2024 IEEE Intelligent Vehicles Symposium (IV)
June 2-5, 2024. Jeju Island, Korea

2024 IEEE Intelligent Vehicle Symposium (IV) | 979-8-3503-4881-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/IV55156.2024.10588863

Autonomous Intersection Management with Heterogeneous Vehicles: A Multi-Agent Reinforcement Learning Approach

Kaixin Chen, Bing Li, Rongqing Zhang, *Member, IEEE,* and Xiang Cheng, *Fellow, IEEE*

***Abstract*—While autonomous intersection management (AIM) emerges to facilitate signal-free scheduling for connected and autonomous vehicles (CAVs), several challenges arise for planning secure and swift trajectories. Existing works mainly focus on addressing the challenge of multi-CAV interaction complexity. In this context, multi-agent reinforcement learning-based (MARL) methods exhibit higher scalability and efficiency compared with other traditional methods. However, current AIM methods omit discussions on the practical challenge of CAV heterogeneity. As CAVs exhibit different dynamics features and perception capabilities, it is inappropriate to adapt identical control schemes. Besides, existing MARL methods that lack heterogeneity adaptability may experience a performance decline. In response, this paper exploits MARL to model the decision-making process among CAVs and proposes a novel heterogeneous-agent attention gated trust region policy optimization (HAG-TRPO) method. The proposed method can accomplish more effective and efficient AIM with CAV discrepancies by applying a sequential update schema that boosts the algorithm adaptability for MARL tasks with agent-level heterogeneity. In addition, the proposed method utilizes the attention mechanism to intensify vehicular cognition on disordered ambience messages, as well as a gated recurrent unit for temporal comprehension on global status. Numerical experiments verify that our method results in CAVs passing at the intersection with fewer collisions and faster traffic flow, showing the superiority of our method over existing benchmarks in terms of both traffic safety and efficiency.**

I. Introduction

As an essential scenario in intelligent transportation systems, autonomous intersection management (AIM) enables efficient and automatic scheduling for connected and autonomous vehicles (CAVs) without rigid signal controls.

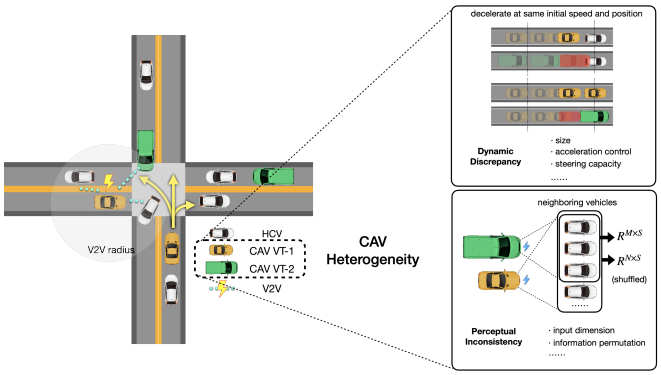

Fig. 1. An illustrated AIM scenario with CAV heterogeneity

However, it remains several difficulties for resolving vehicular decision-making in versatile intersection ambiences.

One of the essential challenges is the interaction complexity regarding multiple vehicles in the ever-changing intersection. To address this challenge, some previous works adopt rule-based methods like the first-come-first-serve protocol [1] and time-to-collision (TTC [2]) precaution rule [3] to dispatch CAVs under fixed logics, but the traffic efficiency of these methods is limited owing to the conservative nature of fixed rules [4]. Meanwhile, some other methods seek scheduling plans with minimized passing delays through optimization-based methods [5] or tree search techniques [6], but these methods encounter bottlenecks in computation speed and generalization when AIM involves large numbers of CAVs [7]. Recently, multi-agent reinforcement learning-based (MARL) techniques have been widely applied in the field of vehicular cooperation [8] [9]. Compared with the above methods, MARL is promising to resolve these defects by enabling CAVs to make rapid decisions with respect to the sophisticated ambience [10] [11] [12]. By efficient inter-vehicle cooperation, MARL-based methods can avoid collisions and minimize passing delays with high scalability.

In addition, CAV heterogeneity is another vital challenge in practical AIM problems. As shown in Fig. 1, due to the common variations of vehicle models in reality, CAVs usually manifest differences in dynamic features like width, length, acceleration capacity, etc. Moreover, CAVs also exhibit perceptual inconsistencies in observation dimensions and messa-

This work is supported in part by the National Key R&D Program of China under Grant 2021ZD0112700, in part by the National Natural Science Foundation of China under Grant 62271351, 62125101, 62201390, and 62341101, in part by the Natural Science Foundation of Shanghai under Grant 22ZR1463600, in part by the Chenguang Program of Shanghai Education Development Foundation and Shanghai Municipal Education Commission under Grant 21CGA24, in part by the New Cornerstone Science Foundation through the XPLORER PRIZE, and the Fundamental Research Funds for the Central Universities.

