energies

Article

MMD-TSC: An Adaptive Multi-Objective Traffic Signal Control for Energy Saving with Traffic Efficiency

Yuqi Zhang, Yingying Zhou, Beilei Wang and Jie Song *

Software College, Northeastern University, Shenyang 112000, China; 2190088@stu.neu.edu.cn (Y.Z.); 20193043@stu.neu.edu.cn (Y.Z.); wangbl@swc.neu.edu.cn (B.W.)
* Correspondence: songjie@mail.neu.edu.cn

Abstract: Reducing traffic energy consumption is crucial for smart cities, and vehicle carbon emissions are a key energy indicator. Traffic signal control (TSC) is a useful method because it can affect the energy consumption of vehicles on the road by controlling the stop-and-go of vehicles at traffic intersections. However, setting traffic signals to reduce energy consumption will affect traffic efficiency and this is not in line with traffic management objectives. Current studies adopt multi-objective optimization methods with high traffic efficiency and low carbon emissions to solve this problem. However, most methods use static weights, which cannot adapt to complex and dynamic traffic states, resulting in non-optimal performance. Current energy indicators for urban transportation often fail to consider passenger fairness. This fairness is significant because the purpose of urban transportation is to serve people's mobility needs not vehicles. Therefore, this paper proposes **M**ulti-objective Adaptive **M**eta-**DQN** **TSC** (MMD-TSC), which introduces a dynamic weight adaptation mechanism to simultaneously optimize traffic efficiency and energy saving, and incorporates the per capita carbon emissions as the energy indicator. Firstly, this paper integrates traffic state data such as vehicle positions, velocities, vehicle types, and the number of passengers and incorporates fairness into the energy indicators, using per capita carbon emissions as the target for reducing energy consumption. Then, it proposes MMD-TSC with dynamic weights between energy consumption and traffic efficiency as reward functions. The MMD-TSC model includes two agents, the TSC agent and the weight agent, which are responsible for traffic signal adjustment and weight calculation, respectively. The weights are calculated by a function of traffic states. Finally, the paper describes the design of the MMD-TSC model learning algorithm and uses a SUMO (Simulation of Urban Mobility) v.1.20.0 for traffic simulation. The results show that in non-highly congested traffic states, the MMD-TSC model has higher traffic efficiency and lower energy consumption compared to static multi-objective TSC models and single-objective TSC models, and can adaptively achieve traffic management objectives. Compared with using vehicle average carbon emissions as the energy consumption indicator, using per capita carbon emissions achieves Pareto improvements in traffic efficiency and energy consumption indicators. The energy utilization efficiency of the MMD-TSC model is improved by 35% compared to the fixed-time TSC.

Keywords: sustainable transition; energy saving; reinforcement learning; meta-learning

check for updates

Citation: Zhang, Y.; Zhou, Y.; Wang, B.; Song, J. MMD-TSC: An Adaptive Multi-Objective Traffic Signal Control for Energy Saving with Traffic Efficiency. *Energies* **2024**, *17*, 5015. https://doi.org/10.3390/en17195015

Academic Editors: Sheng Du, Xiongbo Wan, Zixin Huang and Li Jin

Received: 3 September 2024
Revised: 5 October 2024
Accepted: 6 October 2024
Published: 9 October 2024

1. Introduction

Severe global warming has motivated people to increase their awareness of environmental protection. The development trend of smart cities is gradually shifting to-wards a sustainable transition [1]. Sustainable transition aims to improve traffic efficiency and focuses on saving energy and ensuring fairness. This includes reducing carbon emissions during driving and ensuring that all passengers have equal rights to road usage [2]. Therefore, fairness and carbon emissions have become key indicators to evaluate the development level of smart cities and the effectiveness of their sustainable transition initiatives.

Traffic signal control (TSC) provides a way to manage mixed traffic flow effectively. TSC can optimize traffic flow by reasonably setting each signal's phases and duration [3]. Therefore, TSC has become a potentially effective method to promote energy saving. However, traffic efficiency will be decreased if energy saving is the only goal. Traffic efficiency is one of the most important goals of a smart city [4]. An efficient transportation system can alleviate congestion and save travel time. TSC is a crucial method to improve traffic efficiency by regulating traffic flow in different directions at intersections based on traffic states. Therefore, the research object is a TSC method that can simultaneously improve traffic efficiency and reduce per capita carbon emissions under the traffic states of mixed traffic flow and vehicle types.

