# E2E Autonomous Parking for Hyper-Realistic Truck and Trailer Dynamics

**Oliver Berton***
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
oberton@andrew.cmu.edu

**Rodrigo Lopes Catto***
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
rlopesca@andrew.cmu.edu

## Abstract

This work presents end-to-end autonomous reverse parking for a tractor–semitrailer rig in a hyper-realistic simulated logistics yard. While recent work on autonomous parking has largely focused on passenger cars in simplified environments, the "last-meters" depot parking problem for articulated trucks remains challenging due to unstable trailer dynamics, jackknife constraints, and tight spatial tolerances. We build a full pipeline in CARLA 0.9.14, integrating a realistic tractor–trailer vehicle and customizing an industrial parking map, and collect more than 300 expert teleoperation trajectories using a Logitech G29 wheel. On top of this environment we design *TruckNet*, a low-dimensional policy trained with Proximal Policy Optimization (PPO), warm-started by a behavior cloning (BC) baseline and optimized with a curriculum over spawn difficulty. Our best policy achieves a success rate of 0.86 on a reverse-parking task into a single docking bay, while satisfying jackknife and collision constraints. We analyze the role of reward shaping and curriculum design, and discuss the remaining gaps to robust performance under harder spawn distributions and more cluttered layouts.

## 1 Introduction

Most industrial autonomous trucking research has focused on long-horizon highway driving and lane keeping, where the environment is structured and high-level autonomy can be layered on top of well-understood control stacks. In contrast, the "last-meters" problem, precise depot parking and docking, is still largely handled by human drivers, despite being a frequent source of time loss, damage, and safety incidents in logistics hubs.

Reverse parking a tractor–trailer rig into a loading bay is particularly challenging: the kinematics are highly non-linear and unstable when backing up, the trailer can jackknife if the steering is not carefully managed, and the vehicle must plan over long horizons with tight spatial tolerances and occlusions. Traditional control solutions typically rely on handcrafted maneuvers or carefully tuned controllers, and do not easily generalize across yard layouts and loading configurations.

Recent work has considered deep reinforcement learning (RL) and end-to-end policies for autonomous parking, but primarily in the context of passenger cars in CARLA or similar simulators [1–3]. These works typically assume simpler vehicle dynamics and focus on visual perception rather than articulated truck behavior. To the best of our knowledge, end-to-end RL for depot parking with a realistic tractor–semitrailer model in CARLA is underexplored.

In this project we take a first step toward this setting and study the following question:

*Equal contribution. Work done as part of the 10-703 Deep Reinforcement Learning course at Carnegie Mellon University.

Preprint. Under review.

*Can a low-dimensional RL agent, trained only on compact state features and without any handcrafted controller, reliably learn to reverse park a tractor–trailer rig in a realistic simulated logistics yard?*

### 1.1 Contributions

Our main contributions are:

- We integrate a validated tractor–semitrailer model into CARLA 0.9.14 and customize an industrial parking map by editing collision geometries and static objects to enable realistic reverse-parking maneuvers.
- We build a teleoperation data-collection pipeline with a Logitech G29 steering wheel and collect over 300 high-quality reverse-parking trajectories from randomized start and goal states, logging both continuous control actions and low-dimensional state features.
- We propose *TruckNet*, a compact MLP-based policy trained with PPO on a low-dimensional state space, warm-started from a behavior cloning baseline and optimized with a simple but effective curriculum over spawn difficulty.
- We empirically evaluate TruckNet on a reverse-parking task into a single docking bay and achieve a success rate of $0.86$ in the hardest curriculum stage we considered, analyzing the effects of our reward shaping and curriculum design and highlighting limitations.

## 2 Related Work

Autonomous parking has been widely studied for passenger vehicles, both with classical planning and control and with end-to-end learning. Recent works have used CARLA to benchmark deep RL approaches for car parking. Lazzaroni et al. propose an RL-based automated parking system in CARLA with a three-phase curriculum and sparse visual inputs [1]. Yang et al. present an end-to-end neural network for visual autonomous parking in CARLA [2]. Chen et al. introduce a control-aided attention mechanism for end-to-end visual autonomous parking [3].

