2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Towards Enhanced Fairness and Sample Efficiency in Traffic Signal Control

Xingshuai Huang[1], Di Wu[1], Michael Jenkin[2], and Benoit Boulet[1]

***Abstract*— Traffic signal control (TSC) has seen substantial advancements through the application of reinforcement learning (RL) algorithms, which have shown remarkable potential in enhancing traffic flow efficiency. These RL-based approaches often surpass traditional rule-based methods, particularly in dynamic traffic environments. However, current RL solutions for TSC predominantly rely on model-free methods, necessitating extensive environmental interactions during training. This requirement can be prohibitively expensive or unfeasible in real-world implementations. Furthermore, existing methods have frequently neglected the issue of fairness in multi-intersection control, resulting in unbalanced congestion across different intersections. To address these challenges, we present FM2Light, a fairness-aware model-based multi-agent RL framework for TSC. Our approach leverages an ensemble of global world models for generating synthetic samples to enhance sample efficiency, thereby mitigating the data-intensive nature of the training process. Additionally, FM2Light incorporates a refined reward structure to promote fairness and improve coordination across multiple intersections. Extensive evaluations conducted in diverse real-world scenarios demonstrate that FM2Light achieves performance comparable to or exceeding that of model-free RL (MFRL) methods, while significantly reducing sample requirements and ensuring more equitable control among multiple agents.**

I. Introduction

Traffic congestion has become one of the main bottlenecks hindering urban development. Stop-and-go delays caused by signalized intersections account for 12–55% of citizens' commuting time, according to studies in urban areas [1]. Traffic congestion not only affects commuting efficiency but also exacerbates fuel consumption and pollutant emissions induced by vehicle idling, which is detrimental to the environment. Traffic signal control (TSC), a promising solution that necessitates minimal infrastructure modifications, has proven to be highly effective in addressing these issues and has been widely studied (see [1], for example).

Traditional pre-timed methods directly control traffic signals using simple timers, which can work effectively in areas with constant traffic patterns [1]. With the growing volume of automobiles on roads, modern traffic becomes increasingly complex and unpredictable, rendering simple timers less effective for managing traffic patterns. Recent advances [2], [3], [1] in intelligent and adaptive TSClers have shown their capability to handle such problems. RL has emerged as a powerful tool for traffic management, drawing researchers' attention for its ability to intelligently control traffic signals. This approach learns optimal strategies by adapting to changing traffic patterns, offering a dynamic solution to complex traffic problems.

Many previous RL-based TSC methods [4], [5] significantly improve overall vehicle passing efficiency in complex traffic scenarios compared to traditional pre-timed controllers, but are limited to a single intersection in isolation. They fail to take into account the impact of TSC on neighboring intersections, which might lead to conflicting effects between intersections due to a lack of cooperation. Researchers have turned to multi-agent RL (MARL) algorithms with decentralized architectures to manage traffic signals in large-scale road networks. In this type of approach, each local RL agent controls a single intersection through partial observation and restricted communication [6]. Most existing MARL approaches focus on addressing nonstationarity caused by the evolving actions of other RL agents [7], and improving agent communication [3].

Even taking multiple intersections and their interaction into account, there are still two main problems with the existing TSC methods. **(1)** These algorithms require a great amount of training data collected through interactions with the environment, which is infeasible in practice due to the excessive training time and severe congestion that might be induced during the learning process. Previous studies attempt to reduce required training data by taking advantage of meta-learning methods [8], [2] and improve sample efficiency using model-based RL (MBRL) approaches [8], but most of them still require a vast amount of training data from source tasks or can only be applied to single-intersection scenarios. **(2)** Given that each agent in a decentralized architecture focuses on its own cumulative reward, existing approaches typically ignore fairness in controlling different intersections. Fairness, encompassing various dimensions like Pareto-efficiency, equity, impartiality, and envy-freeness [9], has emerged as a significant consideration in diverse fields, such as robotics [10] and unmanned aerial vehicle (UAV) [11]. Existing work that considers fairness in TSC is typically limited to vehicle level [12], or lane level but at an isolated intersection [13], [14].

To address sample deficiency and unfairness problems in current MARL-based TSC algorithms, this study introduces FM2Light, a fairness-aware model-based MARL algorithm novelly extending fairness to the network level. Specifically, an ensemble of probabilistic global world models of the environment are learned, and a model-based method is adopted

[1]Xingshuai Huang, Di Wu, and Benoit Boulet are with the Department of Electrical and Computer Engineering, McGill University, Montreal, QC H3A 0G4, Canada xingshuai.huang@mail.mcgill.ca, di.wu5@mcgill.ca, benoit.boulet@mcgill.ca

[2]Michael Jenkin is with the department of Electrical Engineering and Computer Science, York University, Toronto, ON M3J 1P3, Canada jenkin@yorku.ca

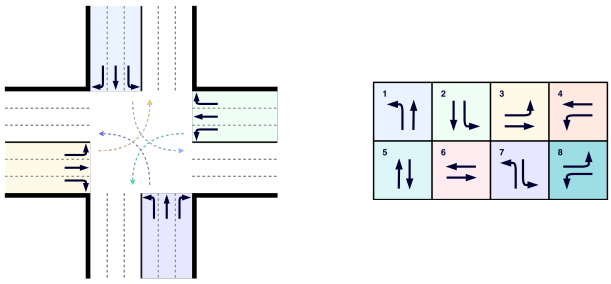

Fig. 1. A standard intersection with 12 approaching lanes. Left: the intersection layout; right: traffic phases excluding the right-hand turn.

to improve the policy optimization of a MARL algorithm. The fairness-aware reward is designed to improve both traffic efficiency and fairness among all the signalized intersections in the road network.

