Proceedings of CONF-SEML 2025 Symposium: Machine Learning Theory and Applications
DOI: 10.54254/2755-2721/2025.TJ23323

Reinforcement Learning Methods for Autonomous Driving: A Survey

Yujie Lin

Department of Computer Science and Technology, Beijing Jiaotong University, Beijing, China
22722019@bjtu.edu.cn

Abstract: In recent years, with the rapid development of intelligent transportation, Reinforcement Learning (*RL*), as an adaptive decision-making method, has gradually permeated into various levels of Autonomous Driving (*AD*). Therefore, this paper reviews the latest advances in the application of *RL* in *AD*. In terms of high-level decision-making and behavioral planning, *RL*, combined with visual-language models, imitation learning, multi-stage training, and autoregressive trajectory planning, systematically improves planning accuracy and task success rates. At the motion control level, the synergistic optimization of deep reinforcement learning (*DRL*) based continuous control strategies and robust control methods enhances performance in path tracking, dynamic obstacle avoidance, and multi-sensor information fusion. Meanwhile, end-to-end autonomous driving leverages novel frameworks such as closed-loop *RL*, World Model (*WM*), and multimodal decision fusion, effectively narrowing the gap between simulation and real-world environments while achieving significant improvements in safety and smoothness. Additionally, the paper discusses the limitations of *RL* applications, including data dependency, training efficiency, safety, and interpretability. Furthermore, it explores the future prospects for achieving more intelligent autonomous driving systems through strategies such as meta-learning, transfer learning, adversarial training, and human-machine collaboration.

Keywords: Reinforcement Learning (*RL*), Autonomous Driving (*AD*), high-level decision-making, motion control, end-to-end.

1. Introduction

As Autonomous Driving (*AD*) advances toward higher levels of intelligence and autonomous decision-making, it has become a critical technical challenge to achieve accurate perception, rapid decision-making, and stable control in highly dynamic and complex traffic environments. Traditional *AD* researches often rely on modular designs that independently develop environmental perception, high-level decision-making, and motion control. Nevertheless, in complex road scenarios of the real world, these approaches struggle to achieve efficient coordination among modules. In recent years, Reinforcement Learning (*RL*) has been progressively introduced into *AD* systems due to its unique advantages in handling continuous control and uncertain environments. It can not only optimize the performance of individual modules, but also enhance overall system performance through end-to-end decision training.

Specifically, in high-level decision-making and behavioral planning, *RL* integrates language models, imitation learning, multi-stage training, and autoregressive trajectory planning to improve

DOI: 10.54254/2755-2721/2025.TJ23323

semantic understanding and behavioral diversity, significantly enhancing decision-making accuracy in complex road conditions [1-5]. For motion control, the fusion of *RL* and traditional control theory improves vehicle responsiveness in tasks such as path tracking and emergency obstacle avoidance [6-8]. Meanwhile, end-to-end autonomous driving methods integrate multi-sensor information and leverage closed-loop training and world model algorithms to reduce errors effectively from perception to execution [9-12]. Although current methods have achieved notable success in simulation environments and some real-vehicle testing, their practical applications are still hindered by challenges such as data dependency, training sample complexity, safety, and interpretability.

In the following sections, the paper will comprehensively review the application of *RL* in key aspects of *AD*, analyze existing limitations, and provide an in-depth discussion on future technological directions, thereby offering theoretical references and practical guidance for achieving robust autonomous driving systems.

2. Applications of *RL* in key aspects of *AD*

The development of Autonomous Driving technology is undergoing an evolution from modular architecture to end-to-end systems. Traditional *AD* system is usually composed of three core modules: environment perception, high-level decision-making and behavior planning, and motion control. The end-to-end rule realizes the direct mapping from sensor input to control output through deep learning. In this process, *RL* is gradually becoming the key technology to improve the performance of *AD* systems due to its unique advantages in complex decision-making problems.

This section reviews the progress of *RL* applications in autonomous driving, focusing on three key components: high-level decision-making and behavioral planning, motion control, and end-to-end driving. It also analyzes the characteristics and experimental results of various representative methods.