Reinforcement learning has become a popular research technology for TSC due to its ability to model and make decisions in complex environments [5]. The agents can control the phases of traffic signals based on traffic states. TSC models the traffic state at intersections based on the spatial characteristics of single or multiple intersections and vehicle positions from current or previous times. Most reward functions maximize traffic efficiency [6]. Traffic efficiency is often measured by traffic velocity and average queue length [7]. Few papers consider sustainability, such as harmful gas emissions and safety [8]. Dynamic sequence and fixed sequence are two types of actions. In the dynamic sequence mode, the order of phases is dynamically adjusted based on real-time traffic states and optimization algorithms to seek the optimal control strategy. The fixed sequence mode executes the phases in a predetermined order [9].

The current reinforcement learning methods for TSC primarily focus on ensuring smooth traffic flow [6]. While the results of these studies often include improvements in both traffic flow and energy saving, the main objective is typically to optimize traffic efficiency [10]. However, even when traffic efficiency is maximized, energy saving is not optimal, as additional factors beyond traffic efficiency, such as vehicle types and the number of passengers, influence carbon emissions. Consequently, per capita carbon emissions are not minimized when traffic efficiency is maximized. Traffic efficiency and per capita carbon emissions are distinct objectives that cannot be directly substituted for each other. Furthermore, most methods employ a fixed weighting approach to balance these two objectives, which fails to adapt to dynamic traffic environments.

Therefore, the existing TSC methods have the following limitations: First, they mainly focus on optimizing traffic efficiency, lacking a consideration of energy indicators such as per capita carbon emissions, and cannot directly use energy indicators as optimization objectives, making it difficult for the optimization results to meet the requirements of a sustainable transition. Second, when considering the sustainable transition, existing methods lack the consideration of fairness factors and fail to adequately address the travel needs and experiments of different passenger groups, while fairness is an essential component of the sustainable transition [11]. Moreover, most methods use fixed weights to balance traffic efficiency and energy saving, which cannot adapt to dynamically changing traffic environments. These limitations prevent existing methods from achieving the sustainable transition and going beyond only traffic efficiency optimization.

The research question is: How can a TSC model simultaneously optimize traffic efficiency and energy saving, incorporate fairness factors into the energy indicator, and dynamically balance these objectives through adaptive weights based on complex and varying traffic states?

This paper proposes the Multi-objective Adaptive Meta-DQN TSC model (MMD-TSC). DQN is a value-based deep reinforcement learning algorithm that uses deep neural networks to approximate the optimal Q-function (state-action value function). It can handle large-scale, high-dimensional state spaces and is very suitable for solving complex sequential decision-making problems. A Multi-objective DQN for TSC (TSC-Agent) was built by integrating multiple influencing factors into the state and designing the reward function. Based on the spatial distribution of vehicles as the state, the state adds the dimensions of vehicle types and the number of passengers. These are vital factors that influence traffic efficiency and energy saving. The multi-objective reward function adopts the weighting method. One of the objectives is queue length, and the other is per capita carbon emission. The weights of the two indicators range from zero to one, and the sum of these is one. An adaptive weight calculation method for traffic efficiency and per capita carbon emissions based on the state is designed to solve the problem of static weight design. A neural network is built to calculate the weight. It takes the state as input and weight as output.

The challenge lies in achieving multi-objective weight adaptation; namely, how does the algorithm train the neural network parameters for calculating the weight? Weights are often static values. The objective will be improved if weights are dynamic values obtained from the current traffic state. The dynamic weight is like a teacher, and the DQN is like a student. The student learns how to control traffic signals. The teacher determines the student's learning objectives and adjusts the teaching progress according to the student's learning progress. The two cooperate to make the student's learning better. The teacher will adjust the objectives in time, when the student's current learning objectives are well-learned. This paper adopts the meta-learning approach to construct an agent that dynamically controls weights based on traffic states. The dynamic weight agent is designed as the actor–critic model (weight agent). The actor of the AC model is the weight generator, and the PPO algorithm is applied to train the weight agent. Applying the reward of the TSC agent, the weight is adjusted dynamically according to the traffic states. When the training time of the TSC agent is greater than the threshold, the weight agent starts to train. Then, according to the new weights, the TSC-Agent continues to be trained, realizing cooperative learning. Figure 1 illustrates the differences between TSC and MMD-TSC. Firstly, MMD-TSC focuses more on the passengers inside the vehicle and optimizes based on per capita carbon emissions. Secondly, MMD-TSC constructs a weight agent to implement a dynamic and adaptive multi-objective weight calculation.