In contrast, we focus on articulated truck–trailer dynamics, which introduce additional non-linearities and jackknife constraints often ignored in car-focused work. Prior work on truck–trailer vehicles has been build on top of simplified dynamics models [4]. In contrast, our environment builds on an improved and validated tractor–semitrailer vehicle for CARLA [5]. From a reinforcement learning perspective, our approach follows standard PPO with low-dimensional state inputs and curriculum learning [6], rather than high-dimensional visual policies.

## 3 Method

### 3.1 Problem formulation and environment

### 3.2 CARLA environment and map

We build our experiments on CARLA 0.9.14 using a publicly available tractor–semitrailer vehicle model with realistic articulation and collision geometry given by the vehicle model that was validated using measurement data from a DAF XF95 truck[5]. The base map is an industrial parking facility, which we further customize by:

- removing pole lights, barriers, and static assets whose collision meshes were unrealistically large or misaligned;
- adjusting spawn regions to allow long reverse paths into the selected bay.

This results in a "logistics yard" environment where the truck enters the yard forward and must perform a reverse maneuver into a target bay as shown in Figure 1.

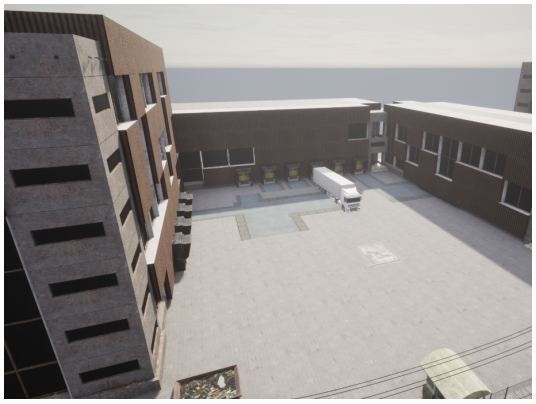


Figure 1: Custom made logistic yard on Carla Sim

### 3.3 Task definition

Each episode is defined by a start pose for the truck and trailer and a target bay pose. At the beginning of an episode, the truck is spawned forward-facing somewhere in a rectangular region in front of the docking bays, with randomized longitudinal offset, lateral offset, and heading. The goal is to reverse into the assigned bay such that:

- the trailer is aligned with the dock within a position tolerance of $0.60$ m and a yaw tolerance of $5^\circ$;
- the tractor–trailer configuration is not jackknifed (articulation angle between tractor and trailer is within $10^\circ$);
- no collisions occur with buildings or static objects.

Episodes terminate when the goal condition is met, a collision or jackknife occurs, the vehicle exits the yard bounds, or a maximum of $1000$ simulation steps is reached.

### 3.4 State and action spaces

Our initial data-collection pipeline and model design included four camera views around the tractor–trailer. However, we quickly realized that the required computation and storage would exceed our available resources for this project. As a result, we pivoted towards using ray-based range sensors, with the final state-space representation consisting of:

- relative position of the truck to the goal in the yard frame $(x_{\text{rel}}, y_{\text{rel}}, \psi_{\text{rel}})$
- longitudinal and lateral velocities $(v_x, v_y)$ at the truck;
- trailer articulation angle $\theta_{\text{trailer}}$.
- 6 Ray-Based Range Sensors, positioned at the front, left, and right of the cab, and the back, left, and right of the trailer.
- Current gear (forward or reverse),

Thus the state vector $s_t$ combines position, velocity, trailer angle, and external environment information.

The continuous action vector $a_t$ comprises:

- steering angle command (normalized to $[-1, 1]$)
- throttle command (normalized to $[0, 1]$)
- brake command (normalized to $[0, 1]$)
- gear command (boolean, 0 for forward, 1 for reverse)

The control inputs are passed through CARLA's vehicle control interface, with internal limits enforcing maximum steering and acceleration.

### 3.5 TruckNet policy architecture

We employ a feed-forward multilayer perceptron (MLP) policy, *TruckNet*, within the PPO framework. The input is a 60-dimensional stacked observation vector constructed from the current state and its four most recent predecessors, providing short-term temporal context without requiring recurrence. Each state includes kinematic features, trailer articulation, a reverse indicator, and ray-based sensor readings.

TruckNet consists of a shared two-layer MLP (512 units per layer, ReLU activations), followed by separate policy and value heads. The policy head outputs the mean of a Gaussian distribution over four continuous control actions (throttle, brake, steering, gear), with a trainable log-standard-deviation vector. The value head outputs a scalar estimate $V_\phi(s)$.