2.1. High-level decision-making and behavioral planning

In high-level decision-making and behavioral planning, *RL* has evolved from optimizing a single algorithm to an innovative paradigm characterized by the deep fusion of multiple technologies, significantly advancing the system's decision-making capabilities in complex dynamic environments. Current researches primarily focus on improving semantic understanding, optimizing behavioral diversity, enhancing trajectory planning efficiency, and refining environment modeling accuracy. By incorporating state-of-the-art learning algorithms and model architectures, these advances are driving *AD* systems toward higher levels of intelligence and reliability.

In terms of enhancing semantic understanding, AlphaDrive integrates *GRPO*-based algorithm with four structured reward functions (planning accuracy reward, action-weighted reward, planning diversity reward, and planning format reward) to embed the semantic understanding capabilities of a visual language model (*VLM*) into *RL* framework. The framework employs a two-stage training strategy. Specifically, supervised fine-tuning (*SFT*) is used to distill planning inference knowledge from a large-scale model, followed by further optimization of planning performance via *RL*. Experiments indicate that AlphaDrive achieves a planning accuracy of 77.12% on the MetaAD dataset and represents a 25.5% improvement over the suboptimal model, significantly elevating safety and efficiency in complex driving scenarios [1]. Similarly, the LGDRL framework incorporates the Large Language Model (*LLM*) as a driving expert into deep Reinforcement Learning (*DRL*). It constrains *RL* policy deviation from *LLM* expert policy through JS-divergence and intermittently substitutes risky actions of the *RL* agent with *LLM*-mediated interventions, achieving a 90% task success rate in highway scenarios while delivering inference speeds far superior to those of the *LLM* expert [2].

DOI: 10.54254/2755-2721/2025.TJ23323

Regarding the optimization of behavioral diversity, the *MRIC* framework innovatively combines *RL* with imitation learning (*IL*). It leverages a differentiable simulator for efficient state matching, thus addressing the instability issues encountered in traditional training approaches. Furthermore, it introduces a hybrid encoder module that employs discrete embedding space and dynamic prior distribution to capture the diversity of driving behaviors, mitigating the generation of unrealistic actions in low-probability regions. Experiments on the Waymo Open Motion Dataset show that *MRIC* outperforms baseline models in terms of collision rate and minimum *SADE*, which confirms its advantages in modeling diverse and realistic driving behaviors [3].

For trajectory planning efficiency, CarPlanner proposes an autoregressive trajectory planning approach aimed at overcoming the low training efficiency and suboptimal performance of large-scale *RL* in autonomous driving. By introducing a consistent mode representation and generation-selection framework, combined with an expert-guided reward function and an invariant view module (*IVM*), the method achieves efficient and stable multimodal trajectory generation. Specifically, this method ensures coherence between time steps by decomposing longitudinal and lateral behaviors. In addition, the use of non-reactive transformation models further simplifies the training process, improving training efficiency and the generalization ability of strategies. Experimental results on the nuPlan dataset reveal that CarPlanner is superior to existing rule-based, imitation learning, and *RL* methods, which demonstrates its potential in complex real-world driving scenarios [4].

With respect to environment modeling accuracy improvement, Imagine-2-Drive combines a high-fidelity world model with a diffusion policy actor (*DPA*). The world model predicts future states and rewards to provide precise environment modeling, while DPA utilizes the diffusion model to generate multimodal behavior patterns, thereby enhancing the diversity and robustness of trajectory planning. Optimized through a *DDPO* strategy to maximize cumulative rewards, Imagine-2-Drive yields a better result than existing methods in the *CARLA* simulator, achieving improvements of 15% in route completion rate and 20% in overall success rate. This offers an efficient and flexible solution for long-horizon trajectory planning in autonomous driving [5].

Overall, recent applications of reinforcement learning in high-level decision-making and behavior planning for autonomous driving are increasingly driven by synergistic innovations across multiple technologies. While AlphaDrive and *LGDRL* enhance semantic understanding through language models, *MRIC* improves behavioral diversity via imitation learning integration. CarPlanner optimizes planning efficiency, whereas Imagine-2-Drive refines environmental modeling precision. These innovations not only address domain-specific challenges, but collectively provide comprehensive technical support for building more intelligent and reliable *AD* decision systems.

