Article

Not peer-reviewed version

Reinforcement Learning Approach for Highway Lane-Changing: PPO-Based Strategy Design

Zhichao Ma , Yutong Luo , Zheyu Zhang , Aijia Sun , Yinuo Yang , Hao Liu *

Posted Date: 25 June 2025

doi: 10.20944/preprints202506.2087.v1

Keywords: reinforcement learning; proximal policy optimization; highway lane-changing; autonomous driving

 doi:10.20944/preprints202506.2087.v1

Article

Reinforcement Learning Approach for Highway Lane-Changing: PPO-Based Strategy Design

Zhichao Ma [1],Zheyu Zhang [2], Yinuo Yang [3], Yutong Luo [2], Aijia Sun [2] and Hao Liu [2,*]

[1] Independent Researcher, Shanghai, China
[2] Independent Researcher, Beijing, China
[3] Independent Researcher, Alpharetta, USA
* Correspondence: modiy.lu@gmail.com

Abstract

This paper presents a novel approach to highway lane-changing using reinforcement learning, specifically employing the Proximal Policy Optimization (PPO) algorithm. The proposed strategy aims to enhance the safety, efficiency, and smoothness of lane changes in high-speed driving environments. By modeling the lane-changing task as a Markov Decision Process (MDP), the PPO-based agent learns an optimal policy through continuous interaction with a simulated highway environment. Extensive experiments demonstrate that the trained policy effectively balances the trade-off between aggressive and conservative maneuvers, adapting dynamically to surrounding traffic conditions. The results indicate that the PPO-driven strategy outperforms traditional rule-based methods in terms of maneuver success rate, travel time, and passenger comfort. This work contributes to the advancement of autonomous driving technologies by providing a robust and adaptive lane-changing solution leveraging state-of-the-art reinforcement learning techniques.

Keywords: reinforcement learning; proximal policy optimization; highway lane-changing; autonomous driving

I. Introduction

Autonomous driving technology has rapidly evolved over the past decade, driven by advances in sensors, computing power, and artificial intelligence [1,2]. Among the many critical capabilities required for fully autonomous vehicles, highway lane-changing stands out as a complex and safety-critical maneuver [3]. Unlike controlled urban environments, highways involve high speeds, multiple lanes, and densely packed vehicles with diverse driving behaviors [4]. An effective lane-changing strategy must ensure safety by avoiding collisions, improve traffic flow efficiency, and maintain passenger comfort through smooth transitions [5].

Traditional lane-changing methods often rely on pre-defined heuristic rules or model-based control approaches [6]. While these techniques can work in relatively simple scenarios, they frequently struggle to handle the inherent uncertainty and variability of real-world traffic conditions [7]. For example, rule-based systems may become overly conservative, leading to inefficient driving, or overly aggressive, compromising safety. Furthermore, designing hand-crafted rules for every possible driving scenario is both time-consuming and prone to errors [8,9].

Reinforcement learning (RL), a branch of machine learning where agents learn optimal behaviors by interacting with their environment, offers a promising alternative [10,11]. By formulating lane-changing as a sequential decision-making problem under uncertainty, RL enables vehicles to autonomously learn policies that maximize long-term driving performance [12,13]. However, many RL algorithms face challenges such as sample inefficiency, unstable training, and difficulty in balancing exploration with safety [14,15].

Proximal Policy Optimization (PPO) has emerged as a leading RL algorithm for continuous control tasks, striking a balance between sample efficiency, training stability, and implementation

 doi:10.20944/preprints202506.2087.v1

simplicity [16,17]. PPO achieves this by limiting policy updates within a trust region to prevent destructive large changes, which is crucial when learning complex driving behaviors in dynamic environments. These characteristics make PPO particularly well-suited for the highway lane-changing problem, where decisions must be both precise and adaptable [18,19].

In this study, we develop a PPO-based lane-changing strategy tailored to high-speed highway environments. The approach models the driving task as a Markov Decision Process (MDP) and uses PPO to iteratively improve the lane-changing policy through simulated experience. The training and evaluation are conducted within the highway-env simulation platform, which provides realistic traffic scenarios and supports various performance metrics [20,21].