A schematic of the architecture is shown in Fig. 2.

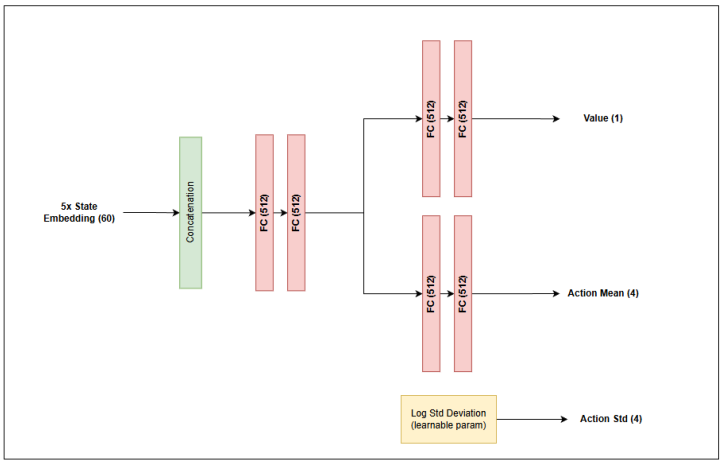



Figure 2: TruckNet policy network. Shared MLP feeds into policy and value heads.

### 3.6 Behavior cloning

As a baseline and warm-start for RL, we first train a behavior cloning (BC) policy on the expert trajectories collected via teleoperation, as shown in Fig. 3. Each trajectory consists of sequences $\{(s_t^{\text{exp}}, a_t^{\text{exp}})\}_{t=1}^{T}$ from randomized start and goal configurations.

Figure 3: Data Collection rig using a Logitech G29 Steering Wheel

To train the BC policy $\pi_\theta^{\text{BC}}(a \mid s)$, we optimize the above TruckNet network $\theta$ using minimum log-likelihood, the objective function of which is given in Eq. (1).

$$\mathcal{L}_{\text{BC}}(\theta) = -\sum_t \log \pi_\theta(a_t^{\text{exp}} \mid s_t^{\text{exp}}). \tag{1}$$

This BC stage provides an initial policy that captures the core reversing and alignment behaviors, which the RL stage then refines.

### 3.7 Reward design for reverse truck parking

Reverse truck parking is challenging due to sparse success signals, nonlinear dynamics, and the need to avoid unsafe behaviors such as jackknifing. We use a structured dense reward composed of progress, alignment, motion, and safety terms, together with a terminal success bonus:

$$r_t = r_{\text{time}} + r_{\text{progress}} + r_{\text{alignment}} + r_{\text{motion}} + r_{\text{safety}} + r_{\text{success}} \tag{2}$$

**Time penalty.** A small constant penalty discourages stalling:

$$r_{\text{time}} = -0.005 \tag{3}$$

**Progress reward.** Reward is provided only when the agent moves closer to the goal. Let $d_t$ denote the distance to the target:

$$r_{\text{progress}} = \begin{cases} 0.25 \cdot \text{clamp}(d_{t-1} - d_t, -1, 1) + 0.02 \cdot \min\left(1, \frac{d_{t-1}-d_t}{0.1}\right), & \text{if } d_{t-1} - d_t > 0.02, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

**Alignment reward.** When progress is made, the agent receives shaping rewards for truck and trailer alignment:

$$r_{\text{alignment}} = 0.05 \cdot \underbrace{0.5\Big(1 - \min(\theta_{\text{heading}}/90, 1) + 1 - \min(\theta_{\text{trailer}}/90, 1)\Big)}_{\text{align\_score}} \tag{5}$$

**Motion reward.** Moderate, nonzero velocity is encouraged:

$$r_{\text{motion}} = \begin{cases} 0.01, & 0.05 < \|\mathbf{v}_t\| < 5.0 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

**Safety penalties.** Large heading or articulation errors incur penalties, and severe trailer angles terminate the episode:

$$\begin{aligned} r_{\text{safety}} = & -0.002 \cdot \min(|\theta_{\text{heading}}|, 30) \\ & -0.004 \cdot \min(|\theta_{\text{trailer}}|, 30) \end{aligned} \tag{7}$$

$$|\theta_{\text{trailer}}| > 55^\circ \quad \Rightarrow \quad r_{\text{safety}} = -3.0 \text{ (termination)} \tag{8}$$

