

# Digital Twin-based Cooperative Autonomous Driving in Smart Intersections: A Multi-Agent Reinforcement Learning Approach

Taoyuan Yu\*, Kui Wang\*, Zongdian Li\*, Tao Yu\*, Kei Sakaguchi\*, and Walid Saad†

\*Department of Electrical and Electronic Engineering, Institute of Science Tokyo, Tokyo, TYO 152-8550, Japan

†Bradley Department of Electrical and Computer Engineering, Virginia Tech, Arlington, VA 22203, USA

Email: {yuty, kuiw, lizd, yutao, sakaguchi}@mobile.ee.titech.ac.jp, walids@vt.edu

**Abstract**—Unsignalized intersections pose safety and efficiency challenges due to complex traffic flows and blind spots. In this paper, a digital twin (DT)-based cooperative driving system with roadside unit (RSU)-centric architecture is proposed for enhancing safety and efficiency at unsignalized intersections. The system leverages comprehensive bird-eye-view (BEV) perception to eliminate blind spots and employs a hybrid reinforcement learning (RL) framework combining offline pre-training with online fine-tuning. Specifically, driving policies are initially trained using conservative Q-learning (CQL) with behavior cloning (BC) on real datasets, then fine-tuned using multi-agent proximal policy optimization (MAPPO) with self-attention mechanisms to handle dynamic multi-agent coordination. The RSU implements real-time commands via vehicle-to-infrastructure (V2I) communications. Experimental results show that the proposed method yields failure rates below 0.03% coordinating up to three connected autonomous vehicles (CAVs), significantly outperforming traditional methods. In addition, the system exhibits sub-linear computational scaling with inference times under 40 ms. Furthermore, it demonstrates robust generalization across diverse unsignalized intersection scenarios, indicating its practicality and readiness for real-world deployment.

**Index Terms**—Digital Twin, Cooperative Driving, Intelligent Transportation System, Generative AI Models, Blind Spot Elimination.

## I. INTRODUCTION

Intersection management remains a critical bottleneck in intelligent transportation systems (ITS) due to intersection complexity and uncertainty [1]. According to the Federal Highway Administration (FHWA) and National Highway Traffic Safety Administration (NHTSA), intersection-related fatalities constitute a significant portion of traffic accident deaths, with unsignalized intersections accounting for 68% in 2024 [2], [3]. Blind spots and unclear interaction protocols make unsignalized intersections particularly dangerous. To address these challenges, the concept of a digital twin (DT) provides a promising solution by creating real-time virtual replicas of physical intersections, providing global perception and intelligent coordination beyond the limited sensing capabilities of individual vehicles [4], [5].

Mixed-traffic scenarios involving autonomous vehicles (AVs) and human-driven vehicles (HDVs) are becoming increasingly common thereby increasing the complexity of coordination among traffic participants. Vehicle-to-everything

(V2X) communications technologies, including vehicle-to-vehicle (V2V), vehicle-to-infrastructure (V2I), vehicle-to-pedestrian (V2P), and vehicle-to-network (V2N), can help enhance traffic safety and efficiency [6], [7]. Among these, V2I communications play a central role in DT systems as they allow real-time synchronization between physical vehicles and roadside units (RSUs), thereby supporting cooperative driving strategies and transforming traditional intersection infrastructure into intelligent control centers [8].

Leveraging V2I communications, DT systems have been applied to intersection management with various architectures. For example, in [9] and [10], the authors developed RSU-based DTs for continuous traffic monitoring and real-time analysis through cloud computing. However, these methods focus on general traffic monitoring and basic perception enhancement, without addressing blind spots and occlusions at intersections. The work in [11] addresses intersection occlusions but the solution of [11] is limited by local vehicle sensing and cannot achieve complete blind spot elimination. Despite the potential of DT technology to provide complete environmental awareness, current implementations [9]–[11] neither eliminate blind spots through global bird-eye-view (BEV) nor support cooperative driving strategies in occluded areas.

To leverage comprehensive DT perception, various intersection coordination algorithms have been explored in [12]–[16]. Traditional methods use optimization and game-theoretic algorithms to manage traffic flow [12], [13], but lack adaptability in dynamic environments. Multi-agent reinforcement learning (MARL) has been proposed as an effective solution for flexible and scalable coordination under partial observability [14]–[16]. Recent works further improve MARL by incorporating self-attention mechanisms to enhance inter-agent communication and decision-making. However, most MARL models adopt uniform policies across agents and fail to model diverse driving intents such as left-turn, straight, or right-turn maneuvers. In addition, these models are rarely evaluated under various vehicle densities, making their robustness in dynamic traffic conditions uncertain. Critically, current MARL methods fail to exploit DT’s global perception for blind-spot elimination, leaving a gap in ITS research.