We present our contributions as follows: (i) We propose a novel model-based MARL framework for controlling multi-intersection traffic signals. The proposed FM2Light framework can naturally be adapted to other MARL algorithms. (ii) A novel fairness-aware reward is designed for the multi-intersection TSC to balance efficiency and fairness among all intersections in a road network, which also facilitates coordination between intersections. (iii) Through rigorous experiments on various real-world TSC scenarios, the proposed FM2Light is shown to significantly enhance sample efficiency and fairness thereby alleviating reliance on enormous real-world interactions and ameliorating heavy traffic congestion.

II. Related Work

A. Reinforcement Learning-Based Traffic Signal Control

With the explosive growth of vehicle numbers on roadways and the rapid development of computing power, RL-based TSC has shown its superior potential over traditional pre-timed controllers in handling more complex traffic scenarios [15]. For multi-intersection scenarios within a given road network of a city, MARL is popular for coordinating traffic flow at different intersections. MPLight [16] scales the decentralized control to expansive urban networks, managing up to 2510 signalized intersections via parameter sharing and enabling coordination with a pressure-based reward. To handle intersections with varied structures, AttendLight [17] proposes a universal controller with attention models, which enables direct control for intersections with any style. UniLight [3] proposes a universal communication form to exploit prediction information across intersections. MABRL [18] replaces deep neural networks with broad networks to accelerate remodeling and proposes a dynamic interaction mechanism to process global information. MetaVIM [2] further employs a meta-learning method to enhance the generalizability of the decentralized policy for the multi-agent TSC system.

B. Fairness in Traffic Signal Control

Fairness in RL typically focuses on societal fairness which mitigates bias caused by learning algorithms and non-societal fairness which enhances fair resource allocation [19]. For TSC, non-societal fairness can be considered as the equitable distribution of waiting times among all traffic streams, minimizing excessive delays for individuals or groups of vehicles. Several previous studies have introduced fairness into RL-based TSC. FairLight [20] proposes a new Fairness Index originated from the user satisfaction index to improve the travel quality of the individual vehicle. FELight [12] proposes a fairness index based on extra phase waiting time, raising attention to less commonly chosen phases. Different from previous studies which focus more on vehicle or lane-level fairness, our work introduces fairness from the multi-agent aspect by coordinating the throughput of intersections using a fairness-aware reward design.

C. Model-Based Multi-Agent Reinforcement Learning

Many recent approaches (e.g., [21]) combine MBRL with MARL to solve multi-agent control problems and improve sample efficiency. MAMBPO [22] uses centralized training to learn an ensemble of global world models, which generate imagined transitions for policy learning and then use decentralized execution to execute control policies. The work proposed by [23] further enhances exploration by adding a centralized exploration policy. MAMBA [24] learns decentralized world models using only communication between agents. MAG [25] addresses the challenge of multi-step predictions and mutual impacts of local models and policies by treating local models as agents while considering the policies as the environment.

Our work novelly employs global world models to enhance the optimization of TSC policies and takes into account fairness among intersections, facilitating coordination between multiple agents.

III. Preliminaries

A. Traffic Signal Control Problem

TSC refers to the selection of combinations of traffic signals at single or multiple signalized intersections. Each intersection is composed of several approaches, lanes, traffic flows, and signal phases.

Approach and Lane. The area where several approaches interact is defined as an intersection [8]. Vehicles head towards an intersection through the incoming approaches while leaving an intersection through the outgoing approaches. Each approach can be divided into different lanes that restrict the movements of vehicles, e.g., go straight, turn left, and turn right. A standard intersection with four three-lane approaches is shown in the left part of Figure 1.

Signal Phase. Signal phases are designed to control vehicle movements in different lanes and prevent conflicts [1]. Each signal phase denotes a combination of non-conflicting traffic signals for various lanes simultaneously. Eight primary signal phases are presented on the right side of Figure 1. Note that each signal phase excludes right-turn movements since it generally provides limited restrictions.

Traffic flow. Traffic flow is formed by the continuous flow of vehicles on the road [1]. Traffic flow is the number of

