Optimizing Roundabout Management via Deep Reinforcement Learning with Safety and Comfort Constraints

Ali Nadar, Jérôme Härri
EURECOM, 450 route des Chappes, 06904 Sophia-Antipolis, France
{ali.nadar, jerome.haerri}@eurecom.fr

***Abstract*—This paper presents a deep reinforcement learning (DRL) framework to optimize autonomous vehicle maneuver during roundabout approaches, with a focus on safety, efficiency, and passenger comfort. The proposed method incorporates a logistic regression-based Roundabout Exit Probability (REP) model to estimate the likelihood that inbound vehicles will exit the roundabout, as well as a regression-based Time-To-Collision (TTC) predictor to model the ego vehicle's controlled maneuver while maintaining comfort constraints. These predictive models are integrated into a Proximal Policy Optimization (PPO) framework, enhanced with a curriculum learning strategy to gradually shape the agent's behavior toward balanced, human-like decision-making. The reward function is designed to penalize unsafe or abrupt actions and encourage smooth, efficient maneuvering. Experimental results in the CARLA simulator demonstrate the effectiveness of the proposed strategy in achieving robust, comfort-aware navigation in roundabout scenarios.**

***Index Terms*—DRL, Curriculum Learning, CARLA, Roundabout, AI-driven autonomous systems, PPO, Logistic-Regression, Passenger-Comfort**

I. Introduction

Navigating roundabouts remains one of the most complex and safety-critical challenges for autonomous vehicles, especially in dynamic urban and semi-urban settings. Unlike traditional intersections, roundabouts require continuous negotiation of entry and exit maneuvers, demanding an understanding of lane assignments, right-of-way rules, surrounding vehicle intentions, and rapidly changing traffic dynamics. Achieving this requires a holistic perception of the driving context and the ability to make real-time, safe, and comfortable decisions.

To address these challenges, recent research has explored various AI and machine learning techniques for modeling roundabout maneuvers. Methods like rule-based systems, tactical planners, and learning-based models emulate human decision-making. Among these, reinforcement learning (RL) has proven effective for enabling adaptive, data-driven decisions in uncertain traffic environments. In our study, we use CARLA[1], an open-source autonomous driving simulator offering a high-fidelity environment for simulating traffic scenarios, ideal for training behavior models in safety-critical tasks such as roundabout navigation.

We propose a two-stage AI-driven framework for roundabout entry, balancing safety, efficiency, and passenger comfort. The first stage utilizes a supervised Logistic Regression model, the Roundabout Exit Probability (REP) model, which predicts whether an inbound vehicle will exit before reaching the ego vehicle's entry leg. By filtering out non-threatening vehicles early, REP reduces unnecessary conservatism in decision-making and allows the ego agent to exploit safe entry gaps more efficiently.

The second stage employs Proximal Policy Optimization (PPO), a state-of-the-art reinforcement learning algorithm, to train a policy that decides whether the ego vehicle should yield or proceed. The RL agent learns to reason under uncertainty, using observations such as ego speed, distance to the roundabout, acceleration, and safety indicators based on Time-To-Collision (TTC). Through repeated interaction with the environment, the agent refines its strategy, balancing assertiveness and caution to ensure smooth and safe roundabout entry.

Our system also integrates Vehicle-to-Vehicle (V2V) communication using Cooperative Awareness Messages (CAMs), as standardized by ETSI (EN 302 637-2) [1]. These messages broadcast key state information—position, speed, acceleration, and heading—among nearby vehicles, enhancing the ego vehicle's ability to assess inbound vehicle dynamics and make contextually appropriate entry decisions, particularly in high-density traffic scenarios.

In this paper, we present a complete methodology for enabling AI-based control of an autonomous vehicle during the roundabout merging phase. Our solution is composed of three interconnected modules: (i) a PID-based comfort-aware TTC estimation module that models the ego vehicle's acceleration profile to generate a realistic and controllable TTC measure; (ii) a supervised REP module that anticipates inbound vehicle behavior, filtering out non-threatening vehicles; (iii) a PPO-based reinforcement learning agent that synthesizes these inputs to learn safe, efficient, and comfort-aware driving policies in real time.

The remainder of this paper is organized as follows: Section II reviews related work on roundabout navigation using reinforcement learning. Section III introduces our methodology for achieving a safe and comfort-aware roundabout merging. Section IV describes the preparation phase on designing and training our regression-based predictors. Section V presents our deep reinforcement learning model for roundabout decision-making maneuver. Section VI discusses the

[1] https://carla.org/

simulation setup and evaluation results. Finally, Section VII concludes the paper and outlines directions for future research. To promote reproducibility and facilitate further research, we have released the complete code-base, including model training, testing routines, and simulation environment setup as open-source, publicly available here[2].