The main contribution of this paper is to address the above limitations by developing a novel DT-based cooperative driv-

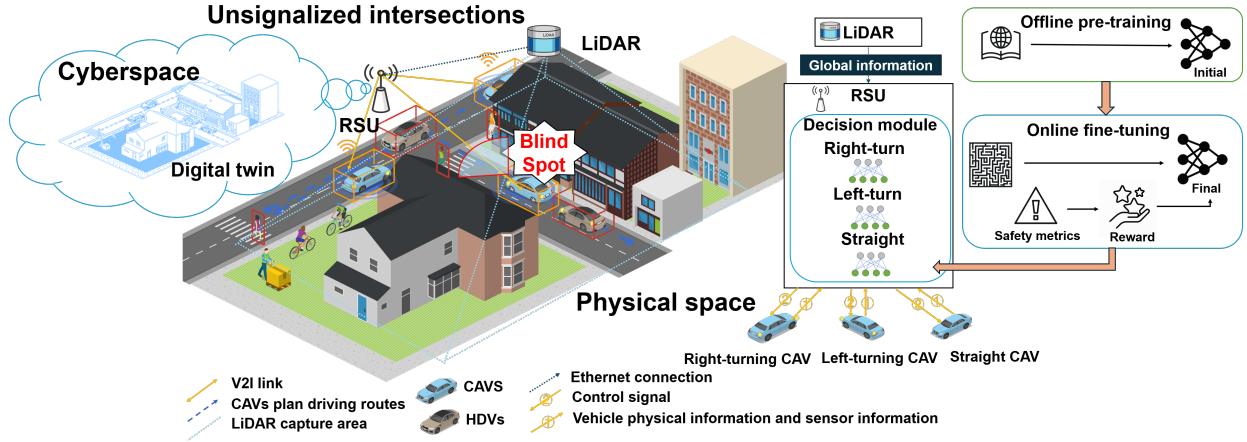


Fig. 1. High-level architecture of the DT cooperative system

ing system for unsignalized intersections. The system leverages RSU-mounted LiDAR to construct comprehensive BEV perception to eliminate blind spots, creating a real-time digital replica of the intersection environment. Our method leverages a centralized MARL decision module with role-specific policy networks and self-attention mechanisms. This method allows robust cooperative driving for various numbers of vehicles. Through a hybrid learning framework combining offline pre-training with online fine-tuning, the system develops decision-making capabilities that can be effectively deployed in real-world scenarios. This design achieves significant improvements in blind spot elimination, system adaptability, and traffic efficiency. In summary, our key contributions include:

- We develop a DT-based MARL framework eliminating blind spots via RSU global perception at unsignalized intersections.
- We introduce role-specific policy networks with self-attention mechanisms to enable adaptive coordination among connected autonomous vehicles (CAVs).
- We propose a hybrid offline-online reinforcement learning method to ensure robust and efficient policy learning.
- We conduct extensive experiments demonstrating system effectiveness and generalization across diverse scenarios.

The rest of this paper is organized as follows: Section II presents the DT system architecture. Section III details the proposed algorithm. Section IV discusses experimental results. Section V concludes the paper and outlines future work.

## II. RSU-CAVs COOPERATIVE SYSTEM

We consider the DT-based cooperative driving system architecture is shown in Fig. 1. The system establishes real-time synchronization between the physical intersection and its DT in cyberspace. RSU-mounted LiDAR sensors provide comprehensive BEV perception which, in turn, can help eliminate blind spots for global traffic monitoring. In contrast to traditional vehicle-centric methods focused on individual vehicles, this DT-based system facilitates centralized decision-making for multiple CAVs at unsignalized intersections. The

RSU's global perception overcomes individual vehicle sensor limitations, enabling a cooperative driving strategy designed to minimize potential conflicts and maximize intersection throughput.

To effectively manage the intersection's complexities and uncertainties, the RSU employs decision-making policies developed through a two-stage learning method. Given the dynamic and partially observable nature of traffic environments, RL provides a framework for sequential decision-making under uncertainty, modeled as a partially observable Markov decision process (POMDP). The training process begins with offline RL on real-world traffic dataset to establish foundational driving strategies, followed by online RL in simulated environments to enhance adaptability and robustness. This hybrid paradigm ensures that the resulting policies can handle diverse traffic scenarios while maintaining safety constraints. Once deployed, the RSU leverages these trained policies within its DT system to make real-time decisions, minimizing onboard compute requirements for CAVs while ensuring low-latency response [17].