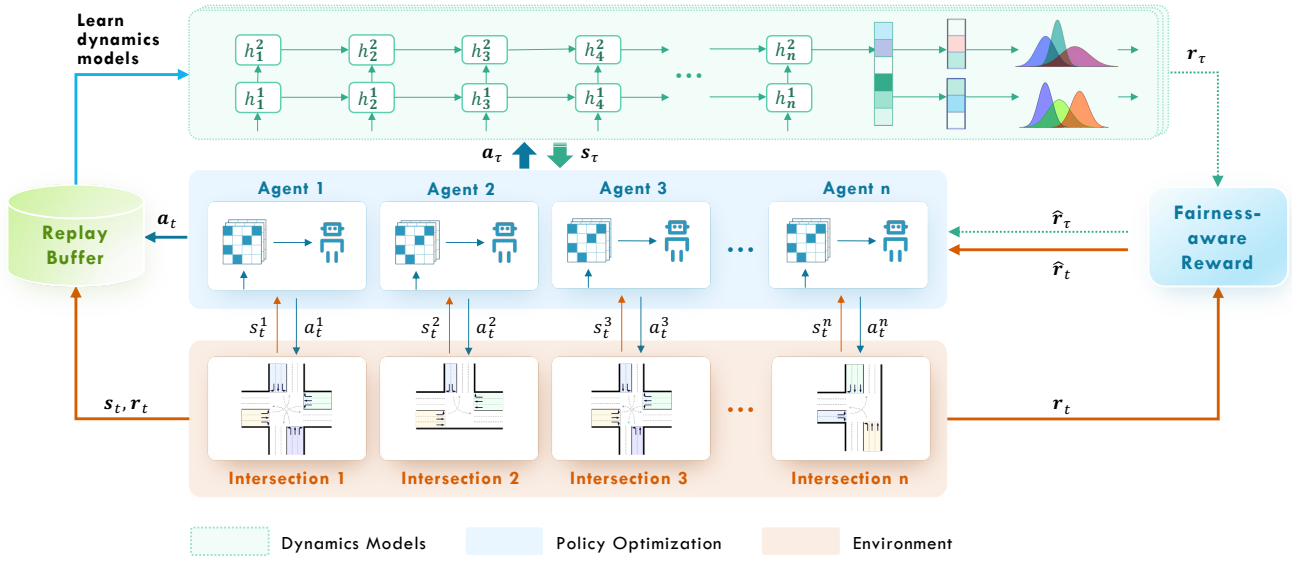

Fig. 2. FM2Light Framework. The designed solution can address TSC for multiple intersections.

vehicles that pass through a lane at a specified location or section per hour, which can be calculated as the multiplication of traffic density and travel speed.

B. Multi-Agent Reinforcement Learning

TSC is generally formulated as a Markov Decision Process (MDP) $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$ composed of state $\mathcal{S}$, action $\mathcal{A}$, state transition $\mathcal{P}$, reward function $\mathcal{R}$, and discount factor γ [15]. The control objective is to maximize the expected future discounted return $\mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r^i \left(s_t^i, a_t^i\right)\right]$ under policy π at time t. The TSC of multiple intersections can be seen as playing a fully cooperative game, whose goal is to optimize the global objective. Decentralized MARL is a potential solution, where each agent i controls an individual intersection and maximizes its own expected return, without or with limited information shared with other agents.

C. Model-Based Reinforcement Learning

In contrast to model-free RL algorithms which operate under the assumption of unknown state transition $\mathcal{P}$ and reward function $\mathcal{R}$, MBRL approaches embark on a two-fold strategy. Initially, they seek to acquire a comprehensive world model $M(S, A)$ of the environment. This learned model is subsequently utilized to facilitate the optimization of the value function or policy, thereby augmenting the efficiency of sample utilization. With respect to the problem of TSC, both the state transition $\mathcal{P}\left(S_{t+1} \mid S_t, A_t\right)$ and the reward function $\mathcal{R}\left(R_{t+1} \mid S_t, A_t\right)$ are derived via supervised learning using a set of experiences denoted by (s, a, s', r).

IV. Methodology

This part presents FM2Light and the framework is shown in Fig 2. FM2Light leverages the learned global world models to facilitate policy optimization. Independent deep Q-network (IDQN) [1] a fully decentralized MARL algorithm that controls each intersection with a separate agent, is adopted as the optimization method. The fairness-aware reward exploits fairness measurement as a penalty to better balance traffic efficiency among different intersections and enhance coordination.

A. Markov Decision Process Formulation

TSC problems are generally formulated as MDPs. This part defines the state, action, and fairness-aware reward function in the MDP setting.

State. Following the state settings in RESCO [1], we utilize four state variables including the current phase, the number of stopped vehicles, and the number and the average speed of approaching vehicles at each signalized intersection. The joint state $\boldsymbol{s}_t := \left\{s_t^i\right\}_{i=1}^I$ is a combination of states of each intersection i, where I is the number of intersections in a road network.

Action. The joint action $\boldsymbol{a}_t := \left\{a_t^i\right\}_{i=1}^I$ is defined as the selection of a set of non-conflicting phases (green light) of all intersections for the next time step. A yellow phase is implemented as a constraint if the chosen phase differs from the currently active phase.

Fairness-aware reward. Given that total travel time can only be computed in hindsight, this work attempts to minimize the total pressure $\sum p_t^i$ at all intersections, where the pressure p_t^i is defined as the difference between the total queue lengths of all upstream lanes and those of all downstream lanes at intersection i at time step t [1]. Therefore, the primary reward for agent i is $r_t^i = -p_t^i$. However, each independent agent aims to maximize its own cumulative rewards, which might cause conflicting effects on the adjacent intersections and lead to severe pressure at some intersections. To avoid locally extreme traffic congestion, we propose a fairness-aware reward function to balance global efficiency and fairness. Specifically, the utility [26] of intersection i at time step t is defined as the average reward over the past t time steps

Algorithm 1 FM2Light: Fairness-Aware Model-Based Multi-Agent Reinforcement Learning

Require: Training episodes K; task horizon H; world model update frequency M; number of imaginary rollouts C; increment of the number of rollouts per episode c; imagined rollout horizon T

Initialize policy π_{θ_i} for agent i, world model p_{ϕ_j} in the ensemble, real transition buffer $D_{real} = \emptyset$, and imagined transition buffer $D_{img} = \emptyset$
for episode $k = 1, ..., K$ **do**
  Initialize environment
  for timestep $h = 1, ..., H$ **do**
    Select joint actions via policy π_{θ_i} of each agent
    Implement selected actions in the environment
    Add real transitions to D_{real}
    if $h\%M = 0$ **then**
      Update each world model $\left\{p_{\phi_j}\right\}$ in the ensemble with real transitions randomly sampled from D_{real}
    end if
  end for
  for C model rollouts **do**
    Randomly sample an initial state s from D_{real}
    Generate T-step imagined short rollout from the ensemble of world models $\left\{p_{\phi_j}\right\}$ using policy π_{θ_i} of each agent; Add imagined transitions to D_{img}
    Update policy π_{θ_i} of each agent with sampled transitions from $D_{real} \cup D_{img}$
  end for
  $C = C + c$
end for

$$u_t^i = \frac{1}{t}\sum_{\tau=0}^{t} r_\tau^i. \tag{1}$$

Three important aspects are considered in our fairness-aware reward: impartiality, efficiency, and equity. Impartiality means permutations of utilities make no impact on the results. Efficiency implies that one specific solution should be selected as a priority if it is preferred by all agents. Equity suggests that a transfer of rewards from richer to poorer agents results in a fairer solution, which is based on *Pigou-Dalton principle* [27]. The fairness can be measured by the coefficient of variation (CV) w.r.t. all the agents' utilities [26]

$$\mathrm{CV} = \sqrt{\frac{1}{I-1}\sum_{i=1}^{I}\frac{\left(u_t^i - \bar{u}_t\right)^2}{\bar{u}_t^2}}. \tag{2}$$

