Appendix

No Author Given

No Institute Given

Appendix A Baseline DRL Models

DQN DQN [1] is a typical DRL model inspired by [2], where the goal of the agent is to interact with the environment by selecting actions in such a way that maximises cumulative future rewards. DQN approximates the optimal action-value function as follows:

$$Q^*(s,a) = \mathbb{E}_{s' \sim \epsilon}\left[r + \gamma \max_{a'} Q^*(s',a') | s, a\right] \tag{1}$$

where $Q^*(s,a)$ is the maximum sum of rewards discounted by γ at each time step, achievable by following any strategy in a state s and taking action a. This function follows the Bellman equation: if the optimal value $Q^*(s',a')$ of the state s' at the next time step is known for all possible actions, then the optimal strategy is to select the action a' that maximises the expected value of $r + \gamma Q^*(s',a')$.

DQN has some advancements that are used for autonomous driving. First, DQN employs a mechanism called experience replay that stores the agent's experience $e_t = (s_t, a_t, r_t, s_{t+1})$ at each time step t in a replay memory $D_t = e_1, \ldots, e_t$. During learning, samples of experience $(s,a,r,s') \sim U(D)$ are drawn uniformly at random from the replay memory, which removes the correlations in the observation sequence and smooths over changes in the data distribution. Second, DQN uses an iterative update to adjust the action values Q towards the target values. A neural network function approximator (Q-network) with weights θ is then used to estimate the Q function. The Q-network updates only the target values periodically to reduce correlations with the target, and the Q-network is trained by minimising the following loss function L at iteration i:

$$L_i(\theta_i) = \mathbb{E}_{(s,a,r,s') \sim U(D)}\left[\left(r + \gamma \max_{a'} Q(s',a';\theta_i{}^-) - Q(s,a;\theta_i)\right)^2\right] \tag{2}$$

where θ_i denotes the parameters of the Q-network at iteration i. $\theta_i{}^-$ are the parameters of the target network at iteration i, which are held fixed between individual updates and are updated only with the Q-network parameters θ_i every specified number of steps.

A2C A2C is a synchronous, deterministic variant of asynchronous advantage actor-critic (A3C) [3]. For autonomous driving, A2C maintains a policy $\pi(a_t|s_t;\theta)$ and an estimate of the value function $V(s_t;\theta_v)$, which uses the same mix of n-step returns to update the policy and the value function. The policy and value

function is updated every t_{max} action or when a terminal state is reached. The update performed by A2C can be denoted as $\nabla_{\theta'} \log \pi\left(a_t s_t; \theta'\right) A\left(s_t, a_t; \theta, \theta_v\right)$, and $A(s_t, a_t; \theta, \theta_v)$ denotes the advantage function, as follows:

$$A(s_t, a_t; \theta, \theta_v) = \sum_{i=0}^{k-1} \gamma^i r_{t+i} + \gamma^k V(s_{t+k}; \theta_v) - V(s_t; \theta_v) \tag{3}$$

where k denotes the number of actions taken since time step t and is upper-bounded by t_{max}.

PPO PPO is a model-free, actor-critic, and policy-gradient approach to maintain data efficiency and reliable performance [4]. In PPO, π denotes the policy network optimised with parameterisation θ, and the policy network takes state s as the input and outputs an action a. PPO uses actor-critic architecture to enable learning of better policies by reformulating reward signals in terms of advantage A. The advantage function measures how good an action is compared to the other actions available in the state s. PPO maximises the surrogate objective function as follows:

$$L(\theta) = \hat{\mathbb{E}}_t[min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)] \tag{4}$$

where $L(\theta)$ is the policy gradient loss under parameterisation θ. $\hat{\mathbb{E}}_t$ denotes the empirically obtained expectation over a finite batch of samples, and $\hat{A}_t$ denotes an estimator of the advantage function at timestep t. ϵ is a hyperparameter for clipping, and $r_t(\theta)$ is the probability ratio formulated as:

$$r_t(\theta) = \frac{\pi_\theta(a_t, s_t)}{\pi_{\theta old}(a_t, s_t)} \tag{5}$$