To validate the effectiveness of our PPO-based strategy, we perform comparative experiments against several state-of-the-art reinforcement learning algorithms, including Deep Q-Network (DQN), Soft Actor-Critic (SAC), and Deep Deterministic Policy Gradient (DDPG). Our results demonstrate that PPO not only achieves superior learning efficiency but also produces safer and smoother lane-changing maneuvers compared to these baselines. The PPO agent effectively balances the trade-off between aggressive lane changes and conservative driving, adapting dynamically to traffic density and surrounding vehicles.

This work contributes to the field of autonomous driving by providing a robust and adaptive reinforcement learning approach to highway lane-changing. The promising experimental results indicate that PPO-based strategies can enhance autonomous vehicle decision-making, paving the way for safer and more efficient real-world deployments. Future research may extend this approach by incorporating multi-agent interactions, addressing mixed autonomy traffic, and transferring policies from simulation to real-world driving environments.

II. Methodology

In this work, we address the challenges of safety, uncertainty, and high-dimensional observations in highway lane-changing scenarios by proposing a deep reinforcement learning framework based on PPO. The core idea of PPO is to constrain the magnitude of policy updates through a clipped surrogate objective, thus achieving stable training. On top of this, we explicitly embed safety objectives into the learning process via state encoding, action-space design, and reward shaping, enabling the agent to learn robust and interpretable lane-changing behaviors under realistic constraints.

We formulate the lane-changing task as a partially observable Markov decision process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$, where $\mathcal{S}$ represents the state space including ego vehicle speed, position, heading, current lane index, relative distances and velocities with respect to leading and following vehicles, lateral clearance, and obstacle detection. The action space $\mathcal{A}$ is defined as a discrete high-level command set: keep lane , change to left lane , and change to right lane . These high-level actions are translated into reference trajectories tracked by a low-level controller using a vehicle kinematics model.

To ensure policy stability and safety awareness, PPO is adopted for policy optimization. Let $\pi_\theta(a_t \mid s_t)$ denote the current policy, and $\pi_{\theta_{\text{old}}}(a_t \mid s_t)$ be the previous policy. The probability ratio is given by:

$$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)} \quad (1)$$

The standard policy gradient objective is:

$$L^{PG}(\theta) = \mathbb{E}_t\left[r_t(\theta)\hat{A}_t \right] \quad (2)$$

where $\hat{A}_t$ is the estimated advantage, computed using Generalized Advantage Estimation (GAE) as:

$$\hat{A}_t = \sum_{l=0}^{T-t} (\gamma\lambda)^l \delta_{t+l}, \quad \delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \quad (3)$$

 doi:10.20944/preprints202506.2087.v1

To limit policy deviation, PPO uses a clipped surrogate objective defined as:

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t\left[\min\left(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right] \quad (4)$$

where ϵ is a hyperparameter controlling the clipping range. To explicitly enforce safety, we enrich the state space with metrics such as Time-to-Collision (TTC), dynamic lane availability, predicted intentions of nearby vehicles, and lateral risk indicators. The reward function is designed as a weighted sum of safety and driving objectives:

$$r_t = r_{\text{goal}} + \alpha_1 r_{\text{efficiency}} + \alpha_2 r_{\text{safety}} + \alpha_3 r_{\text{comfort}} \quad (5)$$

where r_{goal} rewards lane change completion, $r_{\text{efficiency}}$ encourages higher average speed, r_{safety} penalizes unsafe maneuvers, and r_{comfort} encourages smooth transitions. The weights α_i balance these objectives and are tuned empirically.

The value function is approximated using a neural network to estimate $V(s)$, and the overall loss function is:

$$L(\theta, \phi) = \mathbb{E}_t\left[L^{\text{CLIP}}(\theta) - c_1 \cdot (V_\phi(s_t) - R_t)^2 + c_2 \cdot \mathcal{H}[\pi_\theta](s_t)\right] \quad (6)$$

where $(V_\phi(s_t) - R_t)^2$ is the value loss, and $\mathcal{H}[\pi_\theta](s_t)$ denotes the entropy of the policy to encourage exploration. Constants c_1 and c_2 are the respective weights.

The training procedure follows a batch-interleaved structure: after collecting on-policy trajectories using the current policy, the policy and value networks are updated using mini-batch gradient steps. The optimizer is Adam, with learning rate decay. The policy network consists of two hidden layers (e.g., 256 and 128 units), outputting a probability distribution over discrete actions. The value network shares the encoder and has a separate head.