II. Related Work

Navigating roundabouts presents significant challenges for autonomous vehicles due to dynamic multi-agent interactions and the need for real-time decision-making that balances safety and passenger comfort. Recent studies have explored intent prediction, comfort-aware control, and reinforcement learning strategies to address these complexities. In [2], *Deveaux et al.* proposed a knowledge networking approach for AI-driven roundabout risk assessment, introducing the Roundabout Exit Probability (REP) model to predict inbound vehicle exits. We adopt the REP concept to dynamically weight inbound vehicles in the observation space, filtering those likely to exit and reducing unnecessary Time-To-Collision (TTC) computations and waiting time-loss. Comfort-aware control was studied in [3] by *Moradi*, who introduced a PID controller with a transfer function to achieve smooth acceleration and deceleration, minimizing jerk and enhancing passenger comfort during roundabout navigation. Building on this, we integrate a comfort-aware TTC predictor that estimates safe and smooth collision timing for the ego vehicle. For roundabout navigation, *Cuenca et al.* applied Q-learning in [4] to train an agent using trial-and-error interactions within the CARLA simulator. While their work focuses purely on learning maneuvering behavior, our methodology leverages pre-trained REP and TTC models to enrich the agent's situational awareness and improve policy robustness. Intersection navigation via deep RL was tackled by *Elallid et al.* in [5], where the authors addressed uncertainty and dynamic priority management using a DRL policy. Although their work focuses on intersections, it informs our handling of multi-agent interactions in roundabouts, where decision complexity is even greater. In [6], *Gutiérrez-Moreno et al.* trained an agent via PPO to manage intersection types such as traffic lights and stop signs, focusing mainly on collision avoidance. Our work builds upon this by tackling the added complexity of roundabouts and introducing a comfort dimension through smooth, human-like driving strategies based on PID-guided TTC estimation. Recognizing the importance of structured learning, *Anzalone et al.* in [7] proposed Reinforced Curriculum Learning in CARLA to progressively train autonomous agents through increasing task difficulty. Inspired by this, we incorporate a curriculum learning strategy in our PPO-based framework, initially focusing on efficiency and safety before gradually introducing passenger comfort constraints, leading to smoother and more robust roundabout entry behavior. In [8] *Yuan et al.* investigated DRL algorithms—DDPG, PPO, and TRPO—for automated driving through roundabouts. The reward function incorporates safety, efficiency, comfort, and energy consumption. TRPO outperforms others in safety and efficiency, while PPO excels in comfort. The study also demonstrates the adaptability of the TRPO model to different driving scenarios like highway driving and merging. In [9] *Capasso et al.* introduced a maneuver planning module using a novel Delayed A3C (D-A3C) algorithm for negotiating entry into busy roundabouts. The system allows agents to exhibit varying levels of aggressiveness, emulating different driving styles, which is particularly useful in managing congested scenarios.

III. Methodology

Our proposed framework for optimizing autonomous roundabout merging maneuver is structured into two sequential phases: offline pre-training of predictive models, and online decision-making via reinforcement learning. An overview of the framework is illustrated in Fig.1.

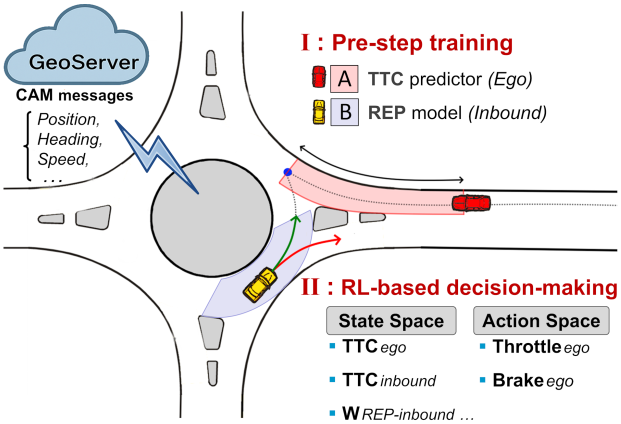


Fig. 1: Overview

In the pre-training phase, we develop two regression-based predictors to enhance the ego vehicle's situational awareness.

First, the Roundabout Exit Probability (REP) model estimates the likelihood, as a continuous probability within [0,1], that an inbound vehicle (zone B) will exit the roundabout before reaching the ego vehicle's trajectory. This probability later serves as a dynamic feature within the reinforcement learning agent's state space, effectively filtering non-threatening vehicles from decision consideration. Second, we train a comfort-aware Time-To-Collision (TTC) predictor for the ego vehicle (zone A). This model outputs a target collision timing that accounts for passenger comfort by adhering to constraints on longitudinal/lateral acceleration and jerk, enabling smoother and more natural driving behavior. To ensure the reliability and quality of these predictive models, we conduct an independent evaluation phase prior to integrating them into the reinforcement learning pipeline. This evaluation validates the models' predictive performance and their suitability for real-time decision support.