For autonomous driving, to increase the sample efficiency, PPO uses importance sampling to obtain the expectation of samples gathered from an old policy $\pi_{\theta old}$ under the new policy π_θ. As π_θ is refined, the two policies diverge, and the estimation variance increases. Therefore, the old policy is periodically updated to match the new policy. The clipping of the probability ratio to the range of $[1-\epsilon, 1+\epsilon]$ ensures that the state transition function is similar between the two policies. The use of clipping discourages extensive policy updates that are outside the comfort zone.

Appendix B Driving Scenario

To investigate the safety of DRL for road traffic junction environments based on the highway simulation system [5], we considered intersection (**Figure 1-A**) and roundabout (**Figure 1-B**) scenarios characterised as relatively challenging driving interflow environments.

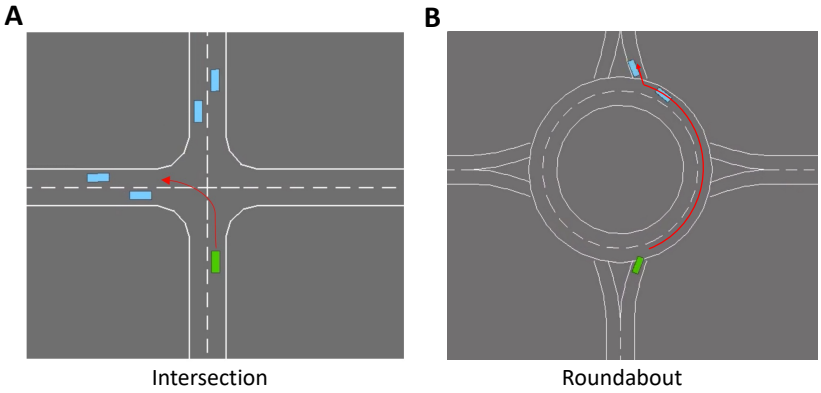

Fig. 1. Two representative driving scenarios in the road traffic junctions: (A) intersection; (B) roundabout. The ego vehicle with green colour is driving as guided by the red arrow to reach the goal.

Intersection Scenario

- *Settings.* The intersection scene comprises two roads crossing perpendicularly and stretching in four directions (north, south, east, and west). The roads are populated with the ego vehicle and several surrounding vehicles. The positions, velocities and destinations of the other vehicles are randomly initialised. The ego vehicle drives from 60 metres south to the intersection, and the task of the ego vehicle is to cross the intersection and turn left (west). The goal is achieved (i.e., destination reached) if the ego vehicle can turn left at the intersection and drive 25 metres or more from the turning point in 13 seconds. Please note that the horizontal road has the right of way in the intersection scenario. **Figure 1-A** is a showcase of the intersection scenario, where the goal of the ego vehicle shown in green is to turn left, as indicated by the red arrow.
- *Observations.* The vehicle observations of the intersection scenario follow the kinematic bicycle model [6], which implements simplistic behaviour and considers the same lane interactions to prevent potential lateral collisions [5]: each vehicle predicts the future positions of neighbouring vehicles over a three-second horizon. If a collision with a neighbour is predicted, the yielding vehicle is decided according to the road priority. The chosen yielding vehicle will brake until the collision prediction clears.
- *Actions.* The ego vehicle in the intersection scenario is designed to operate by selecting one from a finite set of actions $A = \{slower,\ no\text{-}operation,\ faster\}$, where the vehicle can choose to slow down, maintain a constant speed or accelerate.
- *Rewards.* The reward of the ego vehicle for the intersection scenario is designed as follows: the ego vehicle receives a reward of 1 if it drives at the maximum speed or if it arrives at the destination and receives a reward of -5 if a collision

occurs. This reward design encourages the ego vehicle to cross the intersection and arrive at the destination as soon as possible while simultaneously avoiding collisions.