However, it is infeasible to optimize CV with the independent agent of each intersection under a decentralized structure due to the moving-target problem [28]. This work decomposes the fairness measurement to each signalized intersection implicitly by incorporating a fairness penalty into the reward function for each agent and proposing the fairness-aware reward

$$\hat{r}_t^i = r_t^i - \beta|u_t^i - \bar{u}_t|, \tag{3}$$

where $\bar{u}_t$ is the average utility over all intersections at time t, and β is the penalty coefficient. Note this neither communicates full states nor actions, but only rewards, which can be realized when the number of intersections is small. When applying it to large road networks, a gossip algorithm [26] can be adopted to calculate $\bar{u}_t$ iteratively. In the fairness-aware reward, r_t^i encourages each agent to minimize the pressure at the corresponding intersection, while $\beta|u_t^i - \bar{u}_t|$ represents the utility deviation from the mean, which penalizes the agent for any deviating behaviour. The fairness-aware reward not only enables the balance between traffic efficiency and fairness but also enhances coordination between agents' policies in the decentralized structure by allowing agents to coordinate with each other via $\bar{u}_t$.

B. Learning Global World Models

To improve sample efficiency and reduce the required number of interactions with the environment, this study learns world models to represent the dynamics of the control tasks and then facilitates policy optimization with the learned global world models. For multi-intersection TSC tasks with complex and high-dimensional dynamics, expressive neural networks show better representation capacity than Bayesian models such as Gaussian processes and simple time-varying linear models [29]. To further alleviate model bias, this work incorporates aleatoric uncertainty, i.e., the inherent randomness of an environment, into MBRL via probabilistic networks, and captures epistemic uncertainty, i.e., the uncertainty of the model due to lack of training data, by learning an ensemble of world models [29].

Specifically, the j'th world model parameterized by ϕ_j in the ensemble outputs two Gaussian distributions with diagonal covariances for the prediction of the next joint state and reward given the current joint state and action, i.e.:

$$\begin{aligned} p_{\phi_j} &= \Pr\left(\boldsymbol{s}_{t+1}, \hat{\boldsymbol{r}}_{t+1} \mid \boldsymbol{s}_t, \boldsymbol{a}_t\right) \\ &= \mathcal{N}\left(\boldsymbol{\mu}_{\phi_j}\left(\boldsymbol{s}_t, \boldsymbol{a}_t\right), \boldsymbol{\Sigma}_{\phi_j}\left(\boldsymbol{s}_t, \boldsymbol{a}_t\right)\right). \end{aligned}$$

Learning the global world model assumes access to global information, which might be difficult for real-time TSC [7]. However, this restriction is alleviated when each global world model is learned using supervised learning with stored non-real-time experiences $\{(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s'}, \hat{\boldsymbol{r}})\}$ from a replay buffer. Furthermore, centralized MARL algorithms generally suffer from combinatorially large joint action spaces, while representing the world models with neural networks handles this issue well. The negative log-likelihood is selected as the loss function for model learning

$$\mathcal{L}(\boldsymbol{\phi_j}) = -\sum_{n=1}^{N} \log p_{\boldsymbol{\phi_j}}\left(\boldsymbol{s}_{n+1}, \hat{\boldsymbol{r}}_{n+1} \mid \boldsymbol{s}_n, \boldsymbol{a}_n\right), \tag{4}$$

where N is the number of sampled real transitions. To decorrelate different models, each model is randomly initialized and trained with a randomly sampled subset of real