To address the challenges of blind spots, and limited onboard sensing at unsignalized intersections, we introduce a DT-based cooperative system. As CAVs approach the intersection, they are simultaneously represented in the DT through V2I communications. The DT maintains real-time synchronization between physical vehicles and their digital counterparts, enabling the RSU to make decisions based on complete traffic state information. Using real-time data, the RSU determines each vehicle's driving role within the DT. Subsequently, the RSU leverages pre-loaded role-based strategy networks in the centralized decision module to compute control signals. These signals are transmitted in real-time to the corresponding CAVs through V2I communications. Concurrently, the DT continuously monitors traffic conditions, including the states and predicted movements of all traffic participants, traffic flow smoothness, and abnormal situations. This synchronization between physical space and cyberspace provides necessary real-time inputs for the decision networks and facilitates

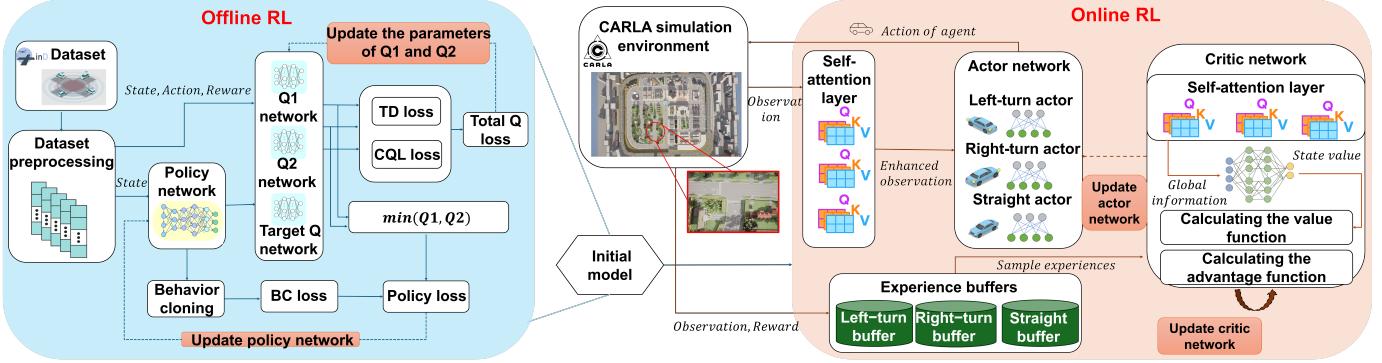


Fig. 2. Offline-online hybrid RL algorithm framework design

performance evaluation.

### III. HYBRID REINFORCEMENT LEARNING FRAMEWORK

As shown in Fig. 2, we propose a DT-based two-stage learning framework to develop cooperative driving strategies for unsignalized intersections. This method first employs offline pre-training on collected datasets, using offline RL to acquire foundational driving skills and traffic priors. Subsequently, online fine-tuning within the CARLA simulator [18] allows agents to adapt to dynamic environments. This hybrid method combines the safety of offline RL with the adaptability of online RL, ensuring that the trained model can achieve real-time decision-making within the RSU's DT system. This section details the methods for offline pre-training and online fine-tuning.

#### A. Observation Space

At each time step  $t$ , the state space  $s(t)$  encompasses all traffic participants monitored by the RSU. The RSU uses global information to construct individual observation vectors  $\mathbf{o}(t)$  for each CAV, capturing partially observable and potentially noisy representations:

$$\mathbf{o}(t) = [\mathbf{o}_{\text{core}}, \mathbf{o}_{\text{veh}}, \mathbf{o}_{\text{ped}}, \mathbf{o}_{\text{role}}, \mathbf{o}_{\text{ctx}}], \quad (1)$$

where  $\mathbf{o}_{\text{core}}$  includes ego-vehicle speed, global position, heading angle, and junction occupancy;  $\mathbf{o}_{\text{veh}}$  includes relative positions and velocities of nearby vehicles;  $\mathbf{o}_{\text{ped}}$  represents pedestrian detection, distance, and angle;  $\mathbf{o}_{\text{role}}$  encodes the agent's driving role; and  $\mathbf{o}_{\text{ctx}}$  contains scenario identifiers.

