Hybrid PPO–DQN for Multi-Objective Adaptive Cruise Control in Eco-Driving: Reward Shaping Toward Safety and Sustainability (Student Abstract)

Tae Hoon Lee, Joongheon Kim

Korea University
Seoul 02841, Republic of Korea
joongheon@korea.ac.kr

Abstract

In adaptive cruise control (ACC), balancing safety, comfort, and sustainability still remains challenging. Accordingly, we propose a hybrid reinforcement learning framework combining proximal policy optimization (PPO) and deep Q-network (DQN) with a multi-objective reward for autonomous carbon-neutral eco-driving. Experimental results revealed the contrasts between eco and non-eco modes, underscoring how reward design shapes driving behaviors.

Introduction

Adaptive cruise control (ACC) is a core function of advanced driver-assistance systems (ADAS), responsible for jointly balancing safety, comfort, and environmental efficiency (Yun et al. 2022; Shin and Kim 2019). Yet improvements in safety and comfort often come at the expense of fuel economy and CO_2 reduction, creating inherent tensions among these objectives.

Among various approaches for carbon-aware ACC operations, prior reinforcement learning (RL) approaches for ACC have typically pursued narrow goals, emphasizing either safety metrics—such as headway regulation and collision avoidance—or eco-driving incentives like emission reduction. Such single-focus strategies fail to capture the coupled dynamics of real driving, where safety, comfort, and sustainability interact. As a result, they often lack robustness and generalization across diverse traffic regimes, particularly when comparing eco and non-eco contexts.

To address these issues, we propose a novel hybrid RL framework that combines proximal policy optimization (PPO) and deep Q-network (DQN) and its performance has been evaluated in an ACC simulator, i.e., simulation of urban mobility (SUMO). The part of PPO stabilizes policy-gradient updates, while the part of DQN enhances exploration through off-policy value backups, jointly improving training resilience. The reward of our hybrid RL is designed with a multi-objective structure, penalizing unsafe behaviors (i.e., gap deviation, relative velocity, acceleration, jerk, shield use) while incorporating environmental terms (i.e., CO_2 emissions, throttle, aerodynamic drag) to encourage eco-driving.

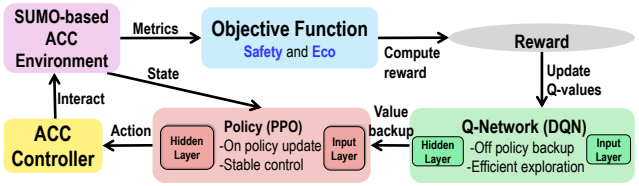


Figure 1: Hybrid PPO–DQN Framework for Multi-Objective ACC Operations.

Although the framework was intended to prioritize eco-driving, comparative evaluation revealed distinct behavioral outcomes: the non-eco agent favored performance and comfort, while the eco agent emphasized sustainability and stability. This contrast highlights how reward shaping can alter behavioral priorities in unexpected ways, offering insights into the design of sustainable autonomous driving systems.

Hybrid Reinforcement Learning for ACC

Unlike conventional RL methods that rely solely on policy gradients or value-based updates, our framework integrates PPO with DQN, as illustrated in Fig. 1.

Hybrid PPO–DQN Framework. The part of PPO constrains policy updates for stability, while the part of DQN provides off-policy value backups for efficient exploration. Combined, they deliver robust short-term control and reliable long-term value estimation, mitigating the weaknesses of each method alone. Accordingly, the overall training objective can be designed as,

$$L(\theta, \phi) = L_{\text{PPO}}(\theta) + \lambda_Q L_{\text{DQN}}(\phi), \tag{1}$$

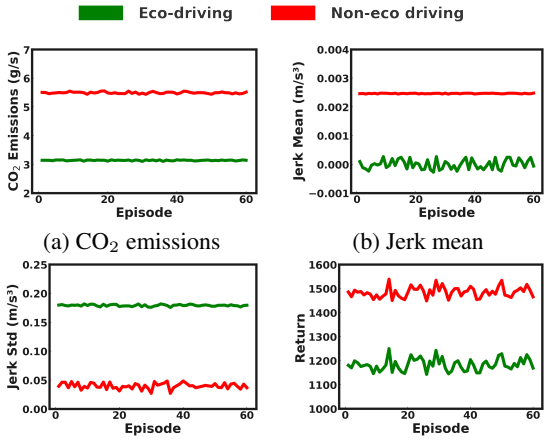
where L_{PPO} is the clipped surrogate loss, L_{DQN} the temporal-difference loss, θ denotes the parameters of PPO, ϕ denotes the parameters of DQN, and λ_Q balances the two components, respectively.