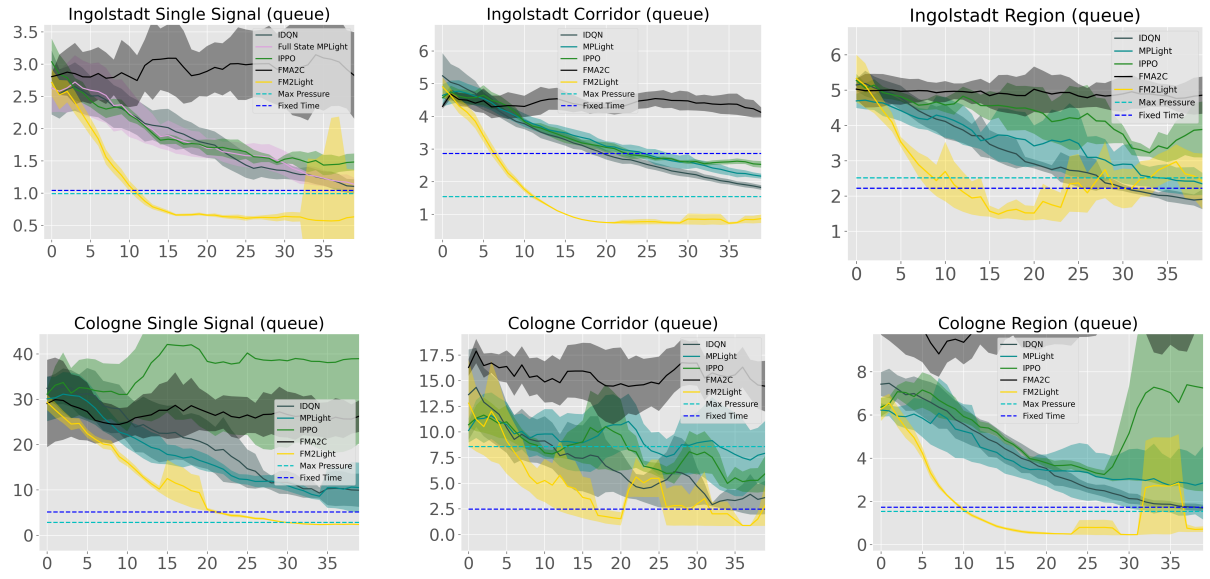

Fig. 3. Learning curves w.r.t. average queue length with means and variances over 5 random seeds. The horizontal and vertical axes represent the number of episodes and average queue length, respectively. The duration of an episode is 3600s. Parts of curves that are out of range are excluded from the figures.

transitions. All world models are continuously retrained with newly collected real transitions to alleviate distributional shift problems [30]. In order to capture the spatial and temporal dependencies in different lanes and intersections, we leverage long short-term memory (LSTM) to model the complex dynamics in the multi-intersection environment. With the learned ensemble of world models, we can either uniformly re-sample a model to make predictions via the selected model every time step or directly output the expected prediction over models. To mitigate accumulated errors caused by world models, we follow the short rollouts technique by generating multiple imagined short rollouts instead of a long-step rollout with the world models [30].

C. Model-Based Multi-Agent Reinforcement Learning

Given that centralized MARL suffers from the dimension curse of action space [28], IDQN [1], a fully decentralized MARL algorithm, is employed as our policy optimization algorithm. Each independent DQN agent i controls an individual intersection and optimizes its own policy by maximizing the cumulative reward. At time step t, agent i observes the partial state s_t^i, takes the optimal action a_t^i, and then receives the local reward $\hat{r}_{t+1}^i$. We use convolutional neural networks (CNNs) to aggregate state information over different lanes and output the approximated Q value for each candidate action according to the Bellman Equation:

$$Q\left(s_t^i, a_t^i\right) = \hat{r}_{t+1}^i + \gamma \max Q\left(s_{t+1}^i, a_{t+1}^i\right). \tag{5}$$

The pseudo-code of FM2Light is presented in Algorithm 1. In policy update iteration, for each rollout, we randomly sample an initial state from the real transition buffer and then collect a rollout of imagined trajectories into buffer D_{img} with a randomly sampled world model. Agents' policies are then updated with the sampled transitions from

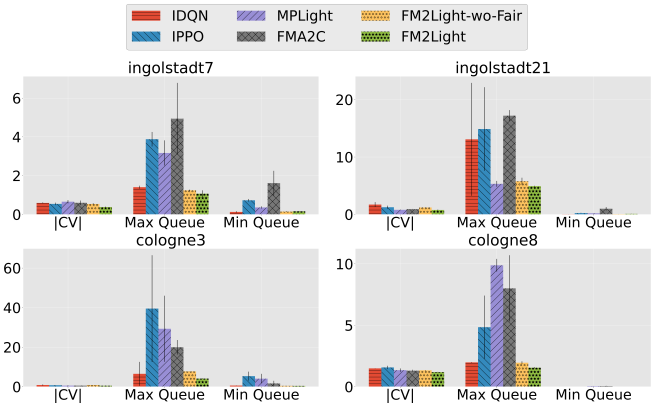

Fig. 4. Fairness comparison of different algorithms with means and variances over 5 random seeds. FM2Light outperforms other baselines on all multi-intersection control tasks on both CV and Max Queue.

both D_{real} and D_{img}. A validation loss threshold is applied to avoid bad rollouts generated by untrusted world models. That is, only when the validation loss is below a threshold, do we use the world models to generate imagined rollouts. As the world models are generally getting better with more training data, the number of rollouts is incremented by c per episode.

D. Implementation Details

The task horizon H for each specific task and time step are 3600s and 10s, respectively. Therefore, we obtain 360 real transitions per episode. Each world model is represented by a 3-layer LSTM with 1024 hidden nodes, followed by 2 fully connected layer-based heads outputting state and reward predictions, respectively. The LSTM-based world model is selected for its better representation accuracy over Transformers, Convolutional Neural Networks, and Multi-

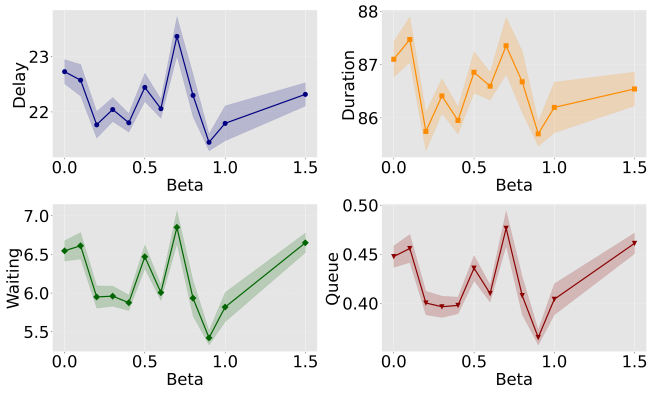

Fig. 5. Sensitivity analysis of penalty coefficient β on performance metrics over five random seeds. The horizontal and vertical axes represent the values of β and values of performance metrics, respectively.