#### B. Action Space

We define a unified two-dimensional continuous action space  $\mathcal{A}$  for the vehicle. It is structured as:

$$\mathbf{a}(t) = [\mathbf{a}_{\text{acc}}, \mathbf{a}_{\text{steer}}] \in \mathbb{R}^2 \quad (2)$$

where  $\mathbf{a}_{\text{acc}}$  is longitudinal acceleration and  $\mathbf{a}_{\text{steer}}$  is steering angular velocity. During offline pre-training, actions are estimated from consecutive state transitions, as ground-truth controls are unavailable. During online fine-tuning, actions are directly predicted by the policy network.

#### C. Reward Function

To enable cooperative driving, we design a structured reward function  $\mathcal{R}_{\text{online}}(s(t), \mathbf{a}(t), s(t+1))$  to translate high-level objectives into real-time feedback. The overall reward  $r(t)$  is defined as:

$$r(t) = \sum w_k r_k(s(t), \mathbf{a}(t), s(t+1)), \quad (3)$$

where  $r_i$  represents individual reward components and  $w_i$  captures the corresponding weight. The reward terms include:

$$r_i \in \{r_{\text{safety}}, r_{\text{eff}}, r_{\text{comfort}}, r_{\text{task}}, r_{\text{yield}}, r_{\text{coop}}, r_{\text{penalty}}\} \quad (4)$$

where  $r_{\text{safety}}$  penalizes hazardous behaviors based on metrics like minimum time-to-collision (TTC);  $r_{\text{eff}}$  encourages speed compatible with traffic flow;  $r_{\text{comfort}}$  penalizes large acceleration changes;  $r_{\text{task}}$  rewards agents reaching navigation targets cooperatively;  $r_{\text{yield}}$  and  $r_{\text{coop}}$  reward compliance with traffic rules and cooperation; and  $r_{\text{penalty}}$  severely penalizes collisions or timeouts. Each term is scaled by its corresponding weight  $w_k$ , where  $w_{\text{safety}}$  and  $w_{\text{penalty}}$  are typically assigned larger values due to their critical importance.

#### D. Offline Pre-training: Networks and Algorithm

The primary goal of offline pre-training is to provide high-quality initialization for online fine-tuning. The model is trained independently for each driving role using subsets of the InD dataset [19], partitioned based on vehicle intentions.

For each subset, we employ an offline RL algorithm combining conservative Q-learning (CQL) and behavior cloning (BC) [20], [21] in an actor-critic framework. The critic uses twin Q-networks  $Q_{\theta_{i,1}}, Q_{\theta_{i,2}}$  with target networks to stabilize learning and reduce overestimation, optimized as:

$$L_Q(\theta_{i,j}) = \mathbb{E}_{(\mathbf{o}, \mathbf{a}, r, \mathbf{o}') \sim \mathcal{D}_{\text{role}=i}} \left[ \frac{1}{2} (Q_{\theta_{i,j}}(\mathbf{o}, \mathbf{a}) - y)^2 \right] + \alpha_{\text{CQL}} L_{\text{CQL\_reg}}(\theta_{i,j}) \quad (5)$$

Here,  $y = r + \gamma(1-d) \min_j Q_{\theta'_{i,j}}(\mathbf{o}', \pi_{\phi_i}(\mathbf{o}'))$  is the temporal difference (TD) target.

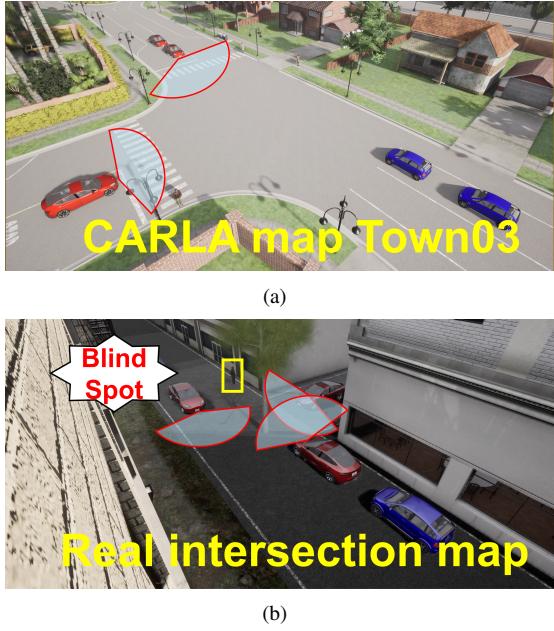


Fig. 3. Experimental scenario and generalization scenario settings (a) CARLA example map, (b) Real intersection map

The policy network  $\pi_{\phi_i}$  minimizes BC loss and maximizes conservative Q-values:

$$L_{\pi}(\phi_i) = \mathbb{E}_{\mathbf{o} \sim \mathcal{D}_{\text{role}=i}} \left[ - \min_{j=1,2} Q_{\theta_{i,j}}(\mathbf{o}, \pi_{\phi_i}(\mathbf{o})) \right] + \lambda_{\text{BC}} \mathbb{E}_{(\mathbf{o}, \mathbf{a}) \sim \mathcal{D}_{\text{role}=i}} [\|\pi_{\phi_i}(\mathbf{o}) - \mathbf{a}\|^2] \quad (6)$$

where  $\alpha_{\text{CQL}}$  and  $\lambda_{\text{BC}}$  denote hyperparameters controlling the strength of CQL regularization and BC imitation.

