iet-its.2019.0457
14. C. You, J. Lu, D. Filev, P. Tsiotras, Advanced planning for autonomous vehicles using reinforcement learning and deep inverse reinforcement learning, *Rob. Auton. Syst.*, **114** (2019), 1–18. https://doi.org/10.1016/j.robot.2019.01.003
15. D. Zhang, X. Han, C. Deng, Review on the research and practice of deep learning and reinforcement learning in smart grids, *CSEE J. Power Energy Syst.*, **4** (2018), 362–370. https://doi.org/10.17775/CSEEJPES.2018.00520
16. Y. H. Khalil, H. T. Mouftah, Exploiting multi-modal fusion for urban autonomous driving using latent deep reinforcement learning, *IEEE Trans. Veh. Technol.*, **72** (2023), 2921–2935. https://doi.org/10.1109/TVT.2022.3217299
17. H. Zhang, Y. Lin, S. Han, K. Lv, Lexicographic actor-critic deep reinforcement learning for urban autonomous driving, *IEEE Trans. Veh. Technol.*, **72** (2023), 4308–4319. https://doi.org/10.1109/TVT.2022.3226579
18. Z. Du, Q. Miao, C. Zong, Trajectory planning for automated parking systems using deep reinforcement learning, *Int. J. Automot. Technol.*, **21** (2020), 881–887. https://doi.org/10.1007/s12239-020-0085-9

19. E. O. Neftci, B. B. Averbeck, Reinforcement learning in artificial and biological systems, *Nat. Mach. Intell.*, **1** (2019), 133–143. https://doi.org/10.1038/s42256-019-0025-4
20. M. L. Littman, Reinforcement learning improves behavior from evaluative feedback, *Nature*, **521** (2015), 445–451. https://doi.org/10.1038/nature14540
21. E. O. Neftci, B. B. Averbeck, Reinforcement learning in artificial and biological systems, *Nat. Mach. Intell.*, **1** (2019), 133–143. https://doi.org/10.1038/s42256-019-0025-4
22. C. Zhu, Y. Cai, J. Zhu, C. Hu, J. Bi, GR(1)-guided deep reinforcement learning for multi-task motion planning under a stochastic environment, *Electronics*, **11** (2022), 3716. https://doi.org/10.3390/electronics11223716
23. J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, preprint, arXiv:1707.06347. https://doi.org/10.48550/arXiv.1707.06347
24. W. Guan, Z. Cui, X. Zhang, Intelligent smart marine autonomous surface ship decision system based on improved PPO algorithm, *Sensors*, **22** (2022), 5732. https://doi.org/10.3390/s22155732
25. J. Han, K. Jo, W. Lim, Y. Lee, K. Ko, E. Sim, et al., Reinforcement learning guided by double replay memory, *J. Sens.*, **2021** (2021), 1–8. https://doi.org/10.1155/2021/6652042
26. H. Liu, A. Trott, R. Socher, C. Xiong, Competitive experience replay, preprint, arXiv:1902.00528. https://doi.org/10.48550/arXiv.1902.00528
27. X. Wang, H. Xiang, Y. Cheng, Q. Yu, Prioritised experience replay based on sample optimization, *J. Eng.*, **2020** (2020), 298–302. https://doi.org/10.1049/joe.2019.1204
28. A. Karalakou, D. Troullinos, G. Chalkiadakis, M. Papageorgiou, Deep reinforcement learning reward function design for autonomous driving in lane-free traffic, *Systems*, **11** (2023), 134. https://doi.org/10.3390/systems11030134
29. B. Geng, J. Ma, S. Zhang, Ensemble deep learning-based lane-changing behavior prediction of manually driven vehicles in mixed traffic environments, *Electron. Res. Arch.*, **31** (2023), 6216–6235. https://doi.org/10.3934/era.2023315
30. M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, Hindsight experience replay, preprint, arXiv:1707.01495.
31. J. Wu, Z. Huang, Z. Hu, C. Lu, Toward human-in-the-loop AI: Enhancing deep reinforcement learning via real-time human intervention for autonomous driving, *Engineering*, **21**(2023), 75–91. https://doi.org/10.1016/j.eng.2022.05.017
32. F. Pan, H. Bao, Preceding vehicle following algorithm with human driving characteristics, *Proc. Inst. Mech. Eng., Part D: J. Automob. Eng.*, **235** (2021), 1825–1834. https://doi.org/10.1177/0954407020981546
33. Y. Zhou, R. Fu, C. Wang, Learning the car-following behavior of drivers using maximum entropy deep inverse reinforcement learning, *J. Adv. Transp.*, **2020** (2020), 1–13. https://doi.org/10.1155/2020/4752651
34. S. Lee, D. Ngoduy, M. Keyvan-Ekbatani, Integrated deep learning and stochastic car-following model for traffic dynamics on multi-lane freeways, *Transp. Res. Part C Emerging Technol.*, **106** (2019), 360–377. https://doi.org/10.1016/j.trc.2019.07.023

AIMS Press