Kaixin Chen, Bing Li, and Rongqing Zhang are with the School of Software Engineering, Tongji University, Shanghai 200092, China (e-mail: 2333108@tongji.edu.cn; lizi@tongji.edu.cn; rongqingz@tongji.edu.cn).

Xiang Cheng is with the School of Electronics, Peking University, Beijing 100871, China (e-mail: xiangcheng@pku.edu.cn).

ge permutation. These heterogeneous factors amplify the difficulty of AIM, as the cooperative methods are required to deliver suitable control schemes for different types of CAVs under varying message contexts acquired through instant communications. Most recent works have discussed the scheduling of heterogeneous CAVs in traffic scenarios like on-ramp and highway [13] [14]. But to the authors' knowledge, no existing works explore the AIM under the more practical heterogeneous scenario. Current methods on AIM still follow the homogeneous assumption, wherein all CAVs are consistent without model distinctions. However, simply adapting identical control methods on diverse CAVs may cause adverse consequences.

Taking the heterogeneity of CAVs into consideration, in this work, we for the first time investigate the AIM problem in a more practical intersection scenario with multiple heterogeneous CAVs. To achieve efficient and effective decision-making in intricate ambiences, we employ MARL to achieve real-time AIM and model the decision-making process of each CAV as a partially observable Markov decision process (POMDP). However, despite the superiority of MARL solutions in complex real-time interactions, existing MARL methods [15] [16] may undergo a performance decline owing to the deficiency in structural adaptability in heterogeneous environments. In response, we propose a novel heterogeneous-agent attention gated trust region policy optimization (HAG-TRPO) method to achieve an effective cooperative decision-making among heterogeneous CAVs. Specifically, the proposed method employs a sequential update schema grounded in the multi-agent advantage decomposition lemma during training to assure heterogeneity elasticity. Additionally, it boosts vehicular cognition of complex ambiences and latent dependencies by utilizing attention-based observation aggregation impervious to shuffled input order, while comprehending potential historical traffic status with the gate recurrent unit (GRU). Extensive experiments are conducted in general AIM scenarios, including both homogeneous and heterogeneous cases. The numerical results demonstrate that compared with existing rule-based and MARL-based methods, our proposed method facilitates CAVs collaborations at the intersection with fewer collisions and faster traffic flow, indicating the superiority of our proposed method in both safety and efficiency.

The rest of the paper is organized as follows. Section II illustrates the AIM problem formulation regarding general traffic settings and vehicular kinematics. Section III proposes our MARL-based method for AIM and Section IV verifies the performance of our method with other benchmarks. Section V concludes this paper.

II. Problem Formulation

A. General Traffic Settings

Consider a four-direction unsignalized intersection depicted in Fig. 1. Various vehicles approach the intersection from any of the four driveway at distinct time steps, and their exit directions are randomly predefined from options including left turn, straight-ahead or right turn. In accordance with the vehicular control scheme, vehicles can be categorized into two types, including the CAVs controlled by standalone decision-making algorithms, and the human-driven connected vehicles (HCVs) with actions determined by the drivers' instant reactions. All vehicles are able to acquire their spatial coordinates and kinematics state, subsequently sharing these data with neighboring vehicles through V2V communications within a limited range. In this scenario, each CAV is required to leverage the aggregated information to maintain its ego trajectory while collaborating with adjacent vehicles for proactively addressing latent conflicts.

B. Vehicular Kinematics

The fundamental kinematics properties of each vehicle are calculated by the kinematic bicycle model [17]. For the CAVs piloted by the algorithm policy that outputs the high-level steering and acceleration action signal, two low-level controllers are implemented to obtain vehicular movements at the longitudinal and lateral level. The longitudinal controller utilizes a simple proportional relationship to manage the vehicle's acceleration:

$$a = \frac{1}{\tau_p}(v_r - v) \tag{1}$$

where a is the vehicle numerical acceleration, v, v_r are the vehicle velocity and the reference velocity, respectively, and $\frac{1}{\tau_p}$ is the controller proportional gain. The lateral controller conducts the vehicle's position control, which can be represented by:

$$\begin{aligned} v_{lat,r} &= -\frac{1}{\tau_{p,lat}}\Delta_{lat} \\ \Delta\psi_r &= \arcsin(\frac{v_{lat,r}}{v}) \end{aligned} \tag{2}$$

as well as the vehicle's heading control represented by:

$$\begin{aligned} \psi_r &= \psi_L + \Delta\psi_r \\ \dot{\psi_r} &= \frac{1}{\tau_{p,\psi}}(\psi_r - \psi) \\ \delta &= \arcsin\left(\frac{1}{2}\frac{l}{v}\dot{\psi_r}\right) \end{aligned} \tag{3}$$

where Δ_{lat} is the lateral position of the vehicle relative to the lane center-line, $v_{lat,r}$ is the lateral velocity command, $\Delta\psi_r$ is a heading variation to apply the lateral velocity command, ψ_L is the lane heading, ψ_r is the target heading to follow the lane heading and position, $\dot{\psi_r}$ is the yaw rate command, $\frac{1}{\tau_{p,lat}}$, $\frac{1}{\tau_{p,\psi}}$ are the position and heading control gains, respectively, and δ is the front wheels angle control.

For the HCVs, the intelligent driver model (IDM) [18] is applied for human-driver-simulated behavior control:

$$\begin{aligned} \dot{v} &= a\left[1 - (\frac{v}{v_0})^\delta - (\frac{d^*}{d})^2\right] \\ d^* &= d_0 + Tv + \frac{v\Delta v}{2\sqrt{ab}} \end{aligned} \tag{4}$$

where v is the vehicle velocity vector, d is the distance to its front vehicle, v_0 is the desired velocity, T is the desired time gap, d_0 is the jam distance, a, b are the maximum acceleration and deceleration, and δ is the velocity exponent.

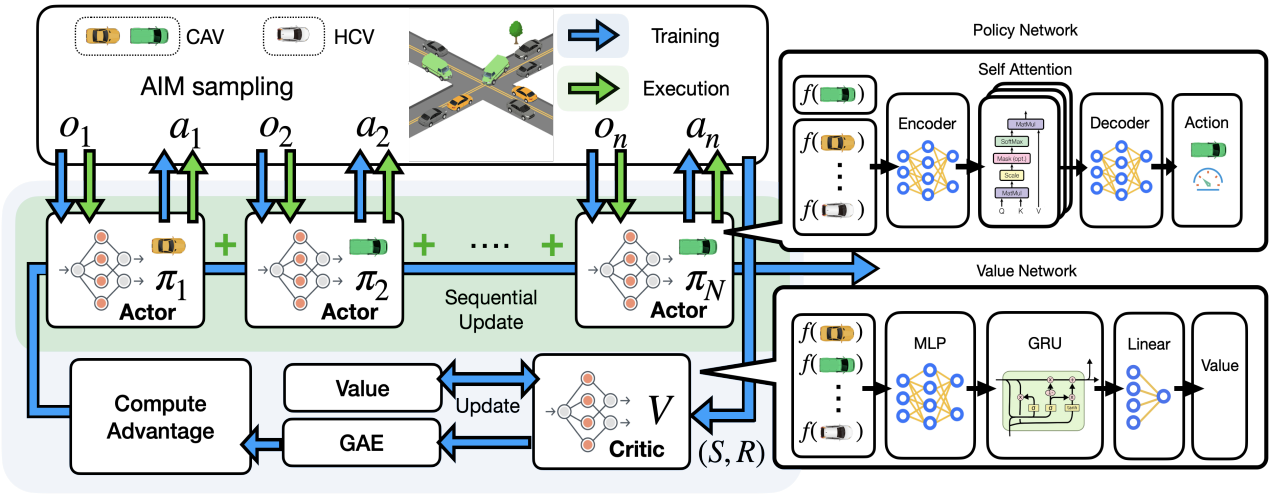

Fig. 2. The overview of HAG-TRPO cooperative AIM structure

TABLE I
VEHICLE MODELS AND COEFFICIENTS

Model Type	V-T1	V-T2
τ_p (s)	0.4	0.8
$\tau_{p,lat}$ (s)	0.4	0.8
$\tau_{p,\phi}$ (s)	0.2	0.4
vehicle length (m)	5.0	6.5
vehicle width (m)	1.8	2.4
max speed (m/s)	30	25
max observed vehicle	7	5

III. MULTI-AGENT REINFORCEMENT LEARNING-BASED METHOD FOR AIM

A. POMDP Problem Formulation

Due to the sequential nature of intersection decision-making and the demand of efficient cooperation, we model the AIM problem as a POMDP. At time step t, agent n receives a localized observation O_n with a reward signal r_n, subsequently executing an action a_n through its policy module π_n to interact with the environment. The elementary definition of the system is as follows:

1) Agent: Each CAV is designated as an agent in the system to pass the intersection. We consider two types of CAVs (named V-T1 and V-T2) that differ in dynamic property and perception input to discuss vehicular heterogeneity. The vehicular dynamics and perception coefficients regarding model V-T1 and V-T2 are displayed in Table I. Compared with V-T1, V-T2 exhibits larger size, bulkier speed and acceleration control, as well as relatively weaker perceptual capability.

