Temporal-Aware Reinforcement Learning for Robust Autonomous Driving Decision-Making

Xin Xing, Jannes Bikker, Sebastian Ohl

Suggested citation:

Xing, Xin, Jannes Bikker, and Sebastian Ohl. 2025. "Temporal-Aware Reinforcement Learning for Robust Autonomous Driving Decision-Making." In *KI-Forum 2025 : KI in Forschung und Lehre an Hochschulen*, edited by Hanno Homann, Cedric Rohbani, and Jens Christian Will, 117–20. Hannover: HsH Applied Academics. https://doi.org/10.25968/opus-3792.

Abstract

Autonomous driving systems require robust and precise decision-making, particularly for safety-critical functions such as Automatic Emergency Braking (AEB) and for comfortoriented features like Adaptive Cruise Control (ACC). Although conventional Reinforcement Learning (RL) methods with Multilayer Perceptrons (MLPs) have shown potential, they lack the temporal awareness needed for complex, dynamic driving scenarios. In this paper, we propose an enhanced decisionmaking architecture that integrates Proximal Policy Optimization (PPO) with sequential models—specifically Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks. Simulation-based evaluations demonstrate that our method improves AEB responsiveness and ACC stability while reducing false activations.

Temporal-Aware Reinforcement Learning for Robust Autonomous Driving Decision-Making

Xin Xing, Jannes Bikker, Sebastian Ohl
Ostfalia University of Applied Sciences, Faculty of Electrical Engineering and Information Technology

***Abstract*—Autonomous driving systems require robust and precise decision-making, particularly for safety-critical functions such as Automatic Emergency Braking (AEB) and for comfort-oriented features like Adaptive Cruise Control (ACC). Although conventional Reinforcement Learning (RL) methods with Multilayer Perceptrons (MLPs) have shown potential, they lack the temporal awareness needed for complex, dynamic driving scenarios. In this paper, we propose an enhanced decision-making architecture that integrates Proximal Policy Optimization (PPO) with sequential models—specifically Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks. Simulation-based evaluations demonstrate that our method improves AEB responsiveness and ACC stability while reducing false activations.**

***Index Terms*—Autonomous Driving, Reinforcement Learning, Sequential Modeling**

I. Introduction

In recent years, autonomous driving technologies have advanced rapidly, and Advanced Driver Assistance Systems (ADAS)—including Adaptive Cruise Control (ACC) [1, p. 24] and Automatic Emergency Braking (AEB) [1, p. 666]—have become increasingly sophisticated. Decision-making for autonomous vehicles is one of the most complex challenges in contemporary artificial intelligence. Decision modules must process continuous streams of sensor data in real time, interpret highly dynamic environments, and execute appropriate actions while accounting for uncertainties in traffic conditions. Reinforcement Learning (RL) has proven to be a highly effective solution for a range of tasks. This framework enables systems to learn optimal policies through interaction with the environment, with the objective of maximising long-term returns, such as safety and efficiency [2]. However, traditional RL approaches based on Multilayer Perceptrons (MLPs) still face significant challenges, particularly in minimizing false activations of safety-critical functions, like AEB [3], [4].

Although both ACC and AEB are essential components of autonomous driving systems, their objectives and operational requirements differ substantially. ACC primarily aims to maintain safe following distances and comfortable driving speeds, requiring smooth and predictable control. Minor errors in ACC typically do not lead to significant safety issues. In contrast, AEB is a critical safety feature that demands binary, high-consequence decisions with extremely precise activation timing. Any false activation may result in abrupt braking, causing discomfort or even hazardous situations. These distinctions explain why our initial MLP-based methods achieved better results with ACC than with AEB. While ACC's continuous nature and tolerance for minor errors align well with standard RL techniques, AEB's high-risk, low-error-tolerance characteristics call for more sophisticated temporal modeling to distinguish genuine emergencies from transient anomalies.

In previous studies [3], [4], we employed Proximal Policy Optimization (PPO)—a state-of-the-art RL algorithm—combined with an MLP architecture to train separate strategies for ACC and AEB behaviors. While the PPO-MLP framework achieved satisfactory performance in maintaining safe following distances for ACC, it revealed shortcomings in AEB.