Role-specific actor  $\pi_{\phi_{\text{role}}}$  and critic  $Q_{\theta_{\text{role}}}$  networks are implemented as multi-layer perceptrons (MLPs). Self-attention is omitted at this stage to ensure robust training stability. The resulting pre-trained weight are reused during online fine-tuning to improve performance and accelerate adaptation.

#### E. Online Fine-tuning: Networks and Algorithm

The online fine-tuning adopts multi-agent proximal policy optimization (MAPPO) [22], integrating role-specific networks ( $\pi_{\phi_{\text{left}}}, \pi_{\phi_{\text{straight}}}, \pi_{\phi_{\text{right}}}$ ) with a shared critic network  $V_{\psi}$ .

To capture dynamic interactions, we augment both actor and critic networks with multi-head self-attention (MHSA). MHSA allows the model to jointly attend to information from different representation subspaces at different positions. The scaled dot-product attention is defined as:

$$\mathbf{A}(Q, K, V) = \text{softmax} \left( \frac{QK^{\top}}{\sqrt{d_k}} \right) V \quad (7)$$

where  $Q$ ,  $K$ , and  $V$  denote the query, key, and value matrices. MHSA computes multiple attention heads in parallel, and concatenates their outputs to form the final embedding, capturing dependencies among observation features.

Online learning proceeds via an interact-learn loop. Agents generate trajectories:

$$\tau = \{(\mathbf{o}_t, \mathbf{a}_t, r_{t+1}, V_{\psi}(\mathbf{o}_t), \log \pi_{\phi_{\text{role}}}(\mathbf{a}_t | \mathbf{o}_t))\}_{t=0}^T \quad (8)$$

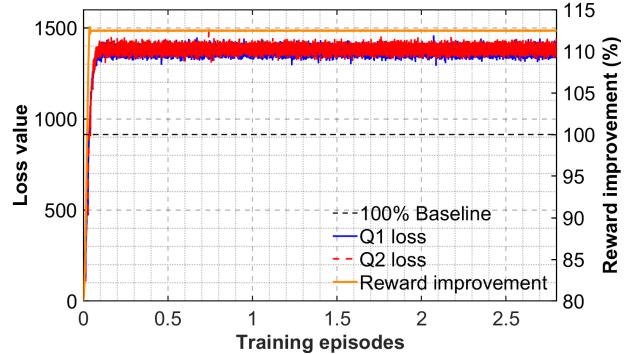


Fig. 4. Offline pre-training results

Advantage estimates  $\hat{A}_t^{\text{GAE}}$  and returns  $\hat{R}_t$  are computed using generalized advantage estimation (GAE), based on temporal-difference (TD) errors  $\delta_t$  calculated from critic values as:

$$\delta_t = r_{t+1} + \gamma V_{\psi}(\mathbf{o}_{t+1}) - V_{\psi}(\mathbf{o}_t) \quad (9)$$

Prioritized experience replay (PER) samples transitions based on priorities proportional to absolute TD errors, with importance sampling (IS) weight correcting sampling bias:

$$w_t = \left( \frac{1}{B \cdot P(t)} \right)^{\beta} \quad (10)$$

where  $B$  is the replay buffer size, and  $\beta$  controls IS correction strength.

Each role-specific actor  $\pi_{\phi_{\text{role}}}$  is trained using the following weighted objective, which includes the PPO clipped surrogate loss and an entropy bonus  $S[\cdot]$ :

$$L^{\text{CLIP+S}}(\phi_{\text{role}}) = \mathbb{E}_{t \sim \text{PER}} \left[ w_t \left( -L_t^{\text{CLIP}}(\phi_{\text{role}}) - c_2 \cdot S[\pi_{\phi_{\text{role}}}](\mathbf{o}_t) \right) \right] \quad (11)$$

The PPO surrogate loss  $L_t^{\text{CLIP}}$  is defined as:

$$L_t^{\text{CLIP}} = \min \left( r_t \hat{A}_t, \text{clip}(r_t, 1 - \epsilon, 1 + \epsilon) \hat{A}_t \right) \quad (12)$$

where  $\epsilon$  is the PPO clipping hyperparameter, and  $r_t$  represents the probability ratio between current and old policies.