2) Observation: During each time step, each CAV observes its pose and kinematic information through sensors and V2V messages transmitted by adjacent vehicles. The observation O_n^t of the CAV n at the time step t can be expressed as:

$$O_n^t = \{ty_n, loc_n^t, v_n^t, head_n^t, ty_{V_n}^t, loc_{V_n}^t, v_{V_n}^t, head_{V_n}^t\} \quad (5)$$

where ty_n is the integer descriptor of the vehicle model type of agent n, loc_n^t is the location coordinates at time step t, v_n^t is the velocity vector, $head_n^t$ denotes the vehicular heading angle, and V_n is a dynamic set representing the vehicles within the observation radius of agent n at time step t.

3) Action: Every CAV selects one action from the action set $\{cruise, slower, faster\}$, subsequently processed by low-level controllers to obtain the steering and acceleration.

4) Reward: the reward of CAV n at time step t is:

$$r_n^t = \eta_1 \frac{v_n^t - v_{min}}{v_{max} - v_{min}} + \eta_2 e_n^t \quad (6)$$

where e_n^t is a mutable signal (1 for arrival reward, -5 for collision penalty, and 0 for other cases). The reward encourages safe driving while sustaining relatively high speed.

B. Cooperative Strategy for AIM with Heterogeneous CAVs

1) Sequential Updating Structure: As general AIM involves cooperation between agents with diverse models, there necessitates a training schema with strong compatibility for variously-structured agent policies to process sampled messages while acquiring a decent performance. In response, we implement the multi-agent trust region optimization algorithm (TRPO) [19] shown in Fig. 2. It provides a sequential updating method during sampling, and promotes an MARL cooperative structure ensuring monotone improvement. The multi-agent TRPO method is theoretically supported by the multi-agent advantage decomposition lemma:

$$A_\pi^{i_{1:n}}(s, \boldsymbol{a}^{i_{1:n}}) = \sum_{m=1}^{n} A_\pi^{i_m}(s, \boldsymbol{a}^{i_{1:m-1}}, a^{i_m}) \quad (7)$$

where $i_{1:n}$ is the agent subset $\{i_1, i_2, ..., i_n\}$ in a specific order, $A_\pi^{i_{1:n}}$ is the advantage function taking the form of the generalized advantage estimation (GAE) [20], and $\boldsymbol{a}^{i_{1:n}}$ is the joint actions taken by agents from number 1 to n. After the buffer collects T steps of trajectory data at episode e, the advantage estimator of agent n can be calculated in a sequential manner:

$$\mathrm{M}^{\mathrm{i_{1:n}}}(\mathrm{s}, \boldsymbol{a}) = \begin{cases} \hat{A}_{s,\boldsymbol{a}}(s, \boldsymbol{a}), & n = 1 \\ \frac{\pi_{\theta_{e+1}}^{n-1}(a_t^{n-1}|s_t^{n-1})}{\pi_{\theta_e}^{n-1}(a_t^{n-1}|s_t^{n-1})} \mathrm{M}^{\mathrm{i_{1:n-1}}}(\mathrm{s}, \boldsymbol{a}) & n > 1 \end{cases} \quad (8)$$

where $\pi_{\theta_e^{n-1}}, \pi_{\theta_{e+1}^{n-1}}$ is the origin and updated policy of agent $n-1$, respectively. Then the gradient of agent n's policy $\boldsymbol{g}$ can be estimated by:

$$\hat{\boldsymbol{g}}_e^{i_n} = \frac{1}{B}\sum_{b=1}^{B}\sum_{t=1}^{T}\nabla_{\theta_e^{i_n}}\log\pi_{\theta_e^{i_n}}\left(a_t^{i_n}|o_t^{i_n}\right)\mathrm{M}^{\mathrm{i_{1:n}}}\left(\mathrm{s_t},\boldsymbol{a}_\mathrm{t}\right) \quad (9)$$

where B is the batch size. To conduct a TRPO step, the Hessian matrix of the average Kullback-Leibler (KL) divergence D_{KL} from the old policy $\boldsymbol{H}$ should be computed so as to obtain the update direction:

$$\hat{\boldsymbol{H}}_e^{i_n} = \nabla^2_{\theta^{i_n}}\frac{1}{BT}\sum_{b=1}^{B}\sum_{t=1}^{T}D_{KL}\left(\pi_{\theta_e^{i_n}}\left(\cdot|o_t^{i_n}\right),\pi_{\theta^{i_n}}\left(\cdot|o_t^{i_n}\right)\right) \quad (10)$$