In the second phase, we integrate the REP and TTC models within a Proximal Policy Optimization (PPO)-based control framework. During online operation, the ego vehicle processes

[2] https://gitlab.eurecom.fr/cats/carla/comfort-rl-roundabout

continuous Cooperative Awareness Messages (CAMs) from nearby vehicles, using the predicted TTC to guide its basic ”yield” or ”proceed” decisions. A curriculum learning strategy progressively introduces comfort considerations into the agent’s learning objective, ensuring robust, safe, and smooth roundabout entries. The following sections detail each phase, beginning with the offline training and evaluation of the REP and TTC predictive models.

IV. Pre-training phase: Regression-based Predictors

A. Roundabout Exit Probability (REP) model

To anticipate the behavior of circulating vehicles and reduce unnecessary conservatism in decision-making, we first train a supervised logistic regression model, referred to as the Roundabout Exit Probability (REP) model [2]. This model estimates the likelihood that an inbound vehicle will exit the roundabout before reaching the ego vehicle’s merging point.

The REP model is trained offline using annotated vehicle trajectories extracted from the CARLA simulator. Each training sample includes features derived from Cooperative Awareness Messages (CAMs), standardized under ETSI EN 302 637-2 [1]. The input features include vehicle position, heading, speed, and acceleration.

At inference time, the REP output is used as a dynamic attention weight within the ego vehicle’s observation space. This mechanism effectively down-weights vehicles that are likely to exit and therefore pose less collision risk, allowing the agent to concentrate computational effort on truly relevant interactions. As a result, the agent achieves faster and more efficient decision-making during roundabout negotiation.

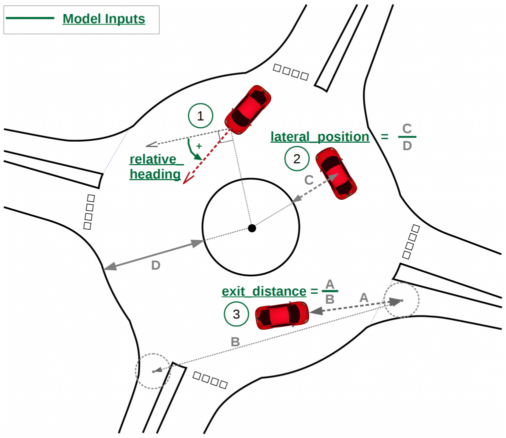


Fig. 2: Input features for the Exit Probability model

As initially outlined in [2], the scenario illustrated in Fig. 2 highlights three spatial and kinematic features critical to predicting exit behavior: *lateral position*, *relative heading*, and *exit distance*.

- **Lateral position** reflects the radial distance from the roundabout center. Vehicles on the outer lane are statistically more likely to exit, while those in inner lanes tend to circulate further.
- **Relative heading** measures the angular deviation between the vehicle’s orientation and the tangent of the roundabout arc, indicating whether the vehicle is preparing to steer outward.
- **Exit distance** captures the normalized arc length remaining before the vehicle’s next exit point. Unlike the Euclidean distance used in [2], our method accounts for roundabout curvature, offering more topologically relevant context.

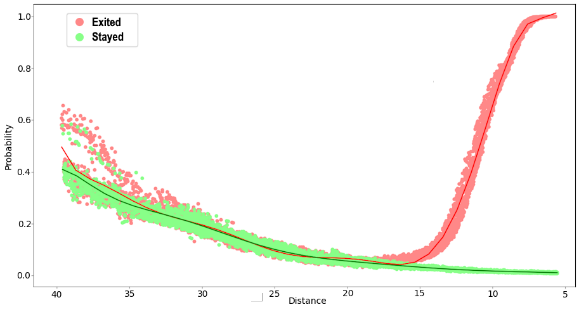


Fig. 3: REP divergence analysis: Polynomial Model

To evaluate the trained model, we conduct a series of experiments in the Town03 two-lane roundabout of the CARLA simulator. In this study, we focus on vehicles circulating in the outer lane, while leaving inner-lane complexities for future investigation.

As shown in Fig. 3, the REP output remains ambiguous at first but begins to diverge significantly approximately 15 meters before the exit. The x-axis indicates the arc distance between two successive exits (about 40 meters), and the y-axis shows the predicted probability of exit. The early divergence of the two trajectory states allows the ego vehicle to anticipate with confidence whether an inbound vehicle will exit, reducing unnecessary waiting time and improving flow.