## IV. EXPERIMENTS AND ANALYSIS

Experiments were conducted in synchronous mode using the CARLA simulator with Unreal Engine. The main test scenario is an unsignalized intersection in Town03, as shown in Fig. 3. In each episode, our system controls 1 to 3 CAVs (red), while background vehicles (blue) are controlled by CARLA's Traffic Manager. Pedestrians are added to simulate realistic urban conditions. For generalization evaluation, we deploy the model on a real intersection map based on the Institute of Science Tokyo campus. The RSU maintains a global state via BEV perception and uses the fine-tuned decision model to compute control commands, which are sent to CAVs through simulated V2I communication.

TABLE I  
PERFORMANCE COMPARISON SUMMARY

approach / Scenario	Failure rate (%)	Avg. time (s)
Ours (1 Agent, Town03)	0.01	5.52
Ours (2 Agent, Town03)	0.03	5.49
Ours (3 Agent, Town03)	0.02	5.25
Autoware (1 Agent, Town03)	5.31	5.77
Ours (3 Agent, Real Map)	0.02	5.15

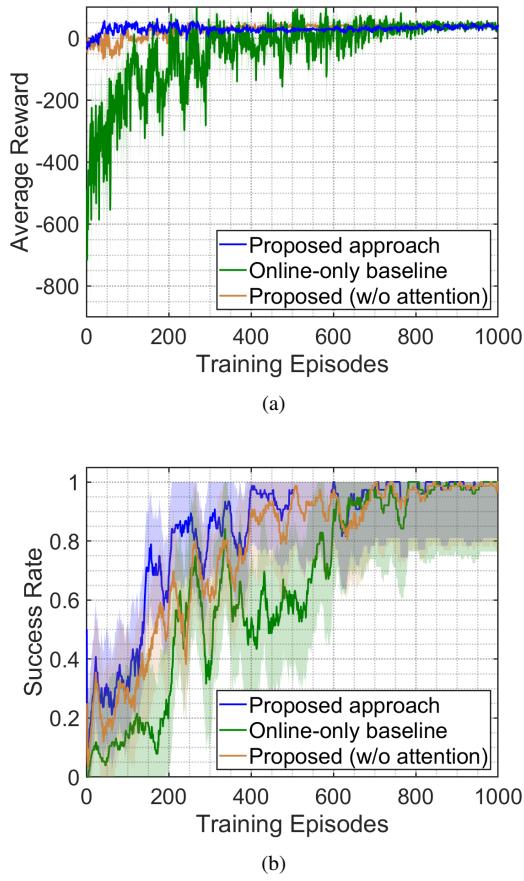


Fig. 5. Comparison of training performance across different approaches, (a) reward, (b) success rate.

#### A. Baselines and Evaluation Metrics

To assess the contribution of each component, we compare our model with several baselines. First, two ablated variants are considered: (1) an online-only MAPPO baseline trained directly; and (2) a variant with offline pre-training but without self-attention or role-specific policies. Both share the same architecture and hyperparameters as full model, isolating the effects of offline pre-training and self-attention respectively. Second, we include Autoware Universe [23], a rule-based autonomous driving stack, configured to control a single vehicle. All approaches are evaluated in terms of convergence speed, failure rate, and average travel time.

#### B. Offline Pre-training Results

The offline pre-training phase aims to extract driving priors from the InD dataset to initialize the model for online fine-tuning. Fig. 4 shows the Q1/Q2 losses and reward improvement over training. The steadily converging losses indicate stable learning of state-action values, while the reward metric stabilizes around 112%, surpassing the 100% baseline. This confirms that the policy learned via CQL combined with BC not only imitates but also improves upon average dataset behavior, offering a strong initialization for the online stage.