Roundabout Scenario

- *Settings.* The scene of the roundabout scenario is composed of a two-lane roundabout with four entrances/exits (north, south, east, and west), and the roads are occupied by the ego vehicle and several surrounding vehicles. The positions, velocities and destinations of the other vehicles are initialised randomly. The ego vehicle drives from 125 metres south to the roundabout, and the task of the ego vehicle is to cross the roundabout and take the second exit (north). The goal is accomplished if the ego vehicle successfully takes the second exit at the roundabout and drives 10 metres or more from the exit point in 13 seconds. Please note that vehicles in the roundabout have the right of way. As presented in **Figure 1-B** for a showcase of the roundabout scenario, the goal of the ego vehicle is to exit the roundabout following the red arrow.
- *Observations.* The vehicle observations of the roundabout scenario are applied to the kinematics type as well. The ego vehicle observes and learns from a $V * F$ array, where V is the number of nearby vehicles and F is the size of the features observed. In our setting, the ego vehicle observes a feature set of size 7, $\{p, x, y, v_x, v_y, cos_h, sin_h\}$, where p represents whether a vehicle is in the row (whether a vehicle was observed), x and y describe the location of the vehicle, v_x and v_y denote the x and y axes of velocity, and cos_h and sin_h describe the heading direction of the vehicle.
- *Actions.* The actions of the ego vehicle in the roundabout scenario are selected from a finite set $A = \{$*lane_left, lane_right, idle, faster, slower*$\}$, implying that the vehicle can choose to change lanes to the left/right, maintain the same speed and lane, and accelerate or decelerate.
- *Rewards.* The reward of the ego vehicle for the roundabout scenario is arranged as follows: the ego vehicle receives a reward of 0.5 if it drives at the maximum speed and receives a reward of -1 if a collision occurs. Every time the vehicle changes lanes, it will be awarded -0.05. We also applied a success reward of 1 if the ego vehicle arrives at the destination. This reward design encourages the vehicle to cross the roundabout and arrive at the destination as soon as possible while simultaneously avoiding collisions and unnecessary lane changing.

Appendix C Experimental Additional Results

Additional Findings in Experiment 1 We observed significant room to improve the safety performance in the testing phase, as shown in **Table 1**, especially the collision rate of the DRL models in DQN. In the intersection scenario, we observed a high collision rate - more than 50% for all three baseline DRL models (56.88% for DQN, 65.51% for A2C, and 64.45% for PPO), suggesting

that the autonomous vehicle in the road traffic junction driving environment is far from safe. Furthermore, the collision rate in the roundabout scenario was lower (¡20%) than that in the intersection scenario, mainly due to the high freezing rate, which does not indeed contribute to safety concerns. Specifically, the freezing robot problem observed in the roundabout scenario was noted during the training phase of all baseline DRL models, especially for DQN, which accounted for more than half of the freezing episodes (51.31%). Considering the above performance results, we chose DQN to investigate further whether the proposed self-attention module could decrease the collision rate and alleviate the freezing robot problem in the intersection and roundabout driving scenarios.

In addition, we demonstrated a showcase in **Figure 2** about testing episodes using DQN in the intersection and roundabout scenarios to further clarify our findings. The goal of the ego vehicle is to turn left in the intersection scenario (**Figure 2-A**), and the training behaviours of the ego vehicle are as follows: (1) the ego vehicle began from the starting point, driving towards the intersection; (2) the ego vehicle reached the intersection entry, and some other vehicles came from the left and right. The surrounding vehicles on the left probably would not collide with the ego vehicle since it was turning right, but the surrounding vehicles from the right might collide with the ego vehicle; (3) the ego vehicle made the decision to turn left and collided with the incoming car from the right. For the roundabout scenario (**Figure 2-B**), the goal of the ego vehicle is to take the north (second) exit from the roundabout, and the training behaviours of the ego vehicle are as follows: (1) the ego vehicle began from the starting point, driving towards the roundabout, and there were other vehicles in the roundabout that might collide with the ego vehicle; (2) the ego vehicle stayed basically where it was until it seemed safe to enter the roundabout; (3) the roundabout was completely clear of vehicles, the ego vehicle finally moved forward before time was out, and the episode ended.