This prediction is embedded directly into the PPO agent’s observation space. By including REP as a dynamic weight, the agent is empowered to reason probabilistically: even if a vehicle appears threatening based on position alone, it may still be deemed safe to proceed if its exit probability is sufficiently high. This integration results in more assertive yet safe decision-making around roundabout entries.

B. Comfort-Aware Time-To-Collision (TTC) Predictor

We develop a comfort-aware Time-To-Collision (TTC) predictor to estimate the ideal timing for the ego vehicle to safely and smoothly reach the roundabout merging point. The method builds upon the longitudinal control strategy proposed in [3], in which a first-order transfer function models vehicle dynamics and a PID controller ensures smooth convergence to a target speed, while satisfying passenger comfort constraints—specifically limiting acceleration and jerk.

The TTC predictor takes three input features; *current speed*, *current acceleration*, and *distance to the merging point* to produce a controlled TTC value. The ego vehicle’s longitudinal maneuver is modeled using a first-order transfer function:

$$\frac{V(s)}{U(s)} = \frac{\kappa}{\tau s + 1} \tag{1}$$

where:

- $V(s)$ is the vehicle speed in the Laplace domain,
- $U(s)$ is the throttle input signal,
- κ is the system gain,
- τ is the time constant.

To regulate acceleration while maintaining passenger comfort, we adopt a classical Proportional-Integral-Derivative (PID) control law:

$$u(t) = k_P e(t) + k_I \int_0^t e(\tau) d\tau + k_D \frac{de(t)}{dt} + \text{bias} \tag{2}$$

where:

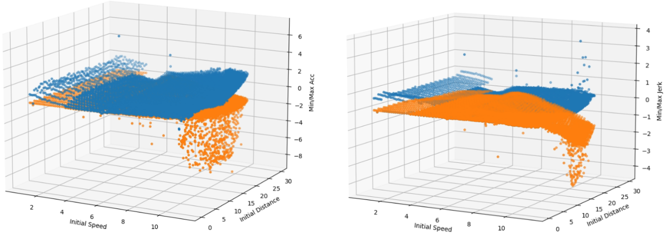
- $e(t) = v_{\text{target}} - v(t)$ is the velocity tracking error,
- k_P, k_I, and k_D are the PID gains,
- bias ensures smooth throttle transitions from initial conditions.

To evaluate the smoothness and comfort compliance of our PID-based control strategy, we visualize the maximum and minimum values of acceleration and jerk obtained during the data generation phase across a wide range of initial vehicle conditions.

To evaluate the comfort compliance of the PID controller, we assess the maximum and minimum acceleration and jerk values across a variety of initial conditions during the data generation phase.

Figure 4a presents the range of acceleration values for varying initial speeds and distances. The values remain within the comfort-compliant bounds, approximately between -2.5 and $+2.5$ m/s^2, confirming that the controller generates smooth transitions aligned with real vehicle dynamics. Some deviations at high initial speeds ($\geq$ 10 m/s) are observed, likely due to transitions between control regimes.

Figure 4b shows the corresponding jerk values. Most cases stay within acceptable thresholds, although a few configurations involving high initial speed and short distance exhibit minimum jerk values below -4 m/s^3, indicating more abrupt deceleration. Nevertheless, overall trends validate the controller's ability to maintain low jerk and ensure comfort-aware control.



(a) Max/Min acceleration vs. initial speed and distance.

(b) Max/Min jerk vs. initial speed and distance.

Fig. 4: Evaluation of comfort-aware PID control outputs.

This control setup enables the TTC predictor to produce comfort-aware reference timings that respect both physical dynamics and passenger comfort. These TTC estimates are later used to guide high-level decision-making in our reinforcement learning module.

Figure 5 illustrates the full two-stage process used to construct the TTC predictor. In the first stage, a PID controller receives high-level control parameters, including target speed, max-jerk limit, curvature radius , and horizon distance, as well as the current speed and throttle values, to generate a throttle command profile $\{u_0, \ldots, u_n\}$, executed in the CARLA simulator. The resulting outputs is recorded over time. In the second stage, these trajectories are used to create a supervised regression dataset. From each simulated run, features such as speed, acceleration, and distance are extracted, and the corresponding time-to-spot value is computed as the target label. This dataset is used to train an XGBoost regression model that forms our comfort-aware TTC predictor. The trained model serves as a key component in assessing safe and smooth roundabout entry conditions in real time.