2.2. Motion control

RL offers novel approaches to address the limitations of traditional control methods in complex and dynamic environments. Current research primarily focuses on three key areas: the design of continuous control strategies based on deep reinforcement learning (*DRL*), the collaborative optimization of robust control and *RL*, and the enhancement of *RL* algorithms tailored for emerging electronic and electrical (*E*/*E*) architectures. These studies effectively enhance autonomous vehicles' performance in path tracking, dynamic obstacle avoidance, and stability control, thereby laying a solid foundation for developing safer and more efficient motion control systems.

In the realm of continuous control strategy design, significant progress has been made using *DRL*. Researchers have conducted systematic comparisons between Deep *Q*-Network (*DQN*) and Deep Deterministic Policy Gradient (*DDPG*) algorithms within the *CARLA* simulation environment. Through constructing comprehensive vehicle models encompassing both longitudinal and lateral dynamics, various *DRL* architectures were explored, incorporating visual inputs, waypoint information, and convolutional neural networks. Experimental results indicate that *DDPG*

DOI: 10.54254/2755-2721/2025.TJ23323

outperforms DQN in training efficiency and path tracking accuracy, closely mirroring human driving behavior [6]. Notably, architectures based on waypoint information demonstrated superior performance across all tested scenarios, which provides valuable insights for developing more efficient motion control algorithms.

From the perspective of integrating robust control with *RL*, recent studies have combined robust control theory with the Proximal Policy Optimization (*PPO*) algorithm to achieve high-performance vehicle dynamics control. A robust controller based on $H\infty$ method was designed, which employs a supervisor to execute constrained quadratic programming tasks, ensuring safety in critical performance metrics such as lateral error. Simultaneously, the *PPO* algorithm optimized dynamic performance indicators, such as high-speed driving and path tracking. Experimental validations demonstrate that this hybrid control approach excels in both simulation and real-world tests, particularly in suppressing lateral acceleration. This offers a novel technical pathway for motion control in autonomous driving [7].

To address the challenges posed by new *E*/*E* architectures, improved *DRL* frameworks have been developed. Researchers first conducted precise modeling of domain-centralized *E*/*E* architectures and vehicle motion control problems, followed by latency analysis multi-hop control loops to determine the theoretical boundary values of loop delays in heterogeneous topologies. Based on this analysis, an enhanced heuristic experience replay mechanism was innovatively proposed, integrating estimated loop delay values into the motion controller optimization process. To promote algorithm convergence, the method combined Nesterov accelerated gradient with adaptive moment estimation techniques. Through a combination of virtual and real-world testing, the framework not only improved control performance but also ensured system robustness against inherent delays in *E*/*E* architectures, which provides a crucial reference for motion control of the next-generation *AD* systems [8].

In summary, the application of *RL* in autonomous vehicle motion control has yielded substantial results. From the collaborative optimization of robust control and *RL*, to algorithm enhancements tailored for emerging *E*/*E* architectures, together with the design of continuous control strategies based on *DRL*, recent studies reflect notable progress on multiple fronts. These works highlight three key characteristics: a strong integration of control theory with learning algorithms, innovative solutions to real-world engineering challenges, and rigorous validation through comprehensive simulation and real-world testing. Collectively, they propel the development of motion control technologies in autonomous driving.

2.3. End-to-end driving

In end-to-end autonomous driving systems, *RL* has gradually emerged as a crucial tool for constructing efficient decision-making and control mechanisms. Unlike traditional modular approaches, end-to-end *RL* methods optimize direct mapping from sensor inputs to control outputs to enhance the performance of the overall system. In recent years, numerous studies have explored how to use *RL* to address dynamic interactions in complex traffic environments, fuse multi-sensor information, and improve the understanding of causal reasoning. In response, researchers have proposed various innovative solutions.