The policy parameters of agent n is finally updated with:

$$\theta_{e+1}^{i_n} = \theta_e^{i_n} + \alpha^j\sqrt{\frac{2\delta}{\boldsymbol{g}_e^{i_n}\left(\hat{\boldsymbol{H}}_e^{i_n}\right)^{-1}\boldsymbol{g}_e^{i_n}}}\left(\hat{\boldsymbol{H}}_e^{i_n}\right)^{-1}\boldsymbol{g}_e^{i_n} \quad (11)$$

where α is a positive backtracking coefficient determined by line search, and δ is the threshold value of KL-divergence constraint. Meanwhile, The global critic network ϕ that outputs the value of current state V is updated by minimizing the optimization objective function. It is shown as:

$$\phi_{e+1} = \arg\min_{\phi}\frac{1}{BT}\sum_{b=1}^{B}\sum_{t=0}^{T}\left(V_\phi(s_t)-\hat{R}_t\right)^2 \quad (12)$$

2) Attention Mechanism for Policy Module: To enhance CAV's capability in perceiving latent intentions of varying nearby vehicles from disordered V2V messages, we employ self-attention mechanism to the actor network $\pi_{\theta^{i_n}}$ of each CAV. As displayed in Fig. 2, the attention layer takes in the observation embedding and trains the matrices of ego query W_Q, collective key W_K and collective value W_V. The weighted attention att_i of agent i can be obtained by:

$$att_i = \sum_{h=1}^{N_h}\mathrm{softmax}(\frac{(W_Q e_i)(W_K e_h)^T}{\sqrt{d_k}})h(W_V e_h) \quad (13)$$

where N_h is the number of attention heads, d_k is the embedding dimension, e_h is the embedding and $h(\cdot)$ is the activate function.

3) Temporal Module for Critic Network: As the potential conflicts among vehicles can be inferred from historical vehicular pose information sequentially, we implement a light gate recurrent unit (GRU) into the global critic network ϕ to capture the temporal correlation within vehicles efficiently.

IV. Simulation and Results

A. Scenario Setups

Based on the traffic settings illustrated in II-A and the vehicle settings illustrated in Table I, the AIM simulation is constructed utilizing an intersection gym-based simulator [21]. The scenarios are categorized into two types contingent upon the model configurations of the four CAV agents. In the agent-homogeneous AIM (abbreviated to HOM), all four CAVs are modeled as V-T1 consistently. Whereas for the agent-heterogeneous AIM (abbreviated to HET), two V-T1 agents and two V-T2 agents are employed. For both scenarios, the system randomly generates a set of HCVs, each with an intended destination, to pass through the intersection. The experiment parameters is showed in Table II.

B. Performance Benchmark

1) *TTC*: Rule-based baseline [2]. Assuming that all observed vehicles maintain their trajectories with constant speed and heading, each CAV predicts the duration of the upcoming collision, enabling them to decelerate when TTC falls below a safe threshold and speed up otherwise.
2) *C-MAPPO*: Multi-agent proximal policy optimization (MAPPO) under the cooperative centralized training and decentralized execution (CTDE) schema [12]. CAVs update the policy individually through PPO [**?**] under the guidance of the centralized value network during training. In the execution stage, CAVs can act independently with their policy networks without value instructions.
3) *A-MAPPO*: MAPPO with the attention mechanism in policy modules. CAVs aggregate the received messages through the attention layer to extract the latent relations of neighboring vehicles.

TABLE II
Experiment Parameters

Parameter	Value
clip threshold	0.2
discount factor γ	0.99
learning rate (actor)	0.005
learning rate (critic)	0.005
number of steps per episode	2000
time step limit per round	120
simulation frequency (Hz)	5

Algorithm 1 HAG-TRPO

Initialize actor and critic network parameters $\theta^{i_{1:N}}$, ϕ.
Initialize total training episodes E, replay buffer $\mathcal{B}$.
Initialize the time step limit per round T.
for episode $e = 0, 1, ..., E-1$ **do**
  for step $t = 0, 1, ..., T-1$ **do**
    Get observation O^t from the environment and actions A^t from attention-based policies.
    Get changed observation $\tilde{O}^t$ and reward R^t.
    Store tuple $(O^t, A^t, R^t, \tilde{O}^t)$ into buffer $\mathcal{B}$.
  end for
  Retrieve a minibatch of experience B from buffer $\mathcal{B}$.
  Calculate initial advantage function $M^{i_{1:1}}$ with (8).
  for agent $i_n = i_1, ..., i_N$ **do**
    Obtain the Hessian of KL-divergence with (10).
    Update the actor network $\pi_{\theta^{i_n}}$ with (11).
    Compute $M^{i_{1:n+1}}$ with (8).
  end for
  Update the critic value network V_ϕ with (12).
end for

C. Performance Evaluation

(1) Reward: Fig. 3 depicts the convergence of the episode reward among all learning-based methods under HOM and HET scenarios. All algorithms are initialized with random parameters, and their initial rewards tend to gather around low values indicating frequent collisions and few success cases.