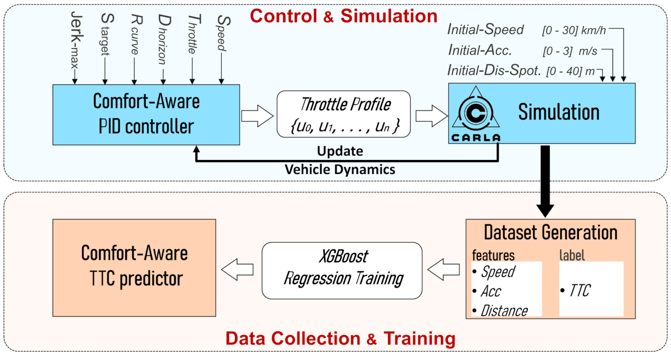

Control & Simulation
Jerk-max
S target
R curve
D horizon
Throttle
Speed
Initial-Speed [0 - 30] km/h
Initial-Acc. [0 - 3] m/s
Initial-Dis-Spot. [0 - 40] m
Comfort-Aware PID controller
Throttle Profile {u0, u1, . . . , un}
CARLA
Simulation
Update Vehicle Dynamics
Dataset Generation
features
label
• Speed
• Acc
• Distance
• TTC
XGBoost Regression Training
Comfort-Aware TTC predictor
Data Collection & Training

Fig. 5: Two-stage process for generating a comfort-aware Time-To-Collision (TTC) predictor.

V. Online Decision-Making: PPO-Based Control

Autonomous vehicle control during the roundabout maneuver involves managing three key control signals: throttle, brake, and steering. In our framework, the longitudinal control (throttle and brake) is learned by a deep reinforcement learning agent, while lateral control (steering) is handled using a rule-based path tracking method based on the Stanley controller.

A. Lateral Control via Stanley Steering

To ensure stable and interpretable lateral motion, we adopt a simplified bicycle model for steering control. The steering input δ is computed based on the heading error $\Delta\psi$—the angular difference between the vehicle's orientation and the path tangent—and the cross-track error e, defined as the lateral distance between the vehicle's front axle and the reference trajectory.

Following the Stanley control law [10], the steering angle is given by:

$$\delta(t) = \Delta\psi(t) + \tan^{-1}\left(\frac{K_e e(t)}{K_v + v(t)}\right) \tag{3}$$

where K_e is the cross-track gain, K_v is the speed-dependent gain, and $v(t)$ is the current vehicle speed. This path-tracking solution is general-purpose and can be applied across different vehicle models. However, the control performance depends on the tuning of K_e and K_v, which are calibrated in our case specifically for the Tesla Model 3 blueprint in the CARLA simulator. The reference trajectory is provided either by a topological planner or retrieved from a fixed map, guiding the ego vehicle along the appropriate lane and entry path. With steering handled through this rule-based strategy, our reinforcement learning agent is dedicated exclusively to learning longitudinal control—deciding when to throttle, brake, or yield—based on real-time traffic context and risk estimation.

B. RL-Based Longitudinal Control with PPO

The main goal of this section is to demonstrate how deep reinforcement learning agents can drive in realistic complex environments by analyzing the design decisions we make for our environment, agent, and network models. The implementation is in Python 3.7 with the package and dependency manager Poetry and PyTorch. All simulations and model training were conducted on a workstation equipped with an Intel Core i7-11700 CPU (8 cores, 16 threads, up to 4.9 GHz), 16 GB of RAM, and an NVIDIA GeForce GTX 1660 GPU with 6 GB of video memory. The experiments were performed under a Linux environment with CUDA 12.3.

We chose Proximal Policy Optimization (PPO) as the deep reinforcement learning algorithm for continuous control problems [11] that worked best in our tests. PPO is a model-free reinforcement learning algorithm based on policy gradients that stops divergence with a first-order trust region criterion.

Proximal Policy Optimization is particularly well-suited to our setting, offering stability in continuous control environments with high-dimensional state spaces. Its clipped surrogate objective prevents unstable policy shifts and supports gradual learning in complex driving scenarios.

We adopt an Actor-Critic reinforcement learning architecture as the foundation for training our Proximal Policy Optimization (PPO)-based agent. This decision is motivated by the continuous action space and the need for stable, sample-efficient learning in high-stakes driving environments such as roundabouts. The actor network generates continuous control actions—specifically throttle and brake signals—based on the ego vehicle's current observation. The critic network evaluates the quality (value) of the current state under the learned policy, providing a baseline that stabilizes policy updates. This dual-network architecture offers several advantages; (i) Efficient Policy Learning: By leveraging the critic's value estimates, the actor can focus on learning a policy that maximizes long-term rewards without high variance in returns. (ii)Support for Continuous Action Spaces: Unlike discrete-action methods (e.g., DQN), the actor directly outputs smooth throttle and brake values, enabling fine-grained longitudinal control necessary for comfort-sensitive driving. (iii) Compatibility with PPO: The architecture is naturally suited for PPO, which uses the actor's outputs and critic's state evaluations to apply clipped, stable policy updates. (iV) Exploration via Learnable Noise: The inclusion of a learnable log-standard deviation ($\log \sigma$) allows the agent to adapt its exploration behavior over time, encouraging wide behavioral sampling early in training and more refined actions in later phases.