Driving is inherently a sequential decision-making process in which current actions influence future states and rewards. Such temporal dependencies imply that optimal decision-making requires not only an accurate understanding of the current environment but also memory of past states and the ability to anticipate future scenarios. Although MLP-based methods are effective for many tasks, they lack explicit mechanisms to capture these temporal relationships. This observation led us to hypothesize that integrating memory mechanisms—specifically Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) networks—can significantly enhance decision quality by preserving and leveraging relevant temporal context [5].

Integrating sequence-modeling techniques such as LSTM or GRU into the PPO framework represents a promising direction to address these challenges. These architectures offer several potential advantages [6]:

- By retaining memory of past states, systems can better differentiate between genuinely hazardous conditions and transient anomalies.
- Sequence models can learn predictive patterns of future states, enabling more informed judgments about the necessity of emergency braking.
- Memory mechanisms can help filter out transient sensor noise by assessing consistency over time.

This integration aims to provide the sophisticated temporal processing capabilities necessary for reliable AEB decision-making, complemented by the stable and efficient learning abilities offered by PPO.

The remainder of this paper is organized as follows. Section II reviews related work. Section III describes the proposed PPO-based sequential modeling framework. Section IV presents the experimental setup and results. Section V concludes the paper and outlines future directions.

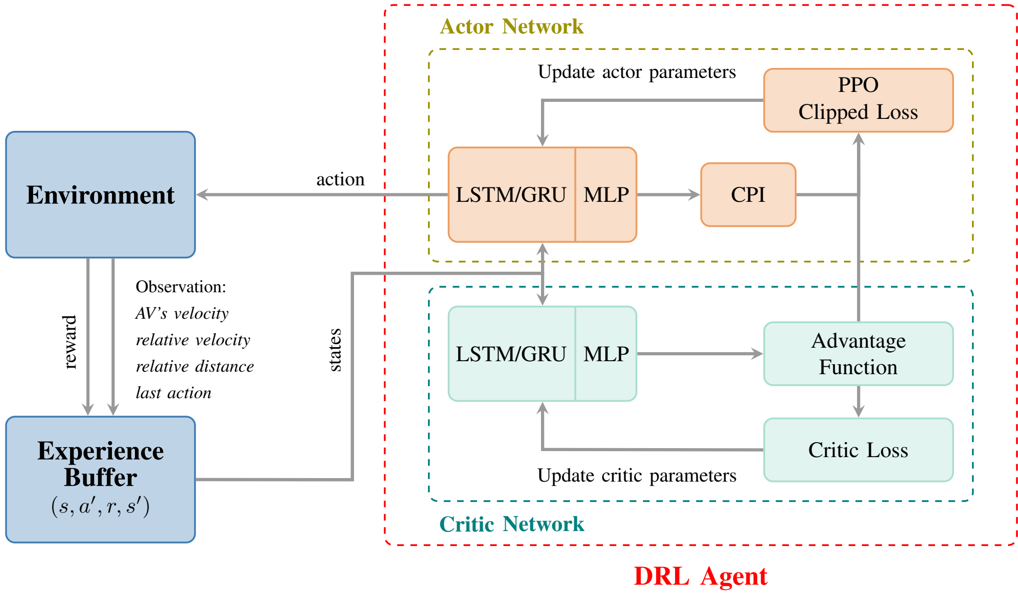

Figure 1. Reinforcement Learning PPO-Clip with Sequential Modeling Integration

II. Related Work

Deep Reinforcement Learning (DRL) has become a pivotal methodology for autonomous driving, particularly for high-level decision-making [2], [7]. Among various DRL algorithms, PPO stands out as a state-of-the-art on-policy approach for decision-making [5], [8].

Proximal Policy Optimization has become a cornerstone on-policy DRL algorithm for autonomous driving due to its balance of simplicity and performance. By clipping policy updates within a trust region, PPO mitigates the instability and divergence issues often seen in Deep Deterministic Policy Gradient (DDPG), while retaining first-order optimization and high sample efficiency. This design leads to faster convergence and reliable continuous control—PPO agents have, for example, been trained in purely synthetic environments to output steering and throttle commands and then deployed successfully on real vehicles [8], [9].