Additional Findings in Experiment 2 Attention-DQN effectively improved the safety of autonomous vehicles in intersection and roundabout driving scenarios. In the intersection scenario, translation as the attention mechanism helped the ego vehicle pay better attention to the other vehicles in traffic, thereby avoiding collisions and completing the task more efficaciously and successfully. In the roundabout scenario, attention-DQN showed a meagre freezing rate of ego vehicles of only 3.28% in the test phase, suggesting that the attention mechanism alleviates the problem of robot freezing because the ego vehicle is more confident in moving forward with supplementary attention information about the surrounding vehicles.

Appendix D Results Discussion

Based on our findings, Attention-DQN lowered the freezing rate of the ego vehicle significantly in the roundabout scenario, from above 50% in both the training and testing results to only 3.28% in testing and 15% in training. This translates

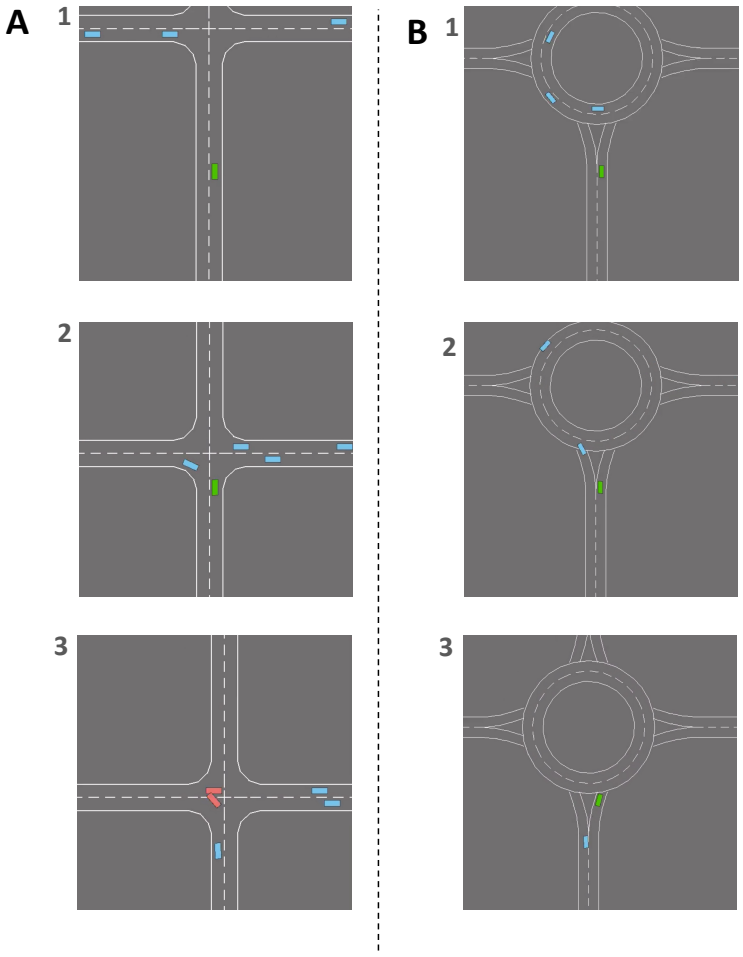

Fig. 2. A showcase of testing episodes using DQN in the (A) intersection and (B) roundabout scenarios.