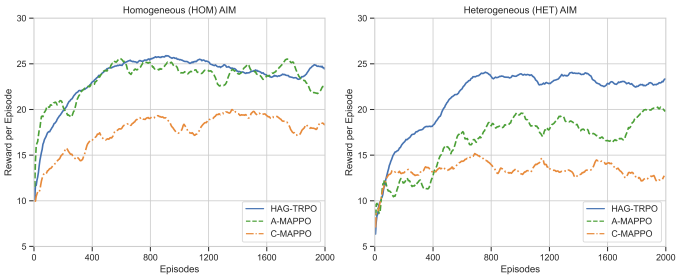

Fig. 3. The total reward per episode of all MARL-based methods

As their reward curve converge, it can be observed that HAG-TRPO and A-MAPPO exhibit similar performances in HOM-AIM, and both of them outperform C-MAPPO with higher rewards. This shows that agents equipped with the attention module have enhanced performance in sensing latent conflicts through disordered messages to avoid collisions through reasonable actions.

Compared with HOM-AIM, we notice an overall decrease of reward in HET-AIM due to its intricacy introduced by the vehicular divergence in dynamics and perception. Nonetheless, the HAG-TRPO method still achieves the highest reward with an obvious positive performance gap compared with the other MARL-based methods, as agents can update their policies with previous advantages considering heterogeneity.

(2) Safety: We proceed 30 rounds of fixed-time experiments for all mentioned methods. Within each round, we conduct a 10000-step traffic simulation, wherein CAVs and HCVs interact to resolve conflicts and reach their desired exit lane. The general traffic safety metric is depicted from both holistic and individual perspectives.

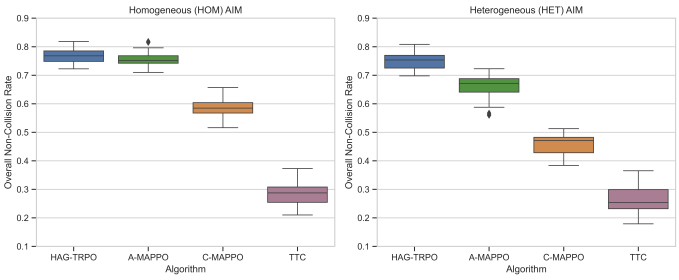

Fig. 4. The intersection holistic non-collision rate of all methods

The holistic intersection safety is described numerically with the overall non-collision rate, which represents the proportion of episodes in which all CAVs successfully pass the intersection without collisions, relative to the total number of episodes. As shown in Fig. 4, the average non-collision rate of TTC stands at merely 28.5% for HOM-AIM and 26.1% for HET-AIM. The reason is that TTC relies on the strong assumption that observed vehicles maintain their current velocity and heading, thus struggles to predict others' intentions when multiple vehicles interact simultaneously.

Both HAG-TRPO and A-MAPPO benefit from the attention mechanism in learning the inter-vehicle weighted relevance and improving traffic safety, as their non-collision rates generally surpass those of C-MAPPO and TTC. Specifically, HAG-TRPO yields a 76.7% non-collision rate in HOM-AIM, similar to A-MAPPO (75.6%). Whereas in HET-AIM, HAG-TRPO achieves the highest non-collision rate (75.1%), and exhibits the minimal performance degradation while other learning-based methods encounter an evident decrease. This proves that our proposed method exhibits its elasticity in handling complex scenarios by learning from prerequisite advantages for the current agent.