LSTM and GRU networks excel at encoding sequential information and preserving long-term dependencies, thereby overcoming vanishing- or exploding-gradient problems inherent to vanilla RNNs. In autonomous driving, they underpin end-to-end systems that must fuse streams of sensor data over time. Frameworks such as Action-Specific Deep Recurrent Q-Networks (ADRQN) integrate past observations and actions via LSTM layers to solve Partially Observable Markov Decision Processes (POMDPs), enabling robust decision-making under uncertainty [5], [7].

By combining PPO's stable, data-efficient policy updates with LSTM/GRU's temporal memory, modern DRL agents can negotiate multi-agent interactions, perform complex maneuvers, and navigate dynamic traffic scenarios safely and efficiently in continuously changing, partially observable environments [9].

III. Methodology

Our approach is centered around a hybrid architecture that combines PPO-based policy optimization with LSTM/GRU-based sequence modeling. The overall framework is illustrated in Figure 1, which shows how the sequence module is integrated into the standard PPO-Clip model. This design enables the system to retain past state information and predict future scenarios, thereby enhancing the decision-making accuracy of critical actions such as AEB activation.

A. Proposed Framework: PPO with Sequential Modeling

The framework operates in three phases:

a) State Representation: The raw sensor inputs (e.g., relative velocity, distance to obstacles) at each time step t are preprocessed into a state vector s_t. To address partial observability, this vector is integrated with historical context through implicit temporal modeling using an LSTM or GRU.

b) Policy Network with Sequence Module: The actor network (policy π) and the critic network (value function V) are enhanced with an LSTM or GRU layer. As shown in Figure 2, the actor network uses the recurrent unit to encode temporal dependencies by transforming the state s_t into a hidden state h_t. This allows the system to learn from sequences of past observations and actions, as emphasized in Section II (e.g., the ADRQN framework for POMDPs). The output of the LSTM—such as the hidden state at the last time step or an aggregated representation across all time steps—is then fed into a MLP. Through fully connected layers, nonlinear transformations are applied to complete the classification task.

c) PPO-Clip Optimization: The policy is optimized using the clipped objective function from PPO to ensure stable

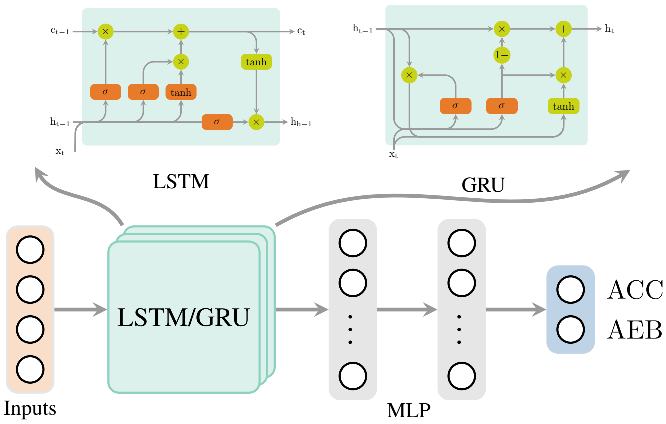

Figure 2. Actor Network for PPO-Clip

updates:

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

where $r_t(\theta)$ is the probability ratio, $\hat{A}_t$ is the advantage estimate, and ϵ (e.g., 0.2) defines the clipping range to prevent large policy deviations. This approach balances exploration and exploitation while maintaining convergence efficiency.

B. Key Components

a) Sequence Model Selection: We empirically evaluate the use of either LSTM or GRU. GRU is favored for its computational efficiency and simpler gating mechanism, while LSTM is used in scenarios requiring longer memory spans. The recurrent layers are initialized with zero hidden states and trained to output hidden representations h_t, which are fed into fully connected layers for action selection.

b) Action Space and Reward: The action space includes continuous control for ACC (e.g., speed control based on Intelligent Driver Model (IDM) [1, p. 148]) and a discrete binary output for AEB (e.g., brake activation). The reward function penalizes false activations (e.g., unnecessary braking) and encourages safety (e.g., high rewards for collision avoidance). This aligns with the distinctions outlined in Section I: ACC tolerates minor errors, whereas AEB requires high precision.

c) Training Protocol: The model is trained in a simulated driving environment (e.g., using Webots [3]) with scenarios reflecting real-world traffic dynamics. We adopt a curriculum learning strategy, starting with simple ACC tasks and gradually introducing AEB-critical scenarios (e.g., sudden obstacles or emergency braking during following). Training is performed using rolling batches of sequential data, with network parameters updated via the Adam optimizer.