First, in order to address the shortcomings of conventional imitation learning (*IL*) techniques in causal understanding and open-loop deployment, a reinforcement learning framework based on 3*D* Gaussian mapping (3*DGS*) was first presented. In particular, the approach builds a realistic digital twin environment to enable the policy to conduct substantial trial-and-error exploration of the state space. In the meantime, a specially built safety-related reward function allows the policy to handle major risk events. Additionally, IL is added as a regularization term during training to bring *RL* closer to human behavior, resulting in a synergistic synergy between *RL* and *IL*. According to the results,

DOI: 10.54254/2755-2721/2025.TJ23323

the method reduces collision rates by about 3 times in closed-loop tests and remarkably increases trajectory consistency and driving smoothness [9].

Second, the CarDreamer platform was developed to support world-model-based *RL* algorithms. This open-source platform integrates advanced world model techniques and provides a standardized Gym interface. It offers highly configurable tasks, an optimized reward function, a flexible task development suite, and a visualization server. Experiments on the *CARLA* simulator demonstrate that the platform effectively enhances system safety and efficiency under various observation modalities and communication configurations [10]. CarDreamer not only offers researchers a convenient tool for development and evaluation but also lays a solid foundation for future RL algorithm research in autonomous driving.

Third, to tackle the challenges of highly interactive traffic environments, the Ramble algorithm adopts a fully end-to-end, model-based *RL* framework. The approach transforms multi-view *RGB* images and LiDAR point cloud data into compact latent representations that encapsulate the overall traffic scene. In addition, it utilizes a transformer-based structure to capture temporal dependencies, enabling precise predictions of future states. Experiments on *CARLA* Leaderboard 1.0 and 2.0 indicate that Ramble achieves exceptional results in route completion and driving scores, underscoring its effectiveness and robustness in complex, dynamic traffic scenarios [11].

In addition, the PolicyFuser method was proposed to solve the challenge of multi-sensor decision fusion. The method preserves each sensor's independent decision to avoid the overhead of complex feature alignment. Specifically, *RL* is used to select the action corresponding to the optimal *Q*-value as the primary decision, while outputs from other sensors fine-tune this decision. To improve fusion stability, a conditional variational autoencoder (*CVAE*) generates pseudo-expert decisions, effectively reducing discrepancies among sensor decisions. Experimental results indicate that PolicyFuser not only attained the highest driving score in *CARLA* tests but also demonstrated excellent sensor fault tolerance, offering robust support for sensor redundancy design in real-world applications [12].

To sum up, current research on *RL* methods in end-to-end *AD* exhibits a diverse trend. Closed-loop *RL* framework built on 3*D* Gaussian mapping, CarDreamer platform that leverages world models, and solutions such as the Ramble algorithm for highly interactive scenarios as well as PolicyFuser's multimodal strategy fusion have each overcome the limitations of conventional methods at different levels. Together, they established an efficient connection between environment perception and control output.

3. Limitations and future prospects

3.1. Limitations

RL has demonstrated significant potential in all key areas of *AD*. However, some limitations remain in its applications in the three aspects mentioned above.

In high-level decision-making and behavior planning, *RL* methods rely heavily on large amounts of labeled data. The collection and annotation of such data are costly, and insufficient data diversity often leads to degraded performance in long-tail scenarios [1]. Moreover, training efficiency is a critical issue. Traditional model-free methods require repeatedly calling simulators for data collection. Coupled with limited generalization beyond closed-loop environments, these factors pose substantial challenges for real-world applications [4]. Additionally, the opaque, black-box nature of *RL* models compromises both safety and interpretability, while inadequate multimodal behavior modeling restricts their adaptation to complex traffic scenarios [1, 3].

In the domain of motion control, current algorithms generally perform well in specific simulated environments. However, when confronted with previously unseen traffic situations or extreme weather conditions, their generalization capabilities are often insufficient [6, 7, 8]. Furthermore, *RL*

DOI: 10.54254/2755-2721/2025.TJ23323

training typically demands a large number of samples. Algorithms like *DQN* and *DDPG* have high sample complexity, and the significant discrepancies between simulated and real-world environments affect the reliability of these models on actual roads. In addition, the requirements for real-time performance, along with the limitations of embedded hardware resources, make it challenging for complex deep networks to meet strict latency constraints. Consequently, safety and robustness under emergency conditions remain deficient [7, 8].