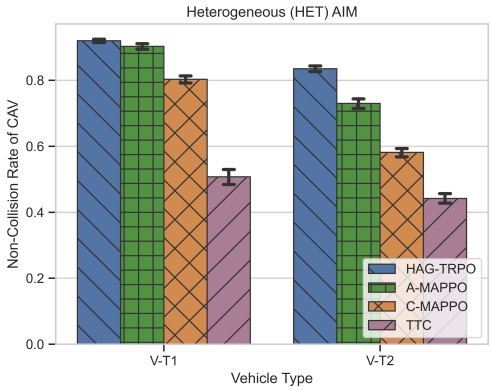

Fig. 5. CAV's non-collision rate of all methods in Heterogeneous AIM

For the individual safety, we assess the non-collision rate of a single CAV concerning various vehicle types in HET-AIM to verify the algorithmic coverage on account of vehicular variability. Through Fig. 5, we observe a general decline on V-T2 non-collision rate for learning-based methods comparing to V-T1 due to the bulkiness in size, kinematic control and perception capacity of V-T2 agents, along with the challenges posed by optimizing heterogeneous agents. Under the CTDE schema, C-MAPPO and A-MAPPO demonstrate the effectiveness of RL-based methods in learning to cooperate by achieving 80.3% and 90.3%. However, they conduct the policy-update tactics without specialized adaptation to heterogeneous agents, leading to their performance drop by 22.2% and 17.3% in V-T2 non-collision rate. Nonetheless, V-T2 agents driven by HAG-TRPO are able to leverage preordinal advantage from V-T1 agents. This helps minimize the heterogeneity learning defect, leading to a superior performance compared to A-MAPPO, C-MAPPO, and TTC, with respective improvements of 10.5%, 25.4%, and 39.3%.

(3) Efficiency: We record the total passing delays for 1500 CAVs to pass through the intersection. Within each episode, the arrival time cost of each vehicle is added to the total delays elapsed since the previous episode. In addition, episodes with collisions are assigned a time penalty equivalent to the maximum time step of a testing round to punish vehicular hastiness. The maximum time step is set to 100000.

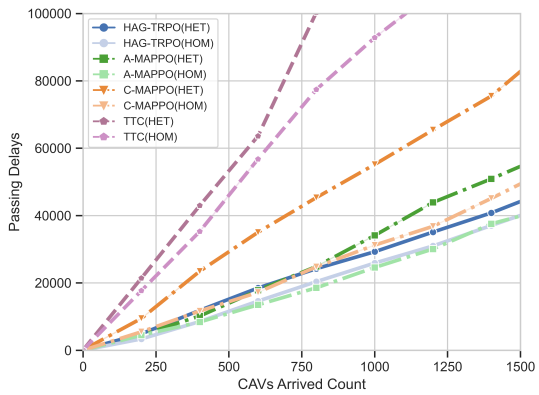

Fig. 6. Total passing delays for successful passage of vehicles

The experiments are carried out in the same initial state with zero CAVs arrival count. As CAVs gradually arrive, the accumulate traffic delay increases in each algorithm, revealing their performance disparities. As shown in Fig 6, the TTC method fails to complete 1500 CAVs' scheduling within a limited time steps of 100000 as it frequently causes collisions that encumber the traffic efficiency. In HOM-AIM cases, HAG-TRPO and A-MAPPO finish 1500 CAVs' passings in 40848 and 41115 time steps, respectively, both superior to C-MAPPO. However, in HET-AIM caes, A-MAPPO's performance is degraded due to the complexity in scenarios with agent-level disparities, resulting in a time-consuming passage of 56032 for total passing delays. Whereas, HAG-TRPO achieves merely 45450 passing delays in HET-AIM, indicating an efficiency increase of 23.3% over A-MAPPO. This is attributed by the advanced cooperation of HAG-TRPO, wherein CAVs notice inter-agent latent discrepancies on current velocity and acceleration capacity through advantage-based sequential update. This enables them to determine a viable plan for conducting an efficient passing.

V. Conclusion

In this paper, our study addressed the challenge of vehicle heterogeneity in AIM with multiple heterogeneous CAVs. To collaborate effectively under agent heterogeneity, we proposed an MARL-based cooperation method that exploits the sequential update schema for policy networks based on the multi-agent advantage decomposition lemma to boost flexibility on heterogeneous AIM, while applying the attention mechanism to reinforce vehicular perception on complicated surroundings. In addition, we implemented a temporal module to capture the implicit relevance for enhanced comprehension of the global status. Numerical results affirmed that our proposed method proceeds intersection tasks with faster traffic flow and high safety standard in both homogeneous and heterogeneous scenarios.

References

[1] K. Dresner and P. Stone, "Multiagent traffic management: A reservation-based intersection control mechanism," in *Proc. 3rd Int. Jt. Conf. Auton. Agents Multiagent Syst. (AAMAS)*, New York, NY, United states, Jul. 2004, pp. 530–537.

[2] R. Van der Horst and J. Hogema, "Time-to-collision and collision avoidance systems," *Verkeersgedrag in Onderzoek*, pp. 59–66, Jan. 1994.

[3] F. D. Salim, L. Cai, M. Indrawan, and S. W. Loke, "Road intersections as pervasive computing environments: towards a multiagent real-time collision warning system," in *Proc. 6th Annu. IEEE Int. Conf. Pervasive Comput. Commun. (PerCom)*, Hong Kong, China, Mar. 2008, pp. 621–626.