Reward Design. The reward integrates safety and environmental objectives. Safety terms cover headway deviation, relative velocity, acceleration, jerk, and shield use, while environmental terms include CO_2 emissions, throttle, and aerodynamic drag. This design balances safety with sustainability. Accordingly, the per-step reward can be designed as,

$$r_t = -\left(\sum\nolimits_{k \in S_k} w_k L_{k,t} + \sum\nolimits_{m \in S_m} w_m C_{m,t} \right), \tag{2}$$

Metric	Eco-driving	Non-eco driving
Mean Return	1183.4 ± 24.8	1485.0 ± 21.4
CO_2 Emissions (g/s)	3.14 ± 0.01	5.49 ± 0.03
Jerk mean (m/s^3)	$(0.0 \pm 1.6) \times 10^{-4}$	$(2.47 \pm 0.01) \times 10^{-3}$
Jerk std (m/s^3)	0.179 ± 0.0015	0.0398 ± 0.0050

Table 1: Eco vs. Non-eco Driving Performance.



(a) CO_2 emissions (b) Jerk mean

(c) Jerk standard deviation (d) Return

Figure 2: Eco vs. non-eco driving. Eco-driving reduces CO_2 while keeping jerk stable, while non-eco driving yields smoother control and higher returns.

where $S_k \triangleq \{\text{gap}, \text{dv}, \text{acc}, \text{jerk}, \text{shield}\}$ and $S_m \triangleq \{\text{CO2}, \text{thr}, \text{drag}\}$. In addition, $L_{k,t}$ and $C_{m,t}$ denote safety and environmental terms; and note that (2) captures the trade-off between stability and efficiency. This trade-off arises because stability-oriented terms (gap, relative velocity, acceleration, jerk, and shield) penalize aggressive maneuvers to ensure safety and smoothness, while efficiency-oriented terms (CO_2, throttle, and drag) encourage minimizing energy consumption and emissions. Emphasizing stability can lead to conservative driving that sacrifices efficiency, whereas emphasizing efficiency is able to reduce responsiveness and compromise stability, therefore, their coexistence in the reward inherently encodes a competing relationship.

Performance Evaluation

We evaluate eco- and non-eco driving in the SUMO-based ACC environment using representative metrics, i.e., mean return, CO_2 emissions, jerk (driving comfort), and in-track/in-norm ratios (safety and stability). Other measures showed largely similar trends across modes.

Evaluation Metrics. We evaluate policies with complementary metrics capturing safety, comfort, and environmental efficiency, as follows: (i) *in-track ratio* (proportion within strict safety bounds (gap and relative velocity)), (ii) *in-norm ratio* (proportion within a relaxed tolerance band), (iii) *mean return* (episodic return), (iv) *CO_2 emissions* (average grams/sec), (v) *jerk mean* (average acceleration change rate (stability)), and (vi) *jerk std* (variability of acceleration change). These metrics capture trade-offs across eco and non-eco modes. The in-track ratio (hard) enforces stringent safety margins, recognizing only steps within tight bounds of headway and relative velocity. In contrast, the in-norm ratio (soft) relies on a probabilistic band similarity with an intentionally permissive threshold, making it easier to satisfy. This dual-level design is deliberate: the hard metric reflects strict, safety-critical precision in vehicle dynamics, while the soft metric captures broader stability and training robustness by accommodating diverse driving patterns. While the soft ratio often yields high values across modes, this is intentional—it is meant to represent overall stability rather than strict safety. Together, the two metrics are complementary by design, since neither alone can fully characterize agent performance.

Eco and Non-eco Setup. To examine these trade-offs, we consider two reward designs: *eco-driving*, which incorporates CO_2, throttle, and drag penalties in addition to safety terms, and *non-eco driving*, which excludes environmental costs. Both modes are trained under identical SUMO conditions, allowing controlled evaluation of how reward design shapes learned behavior.

Comparison of Eco and Non-eco Driving. Both agents achieve nearly identical safety scores, with in-track ratios of 18.5–18.8% and in-norm ratios close to 99%. Analysis of the outcomes revealed a clear divergence: the non-eco agent achieved higher returns and smoother control with lower jerk variability, while the eco agent reduced CO_2 emissions by approximately 43% and maintained stable jerk regulation. Thus, although the framework was designed to encourage eco-driving, the comparative evaluation uncovered contrasting behavioral priorities: eco-driving emphasized sustainability and stability, whereas non-eco driving favored performance and ride comfort.

Conclusion

We proposed a novel hybrid RL framework for adaptive cruise control using a PPO–DQN hybrid with a multi-objective reward in eco-driving. Experiment results showed that while eco-driving reduced CO_2 emissions and stabilized jerk, the non-eco mode yielded higher returns and smoother control. This contrast highlights how reward design can shift behavioral priorities.

Acknowledgements

This research was supported by MSIT (Ministry of Science and ICT), Korea, under ITRC (Information Technology Research Center) (IITP-2024-RS-2024-00436887). The corresponding author is Joongheon Kim and his postal address is as follows: Engineering Building #214, 145 Anam-ro, Seoul 02841, Korea (+82-2-3290-3223, joongheon@korea.ac.kr).

References

Shin, M.; and Kim, J. 2019. Randomized Adversarial Imitation Learning for Autonomous Driving. In *Proceedings of IJCAI*, 4590–4596. Macau, China.

Yun, W. J.; Shin, M.; Jung, S.; Kwon, S.; and Kim, J. 2022. Parallelized and Randomized Adversarial Imitation Learning for Safety-Critical Self-Driving Vehicles. *Journal of Communications and Networks*, 24(6): 710–721.