For end-to-end autonomous driving applications, *RL* faces major challenges including low sample efficiency due to the high-dimensional state space, making real-world tests both expensive and fraught with safety risks [9]. Moreover, the model's generalization performance in new scenarios or adverse weather conditions is often poor; dangerous behaviors may occur, potentially violating traffic rules. The high computational demands further constrain its widespread adoption [10, 11, 12]. In addition, the alignment problem between *RL*-learned policies and human driving behavior needs urgent resolution, as policies often diverge from human driving habits, thereby reducing social acceptance [9].

3.2. Future prospects

Despite these limitations, the future of *RL* in the three aspects analyzed earlier still remains promising, and it is expected to be improved through various technological innovations.

As for high-level decision-making and behavior planning, advanced architectures and algorithms, such as those integrating generative adversarial networks (*GAN*) with diffusion models, can be developed to enhance model generalization and multimodal modeling capabilities. Meanwhile, model-based *RL* methods for multi-step planning may improve training efficiency and closed-loop performance [3, 4]. Besides, the application of data augmentation, synthetic data generation, transfer learning, and meta-learning can help reduce reliance on extensively annotated datasets. Moreover, combining large language models with multimodal perception techniques and incorporating imitation learning can further enhance decision accuracy, interpretability, and safety [1, 4]. Robustness-enhancing mechanisms such as dynamic safety constraints and adversarial training, along with cross-domain deployment approaches, also offer new directions for future development [5].

In the domain of motion control, the integration of multi-agent *RL* and digital twin technologies to construct more realistic simulation environments holds the potential to improve generalization capabilities of models in complex, dynamic scenarios [8]. Research on meta-learning and efficient optimizers, as well as improvements in experience replay mechanisms, is also expected to optimize training efficiency. Furthermore, lightweight network designs and hardware acceleration techniques can effectively satisfy real-time requirements, ensuring rapid decision-making without compromising safety [7, 8].

Regarding end-to-end autonomous driving applications, a key direction lies in improving sample efficiency. By fusing imitation learning, meta-learning, and transfer learning, it is possible to derive high-quality policies with fewer data [10]. Constructing more representative simulation environments and utilizing advanced feature extraction techniques can enhance the model's adaptability to unseen scenarios. Moreover, the incorporation of more intelligent reward design and constraints, along with the optimization at hardware levels can further boost safety and computational efficiency. Finally, leveraging human feedback and preference learning may help align model policies with human driving behaviors, thereby improving ride quality and social acceptance [9, 11, 12].

Therefore, through interdisciplinary technological integration and continuous exploration of innovation methods, *RL* is poised to drive the evolution of *AD* systems toward enhanced safety, robustness and practicality.

DOI: 10.54254/2755-2721/2025.TJ23323

4. Conclusion

In summary, *RL* methods have shown great potential in complex autonomous driving scenarios due to their superior exploration and decision-making capabilities.

In the area of high-level decision-making and behavior planning, current research exhibits a trend toward multi-technology synergistic innovation. Researchers have advanced semantic understanding through language models, improved behavioral diversity with *IL*, optimized trajectory planning efficiency with autoregressive architectures, and enhanced environment modeling accuracy through WMs. Together, these approaches promote the development of more intelligent and reliable decision systems.

In the domain of motion control, *DRL* algorithms (*e.g. DDPG*) have demonstrated continuous control performance that surpasses traditional methods (*e.g. DQN*). Besides, the collaborative design combining *H∞* with *PPO* has improved system stability. Furthermore, optimization algorithms designed for E/E architectures have addressed real-time challenges, resulting in accurate and stable performance in path tracking and dynamic obstacle avoidance.

In end-to-end autonomous driving, 3*DGS* allows closed-loop *RL* training, while the CarDreamer platform provides a standardized development environment based on WMs. The transformer-based Ramble algorithm drives the system to handle complex scenarios, whereas PolicyFuser innovatively addresses the challenge of fusing decisions from multiple sensors. These methods collectively establish an efficient integration among environment perception, decision planning, and motion control, thereby elevating overall system performance.