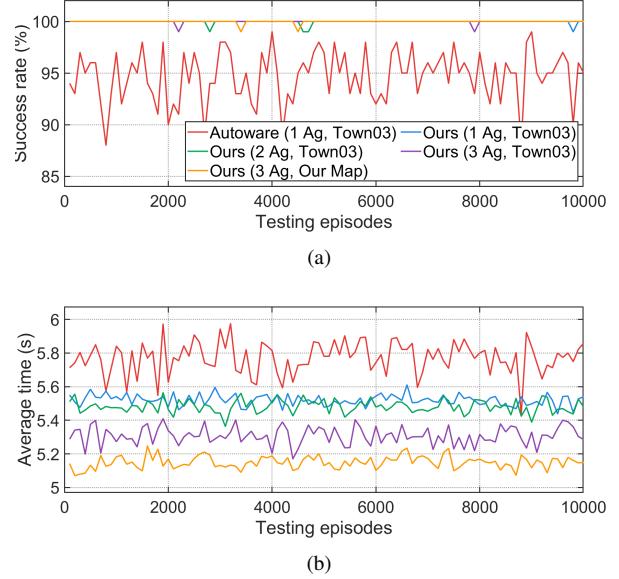


Fig. 6. Final model performance evaluation results (a) success rate by testing episodes, (b) average travel time by testing episodes

#### C. Online Training Results

Fig. 5 presents the training convergence of our proposed model alongside two ablated variants. The full model consistently outperforms all baselines. It reaches stable performance within approximately 250 episodes, whereas the online-only baseline requires over 800 episodes to converge. The ablated variant without self-attention and role-specific policies, despite benefiting from offline pre-training, converges after about 500 episodes. This comparison reveals that both components are essential for achieving optimal performance. Offline pre-training accelerates learning and improves initial performance, while self-attention and role-specific policies further enhance and sustain multi-agent coordination effectiveness. These results confirm that our hybrid method combines the safety of offline RL with the adaptability needed for complex multi-agent coordination at intersections.

#### D. Performance Evaluation and Generalization Analysis

We evaluated the model through 10,000 performance test episodes on both Town03 intersection and the real intersection map, comparing it against baselines. Key performance indicators are summarized in Fig. 6 and Table I, where failure rates represent the percentage of episodes ending in collision or timeout. Our model demonstrates high safety and reliability across all test scenarios on Town03. When controlling single

vehicle, it achieves 0.01% failure rate, outperforming the 5.31% failure rate of the Autoware baseline. Notably, as coordination complexity increases, our system does not exhibit a marked decline in performance. Specifically, the failure rate is 0.03% in the two-vehicle scenario and 0.02% in the three-vehicle scenario. The combination of the BEV perspective and the self-attention mechanism contributes to this robustness, demonstrating our model's effectiveness in handling complex multi-agent cooperative tasks.

In addition, the performance advantage of our model is also demonstrated in terms of traffic efficiency. The average travel time in the single-vehicle scenario was 5.52 seconds, compared to the 5.77 seconds by the Autoware. As the number of controlled vehicles increased, the average travel time slightly decreases, indicating that multi-agents coordinated effectively, establishing efficient cooperative driving strategies that actually improve intersection throughput with more CAVs.

For generalization, the three-vehicle model trained on Town03 is deployed on the real intersection map. It achieves a failure rate of 0.02% and an average travel time of 5.15 seconds in this new environment. This suggests that the BEV effectively eliminates the impact of individual vehicle blind spots. This result validates the excellent generalization capability of our model and provides a solid foundation for the practical application of the approach.

To further validate the system's deployability, we evaluate its computational performance across different coordination scenarios. Experiments are conducted on an NVIDIA RTX 4070 Ti GPU at a 10 Hz control frequency. The average inference times are 23.7 ms for single vehicle, 31.4 ms for two vehicles, 38.2 ms for three vehicles, and a maximum inference time of 42.6 ms across all tests. This sub-linear scaling confirms efficient multi-agent processing. Even in worst cases, inference time stays well within the 100 ms control interval, leaving sufficient margin for V2I communications and safety checks, validating the system's real-time deployability.

## V. CONCLUSION AND FUTURE WORKS

In this paper, we have proposed a DT-based cooperative driving system with RSU-centric architecture at unsignalized intersections. The system leverages BEV perception to eliminate blind spots and employs a hybrid reinforcement learning algorithm for robust multi-agent cooperative driving strategies. We developed role-specific policies and validated the system in diverse scenarios, achieving a 0.03% failure rate and sub-40ms inference time for up to three CAVs. Future primary work involves the proof-of-concept (PoC) experiments to fully validate the system performance in the real world.