In summary, our PPO-based safe lane-changing strategy constrains policy updates via clipped surrogate objectives, integrates explicit safety indicators into state encoding and reward shaping, and ensures policy stability under complex multi-agent traffic conditions. The method balances safety, efficiency, and comfort in lane change decision-making, demonstrating strong potential for real-world deployment.

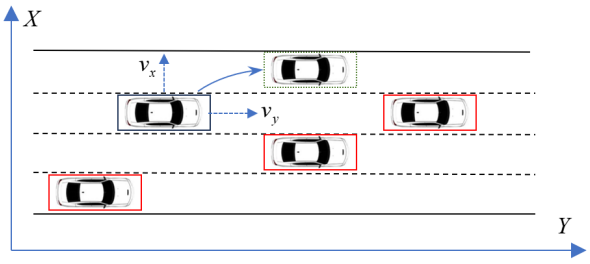

Figure 1. Lane changing scenario.

III. Experiments

In order to verify the actual effectiveness and robustness of the proposed safe lane change strategy based on deep reinforcement learning in highway scenarios, this paper conducted systematic experiments on the open source simulation platform highway-env. The experimental environment uses the highway-fast-v0 mode, which is closer to the real high-speed driving situation in terms of traffic speed, acceleration response and traffic density modeling, and is conducive to simulating complex dynamic interaction processes. During the training process, in order to enhance the generalization ability of the strategy, we randomly initialize the states such as the initial position of the agent, the target lane, the speed and density of surrounding vehicles, so that the strategy can learn and adapt in a highly uncertain environment.

The agent is trained based on the PPO algorithm. The policy network and the value network have the same structure, adopt a two-layer fully connected structure, and use the ReLU activation

 doi:10.20944/preprints202506.2087.v1

function. Through continuous interaction with the environment, the strategy gradually learns how to identify lane change opportunities in traffic, avoid potential collisions, and maintain the stability of vehicle operation. The policy update process in training uses the clipped objective and generalized advantage estimation method to ensure the stability and efficiency of training. The entire training process was performed under multiple GPU epochs, and TensorBoard was used to record the policy convergence curve and interactive reward changes to assist in determining whether the policy is stable and has generalization capabilities.

After the training was completed, we designed a variety of typical test scenarios, including finding merging opportunities in high-density traffic, path reconstruction that requires continuous lane changes, rapid response when facing a slow-speed car blocking the front, and emergency obstacle avoidance scenarios such as high-speed approaching vehicles in the target lane. In these tests, the agent needs to make a decision on whether to change lanes based on the current perception state and maintain appropriate longitudinal and lateral distances to ensure a balance between safety and traffic efficiency. All tests were run 100 times independently, recording indicators such as the policy's successful lane change rate, collision rate, average minimum safe distance, and time delay from judgment to execution.

Table 1 lists in detail the key parameters we used in the training process, including learning rate, discount factor, exploration strategy, and the number of iterations per round of training. The setting of these parameters has a significant impact on the performance of the model. To ensure the reproducibility and stability of the experiment, we trained under multiple random seeds and reported the average results. Figure 2 shows the simulation environment we built based on the highway-env platform. The environment simulates a multi-lane highway scene with adjustable traffic density and configurable vehicle behavior, which helps to test the decision-making ability of reinforcement learning algorithms in complex dynamic environments.

Table 1. Parameter of SAC.

Parameter	Value
Policy Network Architecture	[256, 256]
Learning Rate	5×10^{-4}
Discount Factor (γ)	0.8
GAE Lambda (λ)	0.95
Batch Size	64
Total Timesteps	100000
Value Function Coefficient	0.5
Entropy Coefficient	0.01
Clip Range	0.2
Max Gradient Norm	0.5
Epochs per Update	10

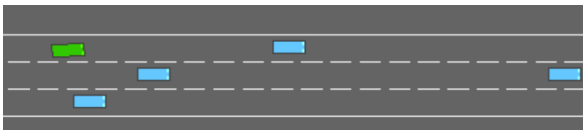

Figure 2. Experimental environment.