C. Action Space

In CARLA simulator, the actions for controlling our agent are usually defined as a tuple of three values *(s, t, b)*. In that tuple *s* is our steer, *t* is the throttle and *b* is the brake. The steer action was described in section V-A. The input representation is then fed into our policy network, which consists of a multi-layer perceptron and outputs $(\hat{t}, \hat{b})$, where $\hat{t}$ is the predicted throttle action and $\hat{b}$ is the predicted brake for that timestep, as shown in Figure 6. The structure of our actor network—with progressively reducing hidden layer sizes (500 → 300 → 100)—was chosen to capture a rich latent representation of the high-dimensional observation space while maintaining computational tractability. Similarly, the critic shares this deep structure to provide accurate value predictions aligned with the policy's complexity.

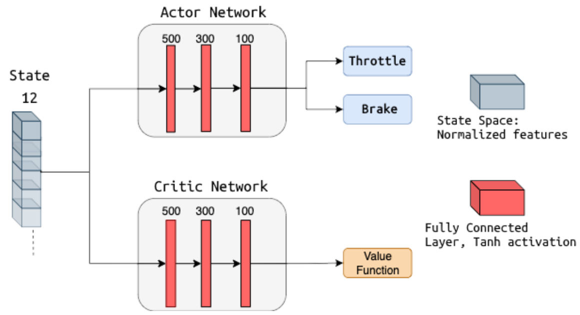


Fig. 6: Proposed Policy Network Architecture

D. State Space

Training is performed using mini-batch gradient descent, guided by Monte Carlo reward estimates. A high initial action standard deviation is used to promote exploration and is gradually reduced to ensure smoother, more deterministic actions as training progresses. The PPO agent's observation space includes:

- Ego-previous-action (throttle and brake) at time (t-1),
- Ego-centric state features (velocity, acceleration, distance to collision point),
- Adversarial-centric state features (velocity, distance to collision point),
- TTC predicted for the ego at time (t),
- TTC predicted for the ego at time (t-1),
- REP-derived exit likelihoods for inbound vehicles.

E. Curriculum Rewarding Strategy

To ensure stable and meaningful policy learning, we adopt a curriculum-based reward shaping strategy that decouples the learning of efficiency from comfort and risk awareness. This design addresses early-stage conflicts between performance, comfort, collision risk.

To facilitate efficient and stable learning, we employ a curriculum-based reward shaping strategy that introduces safety gradually, after the agent masters comfort and efficiency.

a) Step 1: Efficiency-Oriented Reward: Initially, the reward function focuses on core objectives to avoiding unnecessary braking or idling and encouraging forward motion when safe (without traffic inside roundabout). We opted to disable the Comfort-related penalties or weakly weighted during the first 100 episodes to avoid overwhelming early learning.

b) Step 2: Comfort-Oriented and Speed constraint Reward: Once efficiency is consistently achieved, the agent begins learning comfort-aware behavior. The reward function is extended to include; (i) penalties for excessive jerk and abrupt changes, (ii) rewards for smooth acceleration profiles, (iii)penalties for throttle-brake overlap, and (iv) rewards for maintaining a moderate throttle zone ($0.1 < \text{throttle} < 0.6$) with low jerk. A smooth transition is ensured by gradually increasing the weight of comfort.

c) Step 3: Collision-avoidance-Oriented Reward: In the final phase, the agent is trained to develop risk-aware behavior by incorporating safety-critical metrics—particularly Time-To-Collision (TTC)—into the reward function. This step introduces stronger penalties for unsafe decisions and encourages proactive yielding when necessary. The reward function includes:

- A strong penalty for actual collisions or near-miss events;
- A penalty based on the relative TTC difference between ego vehicle and the closest inbound vehicle:

  $$\Delta \text{TTC} = \text{TTC}_{\text{inbound}} - \text{TTC}_{\text{ego}}$$

  Unsafe decisions are penalized when ΔTTC is negative (i.e., the ego vehicle is expected to reach the merging point after an inbound vehicle);
- A scaled reward when the ego vehicle proceeds safely with a sufficiently positive TTC margin;