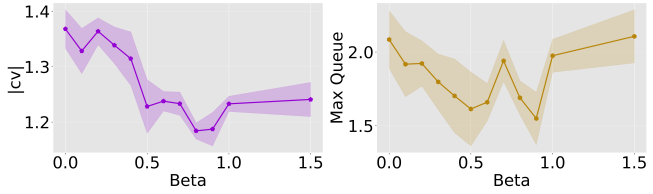

Fig. 6. Sensitivity analysis of penalty coefficient β on fairness metrics.

layer Perceptrons. The Q-network of each agent in IDQN is composed of a single convolutional layer and 3 fully connected layers to aggregate lane information for each agent.

V. Experiments

Our experiments are conducted using SUMO traffic simulator [31], on Intel(R) Core(TM) i9-10900F CPU @ 2.80GHz with 32.0 GB RAM (2933MHz) and a single Nvidia GeForce RTX 3070 GPU. We adopt two SUMO scenarios from real-world cities, Cologne and Ingolstadt (Col. and Ing. for short), which are "TAPAS Cologne" [32] and "InTAS" [33], respectively. Six different TSC tasks are adopted from RESCO benchmark [1]: 1) a single intersection control for each scenario, 2) 3-intersection and 7-intersection coordinated control for corridors of Cologne and Ingolstadt, respectively (Corr. for short), 3) 8-intersection and 21-intersection coordinated control within a downtown region of Cologne and Ingolstadt, respectively (Reg. for short). Following the settings in the RESCO benchmark, sensors at each signalized intersection measure traffic conditions within a 200-meter radius.

We employ several traditional and MARL-based methods as baselines (details can be seen in Appendix): Fixed Time, Max Pressure [16], IDQN [34], IPPO [1], MPLight [16], and FMA2C [1].

A. Comparison of Sample Efficiency

To compare the performance of these algorithms, average queue length over intersections is used as the evaluation metric [1]. More results on different evaluation metrics can be seen in the Appendix. Lower values for these metrics are better.

Figure 3 illustrates the learning curves w.r.t average queue length over 5 random seeds. FM2Light shows significantly improved sample efficiency than other baselines. Specifically, FM2Light requires only a quarter of the training episodes or data of the best baseline, i.e., IDQN, to reach comparable or even better results, which demonstrates the improved sample efficiency over model-free MARL baselines. Compared to IPPO, the FM2Light method achieves even better performance using 40-50$\times$ fewer data. MPLight and FMA2C significantly underperform compared to FM2Light in both sample efficiency and final performance. Among the selected MARL baselines, IDQN and MPLight are more sample-efficient than IPPO and FMA2C. Even though traditional controllers, i.e., Fixed time and Max pressure, respectively achieve comparable results on certain tasks, they are unable to adapt to other tasks. These results highlight the benefits of the FM2Light method, that is, in real-world TSC, FM2Light algorithms can substantially decrease the number of interactions required with the environment during policy training while achieving comparable or superior performance.

B. Comparison of Fairness

To demonstrate improvements in traffic fairness at multiple signalized intersections, we employ two metrics, coefficient of variation (CV) w.r.t. utility (defined in Equation 2) and maximum average queue length (Max Queue for short) across all intersections. We also show Min Queue for additional information. Given that the reward function varies by method, we unify the calculation of utility to be based on queue length. Lower CV means greater fairness and lower Max Queue indicates less likelihood of congestion.

As shown in Figure 4, FM2Light outperforms all other baselines on all multi-intersection control tasks on both fairness dimensions, yielding average improvements of 11.8% and 19.4% over the best baseline with respect to CV and Max Queue, respectively. FM2Light with the original reward setting where fairness is not considered (FM2Light-wo-Fair)achieves a performance comparable to IDQN, which might be due to its use of IDQN-based policy optimization. However, FM2Light-wo-Fair shows worse results on two fairness dimensions compared to FM2Light, which demonstrates the importance of fairness-aware reward. The better fairness of our FM2Light indicates a better efficiency balance across intersections, reducing the likelihood of severe congestion at certain intersections.

C. Hyperparameter Sensitivity Analysis

Figure 5 and 6 illustrate the performance and fairness of FM2Light on Cologne Region task under different penalty coefficients β, respectively. The value of β is chosen to be 0.9 for our FM2Light as it is the value that generates the best performance in most evaluation metrics (queue length, delay, and waiting time). Generally, CV gets better as β increases. However, Max Queue gets lower with larger β when β is smaller than 0.9, while further increasing results in higher

Max Queue. This might be due to that a larger β degrades the overall performance, which can also be reflected in Figure 5.

VI. Conclusion and Discussion

This study describes FM2Light, a novel fairness-aware model-based multi-agent reinforcement learning approach to address low sample efficiency and unfairness issues in RL-based TSC methods. An ensemble of probabilistic networks is learned to represent the global world model of the environment and used to generate imagined transitions for improving policy optimization. A novel fairness-aware reward function is presented to coordinate independent agents in a decentralized structure and constrain fairness over intersections. Under several different real-world TSC tasks and scenarios, experimental results show that FM2Light can significantly reduce required data collected from interactions with the environment to obtain well-trained policies and improve fairness among intersections thus mitigating severe congestion. In our future work, we aim to further enhance data efficiency with a better state representation.

References

[1] J. Ault and G. Sharon, "Reinforcement learning benchmarks for traffic signal control," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[2] L. Zhu, P. Peng, Z. Lu, and Y. Tian, "Metavim: Meta variationally intrinsic motivated reinforcement learning for decentralized traffic signal control," *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[3] Q. Jiang, M. Qin, S. Shi, W. Sun, and B. Zheng, "Multi-agent reinforcement learning for traffic signal control through universal communication method," *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 2022.