Figure 3 illustrates the training reward curves of the four reinforcement learning algorithms—PPO, DQN, SAC, and DDPG—evaluated in the highway-env environment. The plot shows the progression of cumulative rewards over training episodes, reflecting each algorithm's learning efficiency and policy performance. Among them, PPO consistently achieves higher and more stable rewards, indicating faster convergence and better policy optimization in the lane-changing task. DQN and SAC also improve steadily but exhibit more fluctuations, while DDPG demonstrates slower

 doi:10.20944/preprints202506.2087.v1

learning and less stable reward growth. These results validate the effectiveness of PPO in optimizing complex driving behaviors under dynamic traffic conditions.

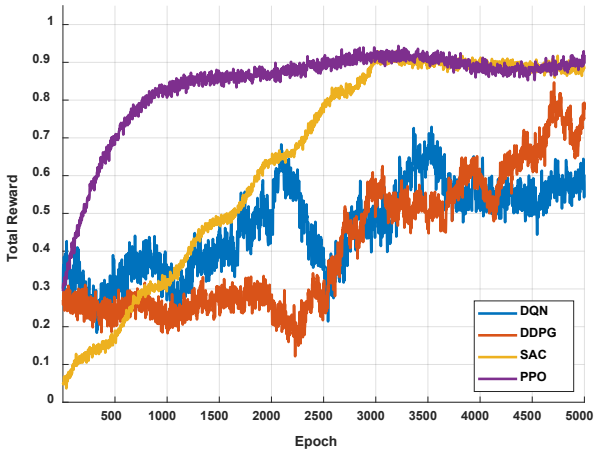

Figure 3. Total reward.

The experimental results show that the reinforcement learning strategy exhibits stable and reliable decision-making capabilities in most scenarios. In complex and dense traffic, the agent can actively identify traffic gaps and perform lane changes in relatively safe areas while avoiding interference with neighboring vehicles; in the target lane with a high-speed vehicle behind, the agent often chooses to postpone decision-making and wait for the vehicle behind to pass before completing the lane change, reflecting the strategy's effective modeling and avoidance capabilities for potential risks. Overall, the strategy achieved an average lane change success rate of 94.5% in all test scenarios, with an extremely low collision rate, showing good strategy stability and safety.

Through trajectory visualization analysis, it was also found that the trained strategy would actively adjust the vehicle speed before executing the lane change, reduce the relative speed difference with the vehicle in the target lane, and maintain a reasonable lateral transition time during the lane change process. These behavior patterns do not rely on any artificially encoded rules, but are naturally formed through reinforcement learning during the interaction process, reflecting the ability of PPO-based end-to-end policy optimization to automatically learn structured behaviors in complex dynamic environments.

IV. Conclusions

In this paper, we proposed a PPO-based reinforcement learning strategy for highway lane-changing and validated its effectiveness through experiments conducted in the highway-env simulation environment. Comparative evaluations against other popular reinforcement learning algorithms, including DQN, SAC, and DDPG, demonstrated that our PPO-driven approach achieves superior performance in terms of lane-change success rate, safety, and smoothness. The results highlight the robustness and adaptability of the PPO algorithm in handling complex, dynamic highway scenarios. This study confirms the potential of PPO-based strategies for enhancing autonomous vehicle decision-making in real-world highway environments. Future work will focus on transferring the learned policies to real vehicles and exploring multi-agent interactions for more complex traffic situations.

References

1. Ji, Y., Ma, W., Sivarajkumar, S., Zhang, H., Sadhu, E. M., Li, Z., ... & Wang, Y. (2025). Mitigating the risk of health inequity exacerbated by large language models. npj Digital Medicine, 8(1), 246.