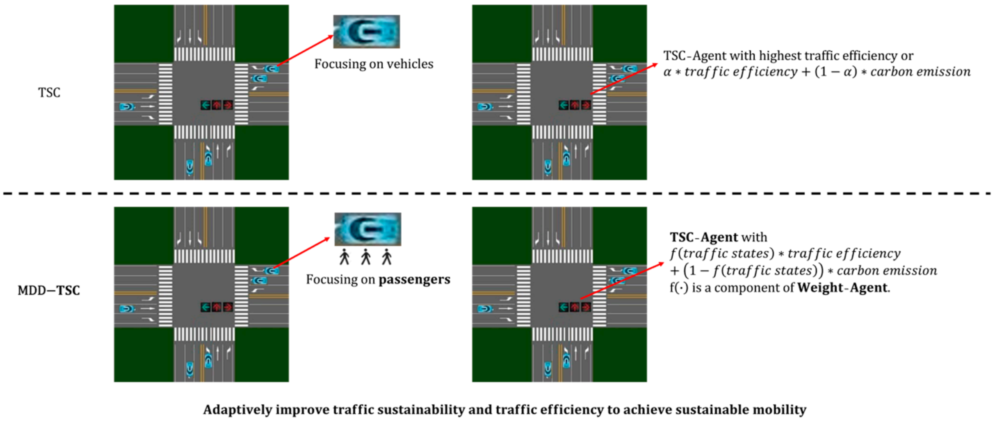

TSC
Focusing on vehicles
TSC-Agent with highest traffic efficiency or
$\alpha * traffic\ efficiency + (1-\alpha) * carbon\ emission$
MDD−TSC
Focusing on **passengers**
TSC-Agent with
$f(traffic\ states) * traffic\ efficiency$
$+ (1 - f(traffic\ states)) * carbon\ emission$
f(·) is a component of **Weight-Agent**.
Adaptively improve traffic sustainability and traffic efficiency to achieve sustainable mobility

Figure 1. Overview of MDD-TSC.

This paper makes the following main contributions:

- We propose an MMD-TSC model. The MMD-TSC model is a dynamic multi-objective optimization model capable of simultaneously optimizing traffic efficiency and energy saving, surpassing the limitations of existing methods that only focus on traffic efficiency.
- We introduce a dynamic weight adaptation mechanism in the MMD-TSC model through a meta-learning weight agent. The model adaptively adjusts the optimization weights of traffic efficiency and energy saving based on traffic states.
- We incorporate the per capita carbon emissions indicator as the energy indicator in the reward function design of the MMD-TSC model to show the development level of the sustainable transition.

The structure of the paper is as follows. Section 2 summarizes related works. In Section 3, it introduces the MMD-TSC background. Section 4 elaborates on the model structure and training algorithms in detail. Section 5 describes the experimental design, shows experiment results, and discusses the results. Section 6 draws the conclusions and proposes future works.