to a significant improvement in the freezing robot problem in the roundabout scenario. One possible reason that Attention-DQN helped to achieve a low freezing rate is that the attention mechanism enables the ego vehicle to be more confident in moving forward by paying more attention to the surrounding vehicles. Interestingly, the attention mechanism did not have the same effect on the ego vehicle in the intersection scenario. The freezing rates of both the training and testing results after the applied attention mechanism were higher than those for the baseline algorithm. The collision rate decreased substantially in the intersection scenario, but not all of the decreased percentage was effectively transferred to the success rate, and some parts of the collision rate were transferred to the freezing rate. It is unclear in the current stage, but one observation during the

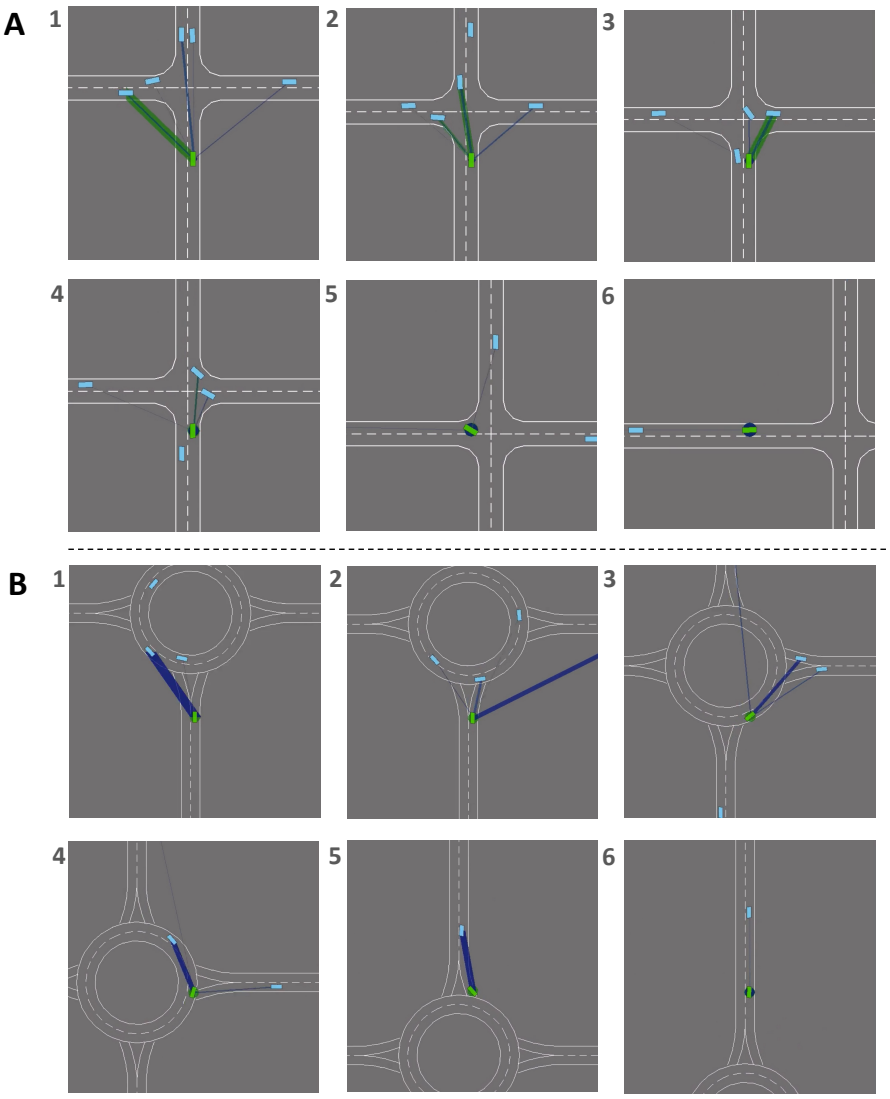

Fig. 3. A showcase of testing episodes using attention-DQN in the (A) intersection and (B) roundabout scenarios. The two attention heads are visualised as green and blue lines connecting the ego vehicle and other surrounding vehicles, and the width of a line corresponds to the weight of the attention.

experiments was that crashes between other vehicles occurred in the intersection and sometimes blocked the path of the ego vehicle. In this case, the ego vehicle without attention collided with the crashed vehicles, but with attention, it had to choose to freeze, as shown in **Figure 4**.