[4] M. Kunjir and S. Chawla, "Offline reinforcement learning for road traffic control," *arXiv preprint arXiv:2201.02381*, 2022.

[5] P. Agand, A. Iskrov, and M. Chen, "Deep reinforcement learning-based intelligent traffic signal controls with optimized co2 emissions," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 5495–5500.

[6] X. Wang, L. Ke, Z. Qiao, and X. Chai, "Large-scale traffic signal control using a novel multiagent reinforcement learning," *IEEE transactions on cybernetics*, vol. 51, no. 1, pp. 174–187, 2020.

[7] T. Chu, J. Wang, L. Codecà, and Z. Li, "Multi-agent deep reinforcement learning for large-scale traffic signal control," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, pp. 1086–1095, 2019.

[8] X. Huang, D. Wu, M. Jenkin, and B. Boulet, "Modellight: Model-based meta-reinforcement learning for traffic signal control," *arXiv preprint arXiv:2111.08067*, 2021.

[9] M. Zimmer, C. Glanois, U. Siddique, and P. Weng, "Learning fair policies in decentralized cooperative multi-agent reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 967–12 978.

[10] Y. Msala, O. Hamed, M. Talea, and M. Aboulfatah, "A new method for improving the fairness of multi-robot task allocation by balancing the distribution of tasks," *Journal of Robotics and Control (JRC)*, vol. 4, no. 6, pp. 743–753, 2023.

[11] X. Luo, J. Xie, L. Xiong, Z. Wang, and Y. Liu, "Uav-assisted fair communications for multi-pair users: A multi-agent deep reinforcement learning method," *Computer Networks*, p. 110277, 2024.

[12] X. Du, Z. Li, C. Long, Y. Xing, S. Y. Philip, and H. Chen, "Felight: Fairness-aware traffic signal control via sample-efficient reinforcement learning," *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[13] M. Raeis and A. Leon-Garcia, "A deep reinforcement learning approach for fair traffic signal control," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021, pp. 2512–2518.

[14] J. Chen, Z. Zhang, J. Feng, and K. Zhu, "Fit: Fairness-aware intelligent traffic signal control with deep reinforcement learning," in *2021 IEEE 23rd Int Conf on High Performance Computing & Communications; 7th Int Conf on Data Science & Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)*. IEEE, 2021, pp. 846–852.

[15] M. Noaeen, A. Naik, L. Goodman, J. Crebo, T. Abrar, Z. S. H. Abad, A. L. Bazzan, and B. Far, "Reinforcement learning in urban network traffic signal control: A systematic literature review," *Expert Systems with Applications*, p. 116830, 2022.

[16] C. Chen, H. Wei, N. Xu, G. Zheng, M. Yang, Y. Xiong, K. Xu, and Z. Li, "Toward a thousand lights: Decentralized deep reinforcement learning for large-scale traffic signal control," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 3414–3421.

[17] A. Oroojlooy, M. Nazari, D. Hajinezhad, and J. Silva, "Attendlight: Universal attention-based reinforcement learning model for traffic signal control," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4079–4090, 2020.

[18] R. Zhu, L. Li, S. Wu, P. Lv, Y. Li, and M. Xu, "Multi-agent broad reinforcement learning for intelligent traffic light control," *Information Sciences*, vol. 619, pp. 509–525, 2023.

[19] P. Gajane, A. Saxena, M. Tavakol, G. Fletcher, and M. Pechenizkiy, "Survey on fair reinforcement learning: Theory and practice," *arXiv preprint arXiv:2205.10032*, 2022.

[20] Y. Ye, J. Ding, T. Wang, J. Zhou, X. Wei, and M. Chen, "Fairlight: Fairness-aware autonomous traffic signal control with hierarchical action space," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2022.

[21] X. Wang, Z. Zhang, and W. Zhang, "Model-based multi-agent reinforcement learning: Recent progress and prospects," *arXiv preprint arXiv:2203.10603*, 2022.

[22] D. Willemsen, M. Coppola, and G. C. de Croon, "Mambpo: Sample-efficient multi-robot reinforcement learning using learned world models," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 5635–5640.

[23] Q. Zhang, C. Lu, A. Garg, and J. Foerster, "Centralized model and exploration policy for multi-agent rl," *arXiv preprint arXiv:2107.06434*, 2021.

[24] V. Egorov and A. Shpilman, "Scalable multi-agent model-based reinforcement learning," *arXiv preprint arXiv:2205.15023*, 2022.

[25] Z. Wu, C. Yu, C. Chen, J. Hao, and H. H. Zhuo, "Models as agents: Optimizing multi-step predictions of interactive local models in model-based multi-agent reinforcement learning," *arXiv preprint arXiv:2303.17984*, 2023.

[26] J. Jiang and Z. Lu, "Learning fairness in multi-agent systems," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[27] H. Dalton, "The measurement of the inequality of incomes," *The Economic Journal*, vol. 30, no. 119, pp. 348–361, 1920.

[28] Z. Li, H. Yu, G. Zhang, S. Dong, and C.-Z. Xu, "Network-wide traffic signal control optimization using a multi-agent deep reinforcement learning," *Transportation Research Part C: Emerging Technologies*, vol. 125, p. 103059, 2021.

[29] K. Chua, R. Calandra, R. McAllister, and S. Levine, "Deep reinforcement learning in a handful of trials using probabilistic dynamics models," *Advances in neural information processing systems*, vol. 31, 2018.

[30] I. Clavera, J. Rothfuss, J. Schulman, Y. Fujita, T. Asfour, and P. Abbeel, "Model-based reinforcement learning via meta-policy optimization," in *Conference on Robot Learning*. PMLR, 2018, pp. 617–629.