Despite the encouraging progress achieved in simulations and partial real-world validations, current research still faces several challenges, including the trade-off between data dependency and training efficiency, the gap between simulated and real-world environments, and the assurance of model safety and interpretability. Future work may address these challenges by integrating meta-learning and transfer learning to develop more efficient model architectures and training paradigms. It may also involve constructing more realistic digital twin environments to reduce the simulation-to-reality gap, improving formal verification and adversarial training techniques to ensure safety, and optimizing human-machine collaboration strategies to enhance social acceptance. These advances will collectively advance the application of *RL* in *AD* toward a higher level of safety, robustness, and practicality, thereby laying a solid foundation for the realization of fully autonomous driving.

References

[1] *Jiang, B., Chen, S., Zhang, Q., Liu, W., & Wang, X. (2025). AlphaDrive: Unleashing the Power of VLMs in Autonomous Driving via Reinforcement Learning and Reasoning. arXiv preprint arXiv:2503.07608.*

[2] *Pang, H., Wang, Z., & Li, G. (2024). Large Language Model guided Deep Reinforcement Learning for Decision Making in Autonomous Driving. arXiv preprint arXiv:2412.18511.*

[3] *He, B., & Li, Y. (2024). MRIC: Model-Based Reinforcement-Imitation Learning with Mixture-of-Codebooks for Autonomous Driving Simulation. arXiv preprint arXiv:2404.18464.*

[4] *Zhang, D., Liang, J., Guo, K., Lu, S., Wang, Q., Xiong, R., ... & Wang, Y. (2025). CarPlanner: Consistent Auto-regressive Trajectory Planning for Large-scale Reinforcement Learning in Autonomous Driving. arXiv preprint arXiv:2502.19908.*

[5] *Garg, A., & Krishna, K. M. (2024). Imagine-2-Drive: High-Fidelity World Modeling in CARLA for Autonomous Vehicles. arXiv preprint arXiv:2411.10171.*

[6] *Lelkó, A., Németh, B., Fényes, D., & Gáspár, P. (2023). Integration of robust control with reinforcement learning for safe autonomous vehicle motion. IFAC-PapersOnLine, 56(2), 1101-1106.*

[7] *Du, G., Zou, Y., Zhang, X., & Zhao, K. (2024, June). Motion Control of Autonomous Vehicle with Domain-Centralized Electronic and Electrical Architecture based on Predictive Reinforcement Learning Control Method. In 2024 IEEE Intelligent Vehicles Symposium (IV) (pp. 1409-1416). IEEE.*

DOI: 10.54254/2755-2721/2025.TJ23323

[8] *Pérez-Gil, Ó., Barea, R., López-Guillén, E., Bergasa, L. M., Gómez-Huélamo, C., Gutiérrez, R., & Díaz-Díaz, A. (2022). Deep reinforcement learning based control for Autonomous Vehicles in CARLA. Multimedia Tools and Applications, 81(3), 3553-3576.*

[9] *Gao, H., Chen, S., Jiang, B., Liao, B., Shi, Y., Guo, X., ... & Wang, X. (2025). Rad: Training an end-to-end driving policy via large-scale 3dgs-based reinforcement learning. arXiv preprint arXiv:2502.13144.*

[10] *Gao, D., Cai, S., Zhou, H., Wang, H., Soltani, I., & Zhang, J. (2024). Cardreamer: Open-source learning platform for world model based autonomous driving. IEEE Internet of Things Journal.*

[11] *Li, Y., Jiang, M., Zhang, S., Yuan, W., Wang, C., & Yang, M. (2024). End-to-end Driving in High-Interaction Traffic Scenarios with Reinforcement Learning. arXiv preprint arXiv:2410.02253.*

[12] *Huang, Z., Sun, S., Zhao, J., & Mao, L. (2023). Multi-modal policy fusion for end-to-end autonomous driving. Information Fusion, 98, 101834.*