C. Advantages and Innovations

Integrating sequence modeling into PPO, as discussed in Section I, brings three key advantages that significantly enhance decision-making in autonomous driving systems. First, by retaining temporal context, the system gains the ability to distinguish genuine threats—such as sudden obstacles—from benign fluctuations, thereby reducing false AEB activations. Moreover, the LSTM/GRU layers enable predictive capabilities by learning temporal patterns, allowing the system to make proactive decisions that minimize the need for emergency interventions. Additionally, sequential consistency checks across time-series sensor data (e.g., changes in velocity or distance) enhance noise robustness by identifying and filtering out transient anomalies that violate physical or environmental continuity, such as single-frame radar misdetections or brief sensor interference. By explicitly modeling temporal dependencies, this approach surpasses earlier MLP-based reinforcement learning methods [3], leading to more reliable performance in both ACC and AEB functionalities under dynamic driving conditions.

IV. Experiments and Results

In experiments, we design two simplified scenarios to evaluate the performance of our reinforcement learning model:

- **Adaptive Cruise Control**: This scenario involves a straightforward car-following task. The autonomous vehicle was required to maintain a safe and constant distance behind a leading vehicle using ACC based on the IDM. The goal was to assess the model's capability in stable driving conditions, ensuring smooth acceleration and deceleration while closely following the leading vehicle.
- **Automatic Emergency Braking**: Building upon Scenario 1, Scenario 2 introduces a dynamic and unpredictable event, where obstacles appeared randomly ahead of the autonomous vehicle during car-following. The appearance of these obstacles necessitated immediate activation of AEB, compelling the model to swiftly transition from ACC to AEB mode, thus testing the robustness and responsiveness of the trained agent.

Hyperparameters were systematically optimized using the **Optuna** framework [10]. To guarantee the comparability and fairness of our experimental results, the optimization procedure and subsequent training utilized an identical environmental setup and reward function. This ensured the hyperparameters selected by Optuna were consistently beneficial across both scenarios.

Figure 3 reveals that the **GRU+MLP**-enhanced policy (**GMLP**) converges after approximately 262 episodes, attaining the highest mean return of 2515.6. By contrast, the plain MLP baseline stabilizes far earlier—around 48 episodes—but stalls at a negative reward of -524.9. Meanwhile, the **LSTM + MLP** (**LMLP**) variant needs 491 episodes to converge and ultimately levels off at 2333.1. Collectively, these results underline both the superior sample efficiency and the higher asymptotic performance of the GRU-based architecture.

The performance metrics presented in Table 1 reflect the comparative effectiveness of the MLP, GMLP, and LMLP reinforcement learning architectures, with the reward values being normalized via **Min-Max** scaling to ensure consistency

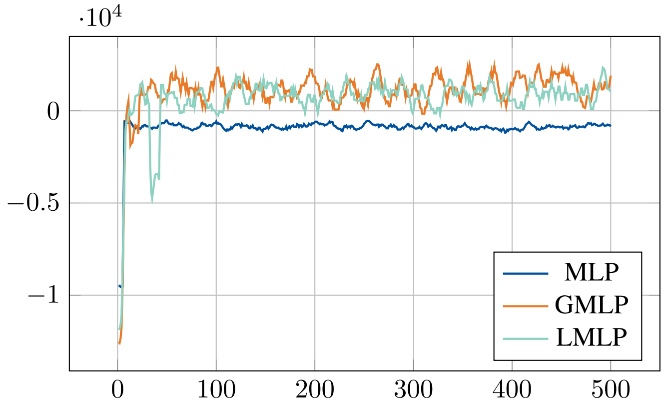

Figure 3. Smoothed rewards (Y-axis) over episodes (X-axis) for MLP, GMLP, and LMLP