**Success reward.** Reaching the goal yields a large terminal reward with additional shaping:

$$r_{\text{success}} = 150 + r_{\text{shaping}} \tag{9}$$

$$r_{\text{shaping}} = \max\left(0, \frac{5(3.5 - d_t)}{3.5}\right) + \max\left(0, \frac{5(9 - \theta_{\text{heading}})}{9}\right) + \max\left(0, \frac{5(12 - \theta_{\text{trailer}})}{12}\right) \tag{10}$$

This reward design encourages steady progress, accurate alignment, and safe maneuvering while maintaining a strong terminal incentive to complete the task. All shaping coefficients were tuned empirically to provide informative gradients without inducing reward exploitation.

### 3.8 Curriculum learning

Directly training from uniformly random initial states across the full yard proved ineffective: sparse rewards and nonlinear truck–trailer dynamics often caused PPO to collapse into oscillatory or jackknifed behaviors. To address this, we employ a simple curriculum. At difficulty 0, the truck and trailer are initialized approximately aligned with the loading bay at a distance of $3 \pm 3$ m and zero lateral offset. We track performance using a moving success-rate window of 200 episodes, and advance the curriculum by 0.1 in difficulty once the average success rate exceeds 0.80. Each increase linearly scales the initialization distance and lateral offset until, at difficulty 1, the truck starts at $15 \pm 3$ m from the bay with up to $\pm 3$ m lateral offset, requiring the agent to discover nontrivial reversing trajectories. This staged procedure enables the policy to gradually acquire meaningful parking behaviors. Figure 4 visualizes the spawn regions at different difficulty levels.

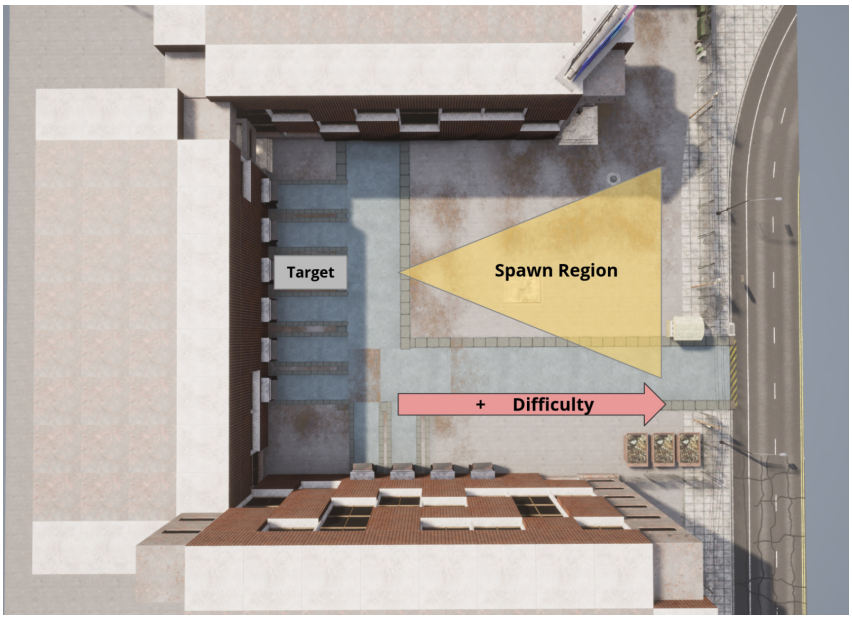



Figure 4: Spawn region relative to target as a function of difficulty.

### 3.9 PPO training

After being initialized via behavior cloning, TruckNet is trained with Proximal Policy Optimization (PPO) [6] using the clipped surrogate objective:

$$\mathcal{L}^{\text{CLIP}}(\theta) = \mathbb{E}_t\Big[\min\big(r_t(\theta)\hat{A}_t,\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\big)\Big], \tag{11}$$

where $r_t(\theta) = \pi_\theta(a_t|s_t)/\pi_{\theta_{\text{old}}}(a_t|s_t)$, $\hat{A}_t$ is the estimated advantage, and $\epsilon$ is the clip coefficient. Returns and advantages are normalized for training stability.