2. Related Work

Deep reinforcement learning (DRL), with its ability to recognize complex environments, can be applied to decision-making problems in dynamic environments. TSC aims to optimize traffic flow and improve efficiency by decreasing queue length and increasing velocity. The DRL's action is to determine which phases should be green. When selecting the action, it is necessary to analyze the complex road environment and the dynamic changes in traffic flow. Therefore, DRL can be applied to TSC. Chu et al. [3] constructed a 3D model of the traffic intersection for state acquisition, using the vehicle position images as the current state. They designed an end-to-end model, and the result is superior to other TSC methods. Lin et al. [12] constructed a lane-level graph to represent the traffic intersection to identify which lane had special situations. This microscopic design method helps represent the relationships between lanes and discover problematic lanes. Razack et al. [7] extracted the spatial and temporal characteristics. The results were better than other TSC models in terms of queue length and the number of waiting vehicles. Antes et al. proposed a hierarchical multiagent reinforcement learning approach for traffic signal control, where regional agents aggregated information and generated recommendations. The proposed method outperforms fixed-time and non-hierarchical RL-based approaches on a synthetic traffic grid [13]. Abdoos et al. proposed a hierarchical reinforcement learning approach to control traffic lights. The upper layer is responsible for predicting traffic flow based on global information, while the lower layer makes traffic light action decisions based on the prediction results and the conditions at the intersection. Experimental results show that this approach is feasible in a scenario involving multiple traffic lights at 16 intersections [14]. Traffic carbon emissions account for one-tenth of the total carbon emissions in China, so low carbon emissions is a goal of smart cities. Low carbon should have been the goal of TSC [15], but Koch et al. [10] found that maximum traffic flow and minimum carbon emissions are not necessarily conflicting goals. That means the two goals can be improved simultaneously. Most of the research considers carbon emissions as a secondary goal while focusing on traffic efficiency. Guo et al. [16] focused on the sustainability of Connected Autonomous Vehicles (CAVs) and designed a multi-agent system that controls both traffic signals and vehicles. The model can balance travel time and fuel use in a large-scale complex road network. It essentially focuses on vehicles rather than traffic flow. Reddy et al. [6] focused on potential route conflicts of lanes at intersections and designed lane-level intelligent TSC to improve throughput and low carbon performance. Górka et al. [17] the control model with priority, adjusts the traffic lights at intersections containing tram to reduce energy consumption, and verifies it on a real dataset using simulation methods. Kolat et al. [18] designed the standard deviation of the number of in-lane parking as a reward function to improve traffic efficiency while also improving sustainability. The electric vehicles with high velocity will have excessive energy consumption. Chen et al. [19] designed a centralized coordinated control method connecting vehicles and traffic signals to achieve a balance between energy consumption and traffic efficiency. Only one study focuses on optimizing carbon emissions through TSC. Kang et al. [8] built a dueling DQN to train TSC agents to minimize harmful gas emissions. The result shows the agent can reduce carbon emissions and improve traffic efficiency.

Fairness for passengers is a core concept of sustainability, so in addition to carbon emissions, fairness is an important component of sustainability. Research also focuses on fairness from perspectives of vehicle and road priority. Sun et al. [20] designed a reward function to avoid excessive waiting for passengers and vehicles on non-main roads, and a priority scheduling TSC method for emergency vehicles. The results increased the throughput of intersections, effectively solved traffic congestion, and ensured fairness. To ensure fairness, a cost function was used to prevent the traffic signal of a lane from being always red. Ye et al. [21] designed a hierarchical TSC model with vehicle fairness. Vehicle fairness is the ratio of waiting time to passing time. Kumar et al. [22] designed three TSC modes: fair mode, priority mode, and emergency mode. Each mode set vehicle priorities according to importance. Cui et al. [23] designed traffic-network fairness from the perspective of phase actuation fairness and network resource utilization fairness.

Traffic efficiency and energy saving are two different goals. For such multi-objective TSC, current research focuses on three methods: (1) Pareto, (2) weighted sum (with static weights, as well as dynamic weight represented by entropy according to traffic state), and (3) other algorithms to balance multiple objectives. Akyol et al. [24] used heuristic algorithms to calculate the Pareto solutions that minimize pedestrian waiting time and carbon emissions. The results showed that MOABC (Multi-objective Artificial Bee Colony) is better. Gong et al. [25] used a multi-objective reinforcement learning method to optimize TSC, aiming to ensure both efficiency and safety. The solution for multi-objective models is to use different models and parameter learning methods for different value functions. Saiki et al. [26] proposed a multi-strategy multi-objective reinforcement learning method. Zhang et al. [27] used NSGA-III (Nondominated Sorting Genetic Algorithm-III) to solve multi-objective TSC with goals of signal control delay and traffic capacity indices. Zhang et al. [28] designed a dynamic multi-objective reinforcement learning model to optimize traffic signals in terms of low-carbon, efficiency, etc. The entropy weight was used to calculate the weights under different states. Fang et al. [29] constructed a multi-agent multi-objective collaborative reinforcement model for traffic signals across the entire road network. The goals include traffic efficiency, safety, and network coordination. The weights are state coefficients. Reyad et al. [30] used a static weighting method to balance traffic efficiency and safety in TSC.

The current methods lack the process of weight optimization. This paper adopts meta learning to train the weight agent to calculate the weight based on traffic state. There are only two researchers studying meta-learning methods of TSC. Wang et al. [31] addressed the problem that TSC cannot effectively make use of the potential relations of spatial–temporal information based on a cooperative strategy. Meta learning is the Meta-Knowledge Learner in this paper. It is applied to identify traffic flow dynamics in the road network and adjust the parameter settings in the original spatio–temporal feature extraction. It is difficult for the intersection agent to fully capture the influence of the adjacent intersection, resulting in a poor agent effect. Zhu et al. [32] used meta-learning to replace the information of adjacent intersections and serve as an intrinsic motivation method to improve performance and stability.