Table 1. Performance Summary for MLP, GMLP, and LMLP

Metric	**MLP**	**GMLP**	**LMLP**
Mean (μ)	0.9529 ± 0.006	0.9006 ± 0.008	0.8788 ± 0.009
Std (σ)	0.0986	0.0995	0.1041
Success (%)	100	100	100
AEB Act. (%)	0	51.8	51.5
AEB False Act. (%)	-	0.5	0

and comparability across models. Among the evaluated models, the MLP architecture exhibited the highest normalized mean reward ($\mu = 0.9529 \pm 0.006$) along with the lowest standard deviation ($\sigma = 0.0986$), indicating strong stability and optimal performance in scenarios predominantly governed by ACC. However, the lack of AEB activation highlights its limited responsiveness in scenarios involving sudden obstacles. In contrast, the recurrent architectures (GMLP and LMLP) showed slightly lower normalized mean rewards (0.9006 and 0.8788, respectively) with marginally higher variability, yet notably exhibited substantial AEB activation rates (51.8% and 51.5%), demonstrating their heightened sensitivity and effectiveness in dynamic driving conditions. Furthermore, the GMLP model's minor false activation rate (0.5%) underlines the challenge of balancing responsiveness with precision, whereas the LMLP model achieved effective obstacle detection without any erroneous activations. These normalized reward insights emphasize the significance of choosing model architectures according to the complexity and unpredictability inherent in real-world driving scenarios.

V. Conclusion

This study presents a sequential modeling framework for enhancing decision-making in autonomous driving, focusing on the integration of LSTM and GRU networks with the PPO reinforcement learning algorithm. The proposed approach addresses the limitations of traditional MLP-based architectures, particularly in detecting and responding effectively to dynamic driving events. Experimental results show that the GMLP or LMLP-enhanced policy not only improves AEB activation accuracy but also maintains ACC performance under varying traffic conditions. These findings confirm the importance of temporal dependencies in autonomous systems and highlight the potential of recurrent neural networks in building safer and more reliable decision-making modules for real-world deployment.

References

[1] A. Eskandarian, Ed., *Handbook of Intelligent Vehicles*. Springer London, 2012, ISBN: 978-0-85729-085-4. DOI: 10.1007/978-0-85729-085-4.

[2] J. Zhao, Y. Wu, R. Deng, S. Xu, J. Gao, and A. Burke, "A survey of autonomous driving from a deep learning perspective," *ACM Comput. Surv.*, vol. 57, no. 10, 2025, ISSN: 0360-0300. DOI: 10.1145/3729420. [Online]. Available: https://doi.org/10.1145/3729420.

[3] X. Xing and S. Ohl, "Application of a maneuver-based decision making approach for an autonomous system using a learning approach," in *ADVCOMP 2024: The Eighteenth International Conference on Advanced Engineering Computing and Applications in Sciences*, Sept. 29–Oct. 3, 2024, Venice, Italy, Sep. 2024, pp. 11–16, ISBN: 978-1-68558-184-8.

[4] X. Xing, M. Molaei, and S. Ohl, "Maneuver-based decision making in autonomous driving via reinforcement learning and simulation-to-real transfer," *International Journal On Advances in Intelligent Systems*, vol. 18, no. 1-2, pp. 46–56, 2025, ISSN: 1942-2679.

[5] I. Smirnov and S. Gu, *Rlbenchnet: The right network for the right reinforcement learning task*, 2025. arXiv: 2505.15040 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2505.15040.

[6] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, *Empirical evaluation of gated recurrent neural networks on sequence modeling*, 2014. arXiv: 1412.3555 [cs.NE]. [Online]. Available: https://arxiv.org/abs/1412.3555.

[7] Y. Chen, C. Ji, Y. Cai, T. Yan, and B. Su, *Deep reinforcement learning in autonomous car path planning and control: A survey*, 2024. arXiv: 2404.00340 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2404.00340.

[8] B. Uppuluri, A. Patel, N. Mehta, S. Kamath, and P. Chakraborty, *Curla: Curriculum learning based deep reinforcement learning for autonomous driving*, 2025. arXiv: 2501.04982 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2501.04982.

[9] J. Hu *et al.*, "A survey of decision-making and planning methods for self-driving vehicles," *Frontiers in Neurorobotics*, vol. Volume 19, 2025, ISSN: 1662-5218. DOI: 10.3389/fnbot.2025.1451923.

[10] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, *Optuna: A next-generation hyperparameter optimization framework*, 2019. arXiv: 1907.10902 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1907.10902.