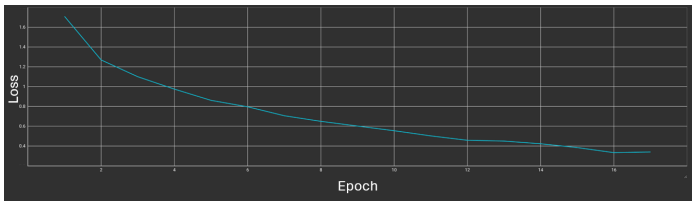
Key hyperparameters are:

- Rollout length: 4096
- Minibatch size: 128
- Clip coefficient: 0.15
- Value loss coef.: 0.2
- Entropy coef.: 0.001
- Max episode length: 1000
- $\gamma$: 0.99
- $\lambda$: 0.98
- Learning Rate: 1e-4
- Epochs Per Update: 6

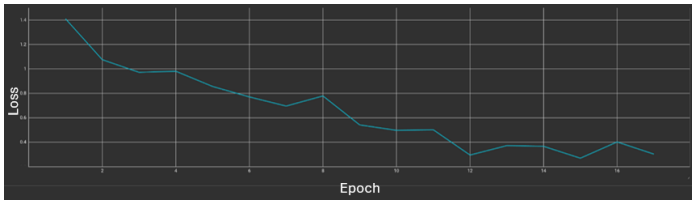
# 4 Experiments and results

## 4.1 Behavior cloning baseline

To train the behavior cloning (BC) policy, we collected 300 rollouts, corresponding to roughly 250,000 state-action pairs. The rollouts were generated using the same spawning logic as difficulty level 1.0 during PPO training (see Figure 5).




(a) Training loss




(b) Validation loss

Figure 5: Behavior Cloning policy losses. The training loss steadily decreases, while the validation loss closely follows, with a final validation loss of 0.28.

Despite effective training, the BC policy achieves a **0% success rate** in completing the reverse parking task. It can perform simple reverse motions when initialized near expert trajectories but fails when starting from out-of-distribution states. Additionally, small deviations often lead to overcorrections, causing the trailer to approach jackknife configurations.

These observations demonstrate that while BC captures basic motion patterns, it does not learn a robust feedback control strategy, motivating subsequent RL fine-tuning.

## 4.2 PPO with curriculum

Using PPO with the curriculum described in Section 3, the agent gradually learns increasingly complex reverse maneuvers. Figure 6 shows the success rate during training, with each black dashed line marking a curriculum difficulty increase. After training, the agent achieves an **86% success rate on difficulty 1.0**.

Qualitatively, successful trajectories resemble human-like reverse maneuvers: the agent pulls forward to straighten the trailer and then executes a controlled reverse arc into the dock. Failures are typically due to sensitivity to braking dynamics, where the agent oscillates between throttle and brake and fails to settle within the goal region before the episode ends.

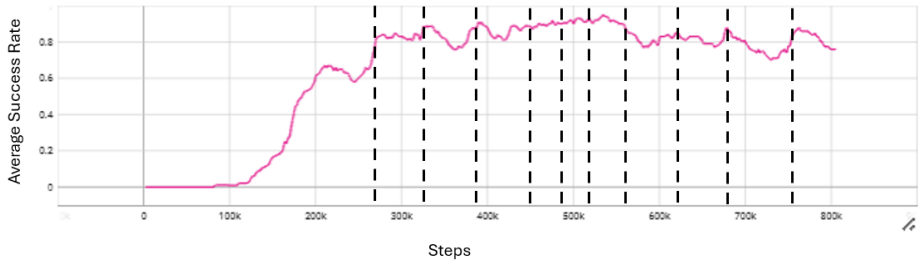



Figure 6: PPO curriculum training success rate. Each black dashed line indicates an increase in difficulty.

### 4.3 Ablations and observations

We conducted small-scale ablations on key reward components and the training curriculum to better understand their impact on learning stability and policy behavior. Due to the high computational cost and variance of training in CARLA, these ablations are qualitative and based on a limited number of runs, but consistently reveal the same failure modes and learning dynamics.