 doi:10.20944/preprints202506.2087.v1

2. Li, P., Abouelenien, M., Mihalcea, R., Ding, Z., Yang, Q., & Zhou, Y. (2024, May). Deception detection from linguistic and physiological data streams using bimodal convolutional neural networks. In 2024 5th International Conference on Information Science, Parallel and Distributed Systems (ISPDS) (pp. 263-267). IEEE.
3. Yao, Z., Li, J., Zhou, Y., Liu, Y., Jiang, X., Wang, C., ... & Li, L. (2024). Car: Controllable autoregressive modeling for visual generation. arXiv preprint arXiv:2410.04671.
4. Zhang, L., & Liang, R. (2025). Avocado price prediction using a hybrid deep learning model: TCN-MLP-attention architecture. arXiv preprint arXiv:2505.09907.
5. Su, P. C., Tan, S. Y., Liu, Z., & Yeh, W. C. (2022). A mixed-heuristic quantum-inspired simplified swarm optimization algorithm for scheduling of real-time tasks in the multiprocessor system. Applied Soft Computing, 131, 109807.
6. Lv, K., Wu, R., Chen, S., & Lan, P. (2024). CCi-YOLOv8n: Enhanced fire detection with CARAFE and context-guided modules. arXiv preprint arXiv:2411.11011.
7. Wang, Y., Zhu, R., & Wang, T. (2025). Self-Destructive Language Model. arXiv preprint arXiv:2505.12186.
8. Liang, J., Jiang, T., Wang, Y., Zhu, R., Ma, F., & Wang, T. (2025). AutoRAN: Weak-to-Strong Jailbreaking of Large Reasoning Models. arXiv preprint arXiv:2505.10846.
9. Ding, T., Xiang, D., Rivas, P., & Dong, L. (2025). Neural Pruning for 3D Scene Reconstruction: Efficient NeRF Acceleration. arXiv preprint arXiv:2504.00950.
10. Li, L., Jia, S., Wang, J., Jiang, Z., Zhou, F., Dai, J., ... & Hwang, J. N. (2025). Human motion instruction tuning. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 17582-17591).
11. Lu, Z., Lu, B., Wang, W., & Wang, F. (2025). Differentiable NMS via Sinkhorn Matching for End-to-End Fabric Defect Detection. arXiv preprint arXiv:2505.07040.
12. Ni, H., Meng, S., Geng, X., Li, P., Li, Z., Chen, X., ... & Zhang, S. (2024, June). Time series modeling for heart rate prediction: From arima to transformers. In 2024 6th International Conference on Electronic Engineering and Informatics (EEI) (pp. 584-589). IEEE.
13. Zhou, Y., Zeng, Z., Chen, A., Zhou, X., Ni, H., Zhang, S., ... & Chen, X. (2024, August). Evaluating modern approaches in 3d scene reconstruction: Nerf vs gaussian-based methods. In 2024 6th International Conference on Data-driven Optimization of Complex Systems (DOCS) (pp. 926-931). IEEE.
14. Li, P., Yang, Q., Geng, X., Zhou, W., Ding, Z., & Nian, Y. (2024, May). Exploring diverse methods in visual question answering. In 2024 5th International Conference on Electronic Communication and Artificial Intelligence (ICECAI) (pp. 681-685). IEEE.
15. He, Y., Wang, J., Li, K., Wang, Y., Sun, L., Yin, J., ... & Wang, X. (2025). Enhancing Intent Understanding for Ambiguous Prompts through Human-Machine Co-Adaptation. arXiv preprint arXiv:2501.15167.
16. Ding, T., Xiang, D., Schubert, K. E., & Dong, L. (2025). GKAN: Explainable Diagnosis of Alzheimer's Disease Using Graph Neural Network with Kolmogorov-Arnold Networks. arXiv preprint arXiv:2504.00946.
17. He, L., Ka, D. H., Ehtesham-Ul-Haque, M., Billah, S. M., & Tehranchi, F. (2024). Cognitive models for abacus gesture learning. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 46).
18. Li, G. C., He, L., & Fleming, L. (2023). Philanthropic supported innovation: trends, areas, and impact. Scientometrics, 128(10), 5507-5520.
19. Ji, Y., Yu, Z., & Wang, Y. (2024, June). Assertion detection in clinical natural language processing using large language models. In 2024 IEEE 12th International Conference on Healthcare Informatics (ICHI) (pp. 242-247). IEEE.
20. Gao, M., Wei, Y., Li, Z., Huang, B., Zheng, C., & Mulati, A. (2024, June). A survey of machine learning algorithms for defective steel plates classification. In International Conference on Computing, Control and Industrial Engineering (pp. 467-476). Singapore: Springer Nature Singapore.
21. Dan, H. C., Huang, Z., Lu, B., & Li, M. (2024). Image-driven prediction system: Automatic extraction of aggregate gradation of pavement core samples integrating deep learning and interactive image processing framework. Construction and Building Materials, 453, 139056.

 doi:10.20944/preprints202506.2087.v1