## REFERENCES

- [1] K. Chu, A. Lam and V. Li, "Traffic Signal Control Using End-to-End Off-Policy Deep Reinforcement Learning," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 7184-7195, 2022.
- [2] U.S. Department of Transportation. [Online]. Available: <https://highways.dot.gov/sites/fhwa.dot.gov/files/2024-08>
- [3] National Highway Traffic Safety Administration. [Online]. Available: <https://www.safercar.gov/press-releases/nhtsa-2023-traffic-fatalities-estimate-april-2024>
- [4] K. Wang et al., "Smart Mobility Digital Twin Based Automated Vehicle Navigation System: A Proof of Concept," in *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 3, pp. 4348-4361, 2024.
- [5] O. Hashash, C. Chacour, W. Saad, T. Yu, K. Sakaguchi and M. Debbah, "The Seven Worlds and Experiences of the Wireless Metaverse: Challenges and Opportunities," in *IEEE Communications Magazine*, vol. 63, no. 2, pp. 120-127, 2025.
- [6] K. Wang, C. She, Z. Li, T. Yu, Y. Li, and K. Sakaguchi, "Roadside Units Assisted Localized Automated Vehicle Maneuvering: An Offline Reinforcement Learning Approach," in *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1709-1715, 2024.
- [7] Z. Li, K. Wang, T. Yu, and K. Sakaguchi, "Het-SDVN: SDN-Based Radio Resource Management of Heterogeneous V2X for Cooperative Perception," *IEEE Access*, vol. 11, pp. 76255-76268, 2023.
- [8] D. Suo, B. Mo, J. zhao, and S. E. Sarma, "Proof of Travel for Trust-Based Data Validation in V2I Communication," *IEEE Internet of Things Journal*, vol. 10, no. 11, pp. 9565-9584, 2023.
- [9] Y. Cui, H. Xu, J. Wu, Y. Sun and J. Zhao, "Automatic Vehicle Tracking With Roadside LiDAR Data for the Connected-Vehicles System," *IEEE Intelligent Systems*, vol. 34, no. 3, pp. 44-51, 2019.
- [10] L. Wang et al., "Multi-Modal 3D Object Detection in Autonomous Driving: A Survey and Taxonomy," in *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 7, pp. 3781-3798, 2023.
- [11] K. Moller, R. Trauth, and J. Betz, "Overcoming Blind Spots: Occlusion Considerations for Improved Autonomous Driving Safety," in *2024 IEEE Intelligent Vehicles Symposium (IV)*, pp. 819-826, 2024.
- [12] Y. Zhu, Z. He and G. Li, "A bi-Hierarchical Game-Theoretic Approach for Network-Wide Traffic Signal Control Using Trip-Based Data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 15408-15419, 2022.
- [13] M. Gallo, "Combined Optimisation of Traffic Light Control Parameters and Autonomous Vehicle Routes," *Smart Cities*, no. 3, pp. 1060-1088, 2024.
- [14] Y. Shi, H. Dong, C. He, Y. Chen and Z. Song, "Mixed Vehicle Platoon Forming: A Multiagent Reinforcement Learning Approach," *IEEE Internet of Things Journal*, vol. 12, no. 11, pp. 16886-16898, 2025.
- [15] S. Iqbal and F. Sha, "Actor-Attention-Critic for Multi-Agent Reinforcement Learning," *arXiv preprint arXiv:1810.02912*, 2019.
- [16] R. Younas, H. M. Raza Ur Rehman, I. Lee, B. W. On, S. Yi and G. S. Choi, "SA-MARL: Novel Self-Attention-Based Multi-Agent Reinforcement Learning With Stochastic Gradient Descent," in *IEEE Access*, vol. 13, pp. 35674-35687, 2025.
- [17] K. Wang, T. Yu, Z. Li, K. Sakaguchi, O. Hashash, and W. Saad, "Digital Twins for Autonomous Driving: A Comprehensive Implementation and Demonstration," in *2024 International Conference on Information Networking (ICOIN)*, pp. 452-457, 2024.
- [18] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An Open Urban Driving Simulator," in *Proceedings of the 1st Annual Conference on Robot Learning*, pp. 1-16, 2017.
- [19] J. Bock, R. Krajewski, T. Moers, S. Runde, L. Vater and L. Eckstein, "The inD Dataset: A Drone Dataset of Naturalistic Road User Trajectories at German Intersections," in *2020 IEEE Intelligent Vehicles Symposium (IV)*, p. 1929-1934, 2020.
- [20] A. Kumar, A. Zhou, G. Tucker, and S. Levine, "Conservative Q-Learning for Offline Reinforcement Learning," *arXiv preprint arXiv:2006.04779*, 2020.
- [21] D. A. Pomerleau, "ALVINN: An Autonomous Land Vehicle in a Neural Network," in *Advances in Neural Information Processing Systems*, 1988.
- [22] C. Yu, A. Veliu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu, "The Surprising Effectiveness of PPO in Cooperative, Multi-Agent Games," *arXiv preprint arXiv:2103.01955*, 2022.
- [23] Autoware. [Online]. Available: <https://autoware.org/>