- **Trailer alignment penalty.** Removing the trailer alignment term substantially increases the frequency of jackknifing and yields policies that reach the target pose while leaving the trailer significantly misaligned, indicating that positional rewards alone are insufficient.
- **Curriculum learning.** Training directly from the full-difficulty distribution results in highly unstable learning. Success rates fluctuate widely and often collapse to near zero as PPO converges to trivial strategies (e.g., minimal motion) that avoid large negative rewards rather than executing meaningful maneuvers.
- **Behavior cloning warm-start.** Initializing PPO from a behavior cloning policy reduces the number of iterations required to achieve high success rates in early curriculum stages. However, this advantage diminishes as the curriculum progresses to the most challenging configurations, where RL fine-tuning dominates performance.

Overall, these ablations suggest that reward shaping, curriculum learning, and BC initialization play complementary roles, and that their combination is important for achieving stable and effective learning in this articulated parking task.

## 5 Limitations and future work

Our study has several limitations:

- **Scope of environment.** We evaluate on a single logistics-yard layout with a fixed row of docking bays. Generalization to different yard geometries, dock positions, and road conditions is not studied.
- **Low-dimensional state only.** We remove camera feed from the loop and assume access to perfect state estimates from the simulator. In real deployments, noise and partial observability would require integrating perception modules or learning directly from raw sensor data.
- **Limited robustness analysis.** Our experiments focus on a single vehicle model and do not explore variations in trailer length, vehicle mass, or tire friction, which are important for real-world robustness.
- **Compute and statistical analysis.** Due to limited time and resources, we did not perform extensive multi-seed runs or hyperparameter sweeps; thus our reported success rate should be interpreted as a proof-of-concept rather than a fully optimized benchmark.

In future work we plan to:

- extend the curriculum and reward shaping to handle cluttered yards with other parked vehicles and obstacles. We have already constructed a CARLA map that can be used for this case.
- increase randomization in spawn configurations and physical parameters to improve robustness.
- incorporate visual inputs and study end-to-end vision-based truck parking policies.
- explore model-based RL or hybrid control/RL schemes to better handle the stiff braking dynamics observed in our simulator.

## 6 Conclusion

This work studies the problem of learning reverse parking for an articulated tractor–trailer in a realistic CARLA depot environment using reinforcement learning. We show that this task is difficult for naïve end-to-end RL due to the unstable backing dynamics and long-horizon planning requirements.

Stable and effective learning only emerges when behavior cloning warm-starting, reward shaping, and curriculum learning are combined. Behavior cloning provides an initial feasible backing policy, shaped rewards encourage trailer alignment and suppress jackknifing, and the curriculum gradually increases task difficulty, allowing PPO to refine long-horizon reverse maneuvers.

Using this structured training approach, the learned policy achieves a 0.86 success rate on the hardest spawn distribution and exhibits human-like reverse parking behavior. Although our experiments are limited to a single yard layout and assume perfect state information, the results suggest that carefully designed training structure is key to making articulated truck backing tractable with reinforcement learning, and provide a foundation for future work on perception-integrated systems, diverse depot layouts, and real-world robustness.

## References


[1] L. Lazzaroni, A. Pighetti, F. Bellotti, A. Capello, M. Cossu, and R. Berta. Automated Parking in CARLA: A Deep Reinforcement Learning-Based Approach. In *Applications in Electronics Pervading Industry, Environment and Society (ApplePies 2023)*, Lecture Notes in Electrical Engineering, Vol. 1110, Springer, 2024, pp. 352–357.

[2] Y. Yang, D. Chen, T. Qin, X. Mu, C. Xu, and M. Yang. E2E Parking: Autonomous Parking by the End-to-End Neural Network on the CARLA Simulator. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, 2024, pp. 2360–2382.

[3] C. Chen, S. Yao, Y. He, F. Tao, R. Song, Y. Guo, X. Huang, C. Wu, L. Ren, and C. Feng. End-to-End Visual Autonomous Parking via Control-Aided Attention. arXiv preprint arXiv:2509.11090, 2025.

[4] E. Bejar and A. Morán. Backing up control of a self-driving truck-trailer vehicle with deep reinforcement learning and fuzzy logic. In *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pages 202–207, 2018. 10.1109/ISSPIT.2018.8642777.

[5] A. Behera. Improved and Validated Tractor-Semitrailer Vehicle for CARLA Simulator. GitHub repository, 2025. `https://github.com/abhijeetbehera97/Carla-Tractor-Semitrailer/tree/main`.

[6] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal Policy Optimization Algorithms. arXiv preprint arXiv:1707.06347, 2017.