[4] M. Cai, Q. Xu, C. Chen, J. Wang, K. Li, J. Wang, and X. Wu, "Multi-lane unsignalized intersection cooperation with flexible lane direction based on multi-vehicle formation control," *IEEE Trans. Veh. Technol.*, vol. 71, no. 6, pp. 5787–5798, Jun. 2022.

[5] X. Pan, B. Chen, S. Timotheou, and S. A. Evangelou, "A convex optimal control framework for autonomous vehicle intersection crossing," *IEEE Trans. Intel. Transp.*, vol. 24, no. 1, pp. 163–177, Jan. 2023.

[6] H. Xu, Y. Zhang, L. Li, and W. Li, "Cooperative driving at unsignalized intersections using tree search," *IEEE Trans. Intel. Transp.*, vol. 21, no. 11, pp. 4563–4571, Nov. 2020.

[7] A. Mirheli, M. Tajalli, L. Hajibabai, and A. Hajbabaie, "A consensus-based distributed trajectory control in a signal-free intersection," *Transp. Res. Part C Emerg. Technol.*, vol. 100, pp. 161–176, Mar. 2019.

[8] Y. Hou, Z. Wei, R. Zhang, X. Cheng, and L. Yang, "Hierarchical task offloading for vehicular fog computing based on multi-agent deep reinforcement learning," *IEEE Trans. Wireless Commun.*, pp. 1–1, 01 2023.

[9] Z. Wei, B. Li, R. Zhang, X. Cheng, and L. Yang, "Many-to-many task offloading in vehicular fog computing: A multi-agent deep reinforcement learning approach," *IEEE Trans. Mobile Comput.*, vol. 23, no. 3, pp. 2107–2122, 2024.

[10] G.-P. Antonio and C. Maria-Dolores, "Multi-agent deep reinforcement learning to manage connected autonomous vehicles at tomorrow's intersections," *IEEE Trans. Veh. Technol.*, vol. 71, no. 7, pp. 7033–7043, Jul. 2022.

[11] C. Huang, J. Zhao, H. Zhou, H. Zhang, X. Zhang, and C. Ye, "Multi-agent decision-making at unsignalized intersections with reinforcement learning from demonstrations," in *Proc. 34th IEEE Intell Veh Symp (IV)*, Anchorage, AK, United states, Jun. 2023, pp. 1–6.

[12] J. Zheng, K. Zhu, and R. Wang, "Deep reinforcement learning for autonomous vehicles collaboration at unsignalized intersections," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOMM)*, Virtual, Online, Brazil, Dec. 2022, pp. 1115–1120.

[13] J. Liu, W. Zhao, C. Wang, C. Xu, L. Li, Q. Chen, and Y. Lian, "Eco-friendly on-ramp merging strategy for connected and automated vehicles in heterogeneous traffic," *IEEE Trans. Veh. Technol.*, vol. 72, no. 11, pp. 13 888–13 900, 2023.

[14] A. Coppola, D. G. Lui, A. Petrillo, and S. Santini, "Eco-driving control architecture for platoons of uncertain heterogeneous nonlinear connected autonomous electric vehicles," *IEEE Trans. Intel. Transp.*, vol. 23, no. 12, pp. 24 220–24 234, 2022.

[15] T. Rashid, M. Samvelyan, C. Schroeder, G. Farquhar, J. Foerster, and S. Whiteson, "Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning," in *Proc. 35th Int. Conf. Mach. Learn. (ICML)*. Stockholm, Sweden: PMLR, 2018, pp. 4295–4304.

[16] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu, "The surprising effectiveness of PPO in cooperative multi-agent games," *36th Adv. Neural Inf. Proces. Syst. (NIPS)*, vol. 35, pp. 24 611–24 624, Dec. 2022.

[17] P. Polack, F. Altché, B. d'Andréa-Novel, and A. de La Fortelle, "The kinematic bicycle model: A consistent model for planning feasible trajectories for autonomous vehicles," in *Proc. 28th IEEE Intell Veh Symp (IV)*, Redondo Beach, CA, United states, Jun. 2017, pp. 812–818.

[18] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat.*, vol. 62, no. 2, p. 1805, Aug. 2000.

[19] J. G. Kuba, R. Chen, M. Wen, Y. Wen, F. Sun, J. Wang, and Y. Yang, "Trust region policy optimisation in multi-agent reinforcement learning," in *Proc. 10th Int. Conf. Learn. Represent. (ICLR)*, Virtual, Online, Apr. 2022, p. 1046.

[20] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," in *Proc. 4th Int. Conf. Learn. Represent. (ICLR)*, San Juan, Puerto rico, May 2016.

[21] E. Leurent, "An environment for autonomous driving decision-making," GitHub repository, 2018.