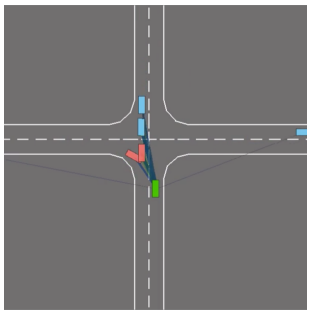

Fig. 4. A showcase of the frozen vehicle in the intersection scenario.

For DRL in a complex autonomous driving environment, our research still presented two limitations: The simulated driving environment might exhibit stochastic behaviour, including the fact that our test driving scenarios are complex scenarios populated with vehicles where one vehicle on the road may collide with another. During the experiments, we observed that crashes sometimes occurred between other vehicles and blocked the path of our ego vehicle. As a result, the ego vehicle could not move forward and could only wait until time ran out, but it was used to calculate into freezing rate, suggesting that the freezing rate might be influenced by this case. Moreover, our proposed model is still not entirely safe for application to real-world autonomous driving, so it is worth exploring the combination of optimisation measurements and more sensing information with DRL-trained agents for further safety improvement.

Appendix E Testing Demonstrations

In addition, we presented a showcase via video replays [1] and **Figure 3** of testing episodes using attention-DQN in the intersection and roundabout scenarios to further clarify our findings. Here, please note that "attention" is visualised as lines connecting the ego vehicle and other surrounding vehicles. The two different colours (green and blue) represent two attention heads, and the width of the line corresponds to the weight of the attention. The goal of the ego vehicle is to turn left in the intersection scenario (**Figure 3-A**), and the training attention behaviours of the ego vehicle are as follows: (1) the ego vehicle paid attention to the surrounding vehicles from the left, front, and right directions; (2) when

[1] https://anonymous.4open.science/r/Replay-videos-30DE/

the ego vehicle approached the intersection, the attention to incoming vehicles strengthened; (3) the surrounding vehicles from the left to right turn or from the front to left turn at the intersection were no longer threats to the ego vehicle, so the main attention switched to the surrounding vehicles from the right whose intention was not clear yet; (4) because the vehicle from the right showed intention to turn right and drove almost out of the intersection, the attention on it decreased; (5) the intersection was clear of all surrounding vehicles, so the attention returned to the ego vehicle itself when the ego vehicle crossed the intersection; (6) the ego vehicle drove in the destination lane with slight attention to the vehicle in front of it to maintain a safe distance. Moreover, in the roundabout scenario (**Figure 3-B**), the ego vehicle must learn to cross the roundabout and take the desired exit, and the training attention behaviours of the ego vehicle are as follows: (1) the ego vehicle paid substantial attention to the surrounding vehicles coming from the left that was likely to collide with the ego vehicle; (2) the ego vehicle waited until the surrounding vehicle from the left passed the entry point and switched its primary attention to the surrounding vehicles coming from the following entry; (3) the ego vehicle entered the roundabout and kept its attention on other surrounding vehicles; (4) the attention to the front vehicle strengthened since the ego vehicle was getting closer to this vehicle; (5) the ego vehicle exited the roundabout but still kept its attention on the front vehicle to maintain a reasonable distance; and (6) the attention switched back to the ego vehicle when it was sufficiently safe.

References

1. Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.
2. Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.
3. Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PMLR, 2016.
4. John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
5. Edouard Leurent. An environment for autonomous driving decision-making. *GitHub repository*, 2018.
6. Philip Polack, Florent Altché, Brigitte d'Andréa Novel, and Arnaud de La Fortelle. The kinematic bicycle model: A consistent model for planning feasible trajectories for autonomous vehicles? In *2017 IEEE Intelligent Vehicles Symposium*, pages 812–818. IEEE, 2017.