VI. Results

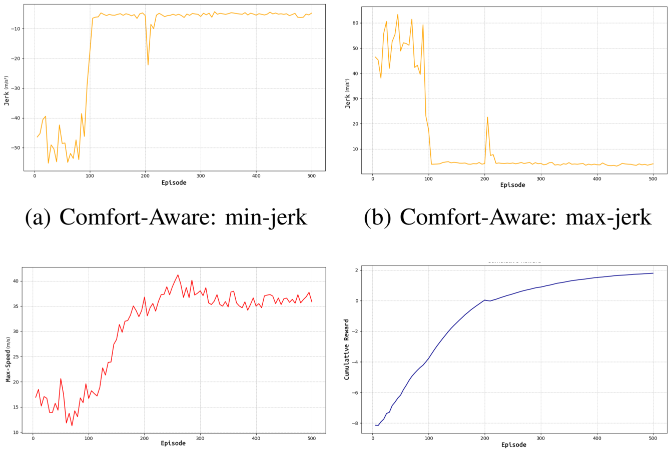
To assess the effectiveness of our proposed framework, we evaluate the trained RL agent under varying traffic conditions and reward configurations, with a particular focus on safety, efficiency, and passenger comfort. The evaluation pipeline is divided into three stages, each designed to test the agent's ability to generalize, adapt, and comply with driving constraints in realistic roundabout scenarios. The sections below describe in detail the training parameters and evaluation metrics used to validate the performance of our PPO-based control policy.

A. Training Parameters and Constraints

We constrain the vehicle's longitudinal behavior with a maximum speed limit of 35 km/h and apply a comfort-aware control policy by penalizing jerk values exceeding 5 m/s^3. To facilitate robust and stable learning, we implement a curriculum-based reward shaping strategy that introduces behavioral objectives progressively: starting with efficiency-focused speed acquisition, followed by comfort-oriented smoothness, enforcement of speed limits, and finally, safe interaction with traffic to encourage collision avoidance. The training takes place in a critical entry zone located approximately 40 meters before the potential collision point representing the area where human drivers typically begin assessing whether to yield or proceed into the roundabout. To ensure realism, we dynamically spawn inbound vehicles ahead of this decision zone, with randomized initial speeds uniformly sampled between 3km/h and 25km/h. This setting exposes the agent to diverse traffic dynamics and enhances generalization to realistic entry conditions.

B. Evaluation: Without-Traffic

In a simplified roundabout environment without circulating vehicles, the agent is initially trained to accelerate efficiently and gain speed. Comfort objectives are then introduced by encouraging smooth acceleration profiles and minimizing jerk. Finally, a speed limit constraint is applied through reward shaping, enabling the agent to learn rule-compliant behavior while maintaining comfort and efficiency. Figs. 7a-7b present the minimum and maximum jerk values during training. Notably, after episode 100, comfort constraints are activated in the reward function, encouraging smooth vehicle behavior. As shown, the jerk values stabilize and remain consistently bounded within the threshold of ±6 m/s^3, validating the agent's ability to respect comfort limits and reduce abrupt acceleration or deceleration, even under varying initial conditions.



(a) Comfort-Aware: min-jerk

(b) Comfort-Aware: max-jerk

(c) Traffic-Rules: max-speed

(d) Reward: without-traffic

Fig. 7: Evaluation of comfort-aware control parameters: speed and jerk under various initial conditions.

Figure 7c illustrates the evolution of maximum speed during training. After episode 200, the reward function introduces a speed-limit constraint set at 35 km/h. The graph shows a noticeable adjustment in the agent's behavior, with average speeds gradually converging toward the limit, demonstrating successful adherence to traffic rules. Finally, Figure 7d shows the cumulative reward trend throughout training. Initially, the reward increases rapidly as the agent learns basic movement and efficiency. However, after introducing the speed-limit penalty at episode 200, the reward curve shows slower but stable improvement, indicating that the agent is adapting to stricter control requirements while preserving overall performance.

C. Evaluation: With-Traffic

Introducing vehicles circulating inside the roundabout significantly increases the complexity of the scene, requiring the agent to learn when to yield to higher-priority inbound vehicles and when to safely initiate entry. Leveraging the ego vehicle's modeled Time-To-Collision (TTC), the TTC of inbound vehicles, and REP-derived exit probabilities, the agent is equipped to make more informed decisions that minimize unnecessary waiting and reduce collision risk. The reward function is designed to encourage cautious decision-making—rewarding braking when required and penalizing unsafe throttle actions. As illustrated in Fig. 8, from episode 600 onward, the reward function begins penalizing the agent for attempting to enter the roundabout under risky conditions. This adjustment results in a temporary drop in cumulative reward, followed by a recovery and convergence approximately 800 episodes later. This behavior confirms that the agent learns to respect safe temporal gaps and prioritize yielding in the presence of other vehicles, thereby improving both safety and traffic negotiation performance in realistic roundabout scenarios.