[31] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.-P. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner, and E. Wießner, "Microscopic traffic simulation using sumo," in *2018 21st international conference on intelligent transportation systems (ITSC)*. IEEE, 2018, pp. 2575–2582.

[32] C. Varschen and P. Wagner, "Mikroskopische modellierung der personenverkehrsnachfrage auf basis von zeitverwendungstagebüchern," *Integrierte Mikro-Simulation von Raum-und Verkehrsentwicklung. Theorie, Konzepte, Modelle, Praxis*, vol. 81, pp. 63–69, 2006.

[33] S. C. Lobo, S. Neumeier, E. M. Fernandez, and C. Facchi, "Intas–the ingolstadt traffic scenario for sumo," *arXiv preprint arXiv:2011.11995*, 2020.

[34] J. Ault, J. P. Hanna, and G. Sharon, "Learning an interpretable traffic signal control policy," *arXiv preprint arXiv:1912.11023*, 2019.
[35] Y. Fujita, P. Nagarajan, T. Kataoka, and T. Ishikawa, "Chainerrl: A deep reinforcement learning library," *Journal of Machine Learning Research*, vol. 22, no. 77, pp. 1–14, 2021. [Online]. Available: http://jmlr.org/papers/v22/20-376.html

APPENDIX

We present the details of all baseline methods in this part.

1) **Fixed Time** controller selects joint phases according to a fixed cycle and held for a fixed duration;
2) **Max Pressure** controller [16] enables joint phases with the maximal joint pressure;
3) **IDQN** (independent DQN) [34] controls each intersection with an independent DQN agent, which is also the policy optimization algorithm of our method. It uses minus waiting time as the reward function while FM2Light adopts minus pressure instead for better performance. IDQN employs 1 CNN layer followed by 3 fully connected layers to aggregate lane information for each agent. Hyperparameters are identical to the default settings in the Preferred RL (PFRL) library [35] while adjusting the target network update frequency to 500 steps per update according to the Atari environment setting.;
4) **IPPO** (independent proximal policy optimization) [1] adopts the same network structure as IDQN while using multiple PPO agents. IPPO follows the default settings in PFRL for the Atari environment.;
5) **MPLight** [16] uses pressure as the reward function and shares parameters over all DQN agents. An extended MPLight [1] with additional states as IDQN is adopted for the control of the Ingolstadt single intersection to get better performance. MPLight uses the same hyperparameters as IDQN;
6) **FMA2C** [1] is built based on MA2C [7] where each intersection is controlled by an A2C agent. Neighborhood information as well as discounted reward and states are proposed to improve coordination between agents. FMA2C adopts the implementation and hyperparameter settings of the open-source MA2C [7].

Table I shows the detailed hyperparameter settings of FM2Light for reproduction. 5 different dynamics models are learned within the ensemble and updated once per episode. When training the dynamics models, learning rate, batch size, dropout rate, and patience of early stopping are set as 0.001, 64, 0.3 and 10, respectively. The initial number of imagined rollouts C is 10 and incremented by $c = 1$ per episode. $T = 18$ imagined transitions are generated for each short rollout. The learning rate, batch size, discount factor and target network update frequency are 0.001, 32, 0.99 and 500, respectively.

Besides the average queue length over intersections, more metrics, e.g., approximated average signal-induced delay (delay), average travel time (duration), and average waiting time (waiting) at intersections are also adopted for evaluation. Specifically, delay measures the difference between actual travel time and ideal travel time at the maximum speed limit, duration measures time spent at intersections, and waiting denotes the time vehicles stop at a red light. Figure7 shows the learning curves w.r.t. all metrics on the Ingolstadt Corridor task. We can see that all metrics of the same algorithm follow similar patterns. Therefore, we can easily get a similar conclusion from any one of these metrics.

TABLE I
HYPERPARAMETER SETTINGS FOR REPRODUCTION.

Hyperparameter	**Value**
Task horizon H	3600s
Time step	10s
Number of dynamics models in the ensemble	5
Initial number of imagined rollouts C	10
Rollout increment c	1
Imagine horizon T	18
Number of LSTM layers (model learning)	3
Learning rate (model learning)	0.001
Batch size (model learning)	64
Dropout rate (model learning)	0.3
Patience of early stopping (model learning)	10
Optimizer (model learning)	Adam
Number of CNN layers (policy learning)	1
Number of fully connected layers (policy learning)	3
Learning rate (policy learning)	0.001
Batch size (policy learning)	32
Discount factor	0.99
Target network update frequency	500

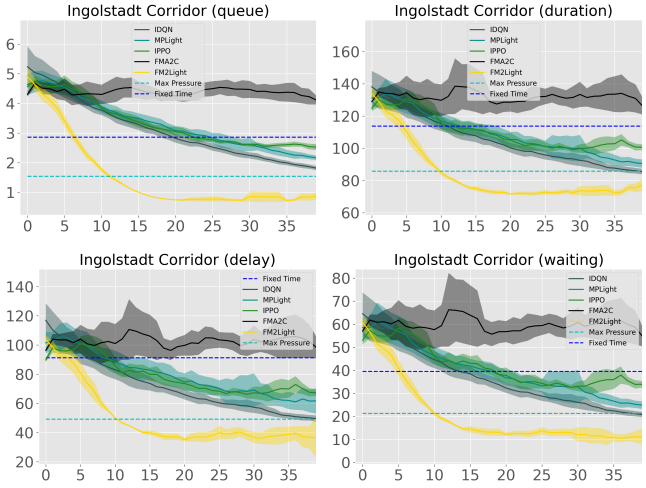

Fig. 7. Learning curves w.r.t. four different evaluation metrics for RL-based methods.