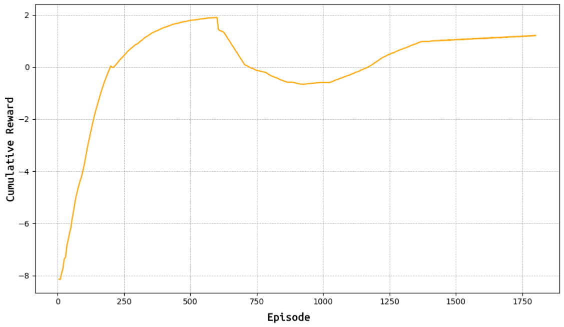


Fig. 8: Cumulative Reward: with-traffic

VII. Conclusion & Future work

This paper introduced a two-phase deep reinforcement learning (DRL) framework to enhance autonomous vehicle decision-making during roundabout entry, emphasizing safety, efficiency, and passenger comfort. In the preparation phase, we trained two predictive models: a logistic regression-based Roundabout Exit Probability (REP) model to anticipate inbound vehicle behavior, and a comfort-aware Time-To-Collision (TTC) predictor that accounts for smooth acceleration dynamics using a PID-based control strategy. These models were integrated into a Proximal Policy Optimization (PPO) framework for online control. A curriculum-based reward shaping strategy is employed to guide the PPO agent through progressive learning stages—starting with efficiency, advancing to comfort-aware control, and culminating in safety-focused decision-making in dynamic, multi-agent environments. Experimental evaluations conducted in CARLA demonstrated that our strategy enables safe and robust roundabout negotiation, while maintaining passenger comfort and driving fluidity. For future work, we aim to extend the agent's capabilities to handle more complex multi-lane lookup, integrate real-time lane detection from perception modules. Additionally, we plan to investigate multi-agent coordination through cooperative reinforcement learning to enable more advanced negotiation strategies between autonomous vehicles in shared traffic scenarios.

VIII. Acknowledgements

This work has benefited from a government grant managed by the "*Agence Nationale de la Recherche*" under the France 2030 program with NF-FITNESS, grant "ANR-22-PEFT-0007".

References

[1] European Telecommunications Standards Institute, *Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Part 2: Specification of Cooperative Awareness Basic Service*, ETSI Std. EN 302 637-2 V1.4.1, Apr. 2019. [Online]. Available: https://www.etsi.org/deliver/etsi_en/302600_302699/30263702/01.04.01_60/en_30263702v010401p.pdf

[2] D. Deveaux *et al.*, "A knowledge networking approach for ai-driven roundabout risk assessment," in *2022 IEEE International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2022, pp. 3249–3254.

[3] S. Moradi, A. Nadar, and J. Härri, "Comfort-based ai-driven roundabout control for automated vehicles," vol. 8, no. 2, pp. 1112–1123, *8th International Conference on Models and Technologies for Intelligent Transportation Systems*, june 14-16, 2023, EURECOM, France.

[4] F. G. Cuenca *et al.*, "Autonomous driving in roundabout maneuvers using reinforcement learning with q-learning," in *2022 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2022, pp. 701–706.

[5] B. Ben Elallid *et al.*, "Deep reinforcement learning for autonomous vehicle intersection navigation," in *2021 International Conference on Advanced Robotics and Mechatronics (ICARM)*. IEEE, 2021, pp. 215–220.

[6] D. Gutiérrez-Moreno, J. A. Álvarez García, and L. M. Soria-Morillo, "Reinforcement learning-based autonomous driving at intersections in carla simulator," *Sensors*, vol. 22, no. 22, p. 8373, 2022.

[7] L. Anzalone, S. Barra, and M. Nappi, "Reinforced curriculum learning for autonomous driving in carla," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 3318–3322.

[8] H. Yuan, P. Li, B. van Arem, L. Kang, and Y. Dong, "Safe, efficient, comfort, and energy-saving automated driving through roundabout based on deep reinforcement learning," *arXiv preprint arXiv:2306.11465*, 2023. [Online]. Available: https://arxiv.org/abs/2306.11465

[9] A. P. Capasso, G. Bacchiani, and D. Molinari, "Intelligent roundabout insertion using deep reinforcement learning," *arXiv preprint arXiv:2001.00786*, 2020. [Online]. Available: https://arxiv.org/abs/2001.00786

[10] G. M. Hoffmann, C. J. Tomlin, M. Montemerlo, and S. Thrun, "Autonomous automobile trajectory tracking for off-road driving: Controller design, experimental validation and racing," in *2007 American Control Conference*, 2007, pp. 2296–2301.

[11] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.