3. Problem Definition

This section explains the existing approaches' limitations to promoting sustainable transition and proves two assumptions: (1) When defining a reward function, traffic efficiency differs from per capita carbon emissions. So, they cannot be directly substituted. (2) Dynamic weights for traffic efficiency and low carbon emission based on traffic state are necessary.

3.1. Limitations of Existing Approaches to Promote Sustainable Transition

There are some practical approaches to promote traffic transition, such as promoting shared mobility and public transportation. The goal of promoting shared mobility and public transportation is to increase the number of passengers per vehicle and use low-carbon vehicles [33]. Its essence is to reduce per capita carbon emissions by improving the utilization rate of carbon emissions through multi-passenger vehicles and low-carbon vehicles. However, this approach does not achieve its goal. On the one hand, ensuring punctuality on public transportation is challenging due to dynamic traffic efficiency. Low service levels will result in insufficient attractiveness for passengers. On the other hand, low traffic efficiency can lead to low-carbon vehicles not being low carbon.

Setting up dedicated lanes for specific vehicle types is an approach to promoting a sustainable traffic transition, such as setting bus lanes [33]. However, this method has limitations. First, there is insufficient road space because the additional lane allocation requires too many road resources. Second, allocating lanes and setting traffic regulations for different types of vehicles can only partially improve traffic efficiency. Finally, dividing lanes by vehicle type for all roads is unsuitable because of the complex traffic flow and traffic regulations.

In summary, promoting shared mobility, public transportation, and dedicated lanes has limitations and is not highly applicable in various scenarios. These methods fail to achieve a sustainable transition. Therefore, there is a need for a comprehensive traffic management solution that can effectively regulate mixed traffic flow, such as traffic signal control. The goal is to design a model that can adaptively balance traffic efficiency and traffic carbon emissions based on traffic states.

3.2. Traffic Efficiency vs. Per Capita Carbon Emissions

Most of the research on TSC to improve traffic efficiency often also reduces per capita carbon emissions [10]. However, traffic efficiency and low per capita carbon emissions are only partially interchangeable. Under different objectives, the actions taken by the TSC agent should be different for the same state. So, the optimal strategies targeting per capita carbon emissions and traffic efficiency are different. This means that as long as one action is different, the strategy is different.

Traffic state changes are Markov processes. The factors that influence traffic efficiency are vehicle distribution and velocity. The factor influencing carbon emissions needs to add vehicle type. The factor influencing per capita carbon emissions needs to add the number of people in each vehicle. According to the Bellman equation as Equation (1), actions with traffic efficiency, carbon emissions, and per capita carbon emissions as reward functions are expressed as Equation (2).

$$action = argmax_a\left(R(s,a) + \gamma \sum P\left(s' \mid s,a\right)V\left(s'\right)\right) \tag{1}$$

$$action' = argmax_a\left(R(s+\Delta s,a) + \gamma \sum P\left(s'+\Delta s' \mid s+\Delta s,a\right)V\left(s'+\Delta s'\right)\right) \tag{2}$$

V represents the state-value function, which estimates the sum of future rewards starting from a given state. S is the velocity value of the vehicle position at the traffic intersection, and Δs is a vehicle type and the number of people in the vehicle. *Action* equals *action*$'$ only if Δs does not affect R and V. That means Δs is constant. However, because Δs changes at any time, *action* equals *action*$'$ is unequal. Since s and Δs are independent of each other, it can be similarly concluded that the actions are not entirely the same. The conclusion indicates that when indicators are substituted for each other, different actions will be generated, leading to various strategies. Therefore, traffic efficiency and per capita carbon emissions are not interchangeable indicators.

3.3. Static Weights vs. Dynamic Weights

Static weights and dynamic weights refer to the fixed settings of static weights and dynamic adaptive settings of weights for traffic efficiency and per capita carbon emissions in the context of constantly changing traffic states [29,31]. If dynamic weights are unnecessary, static weights will be optimal in different states. Due to the large state space, this paper shows two typical scenarios. If the optimal weights in the two scenarios are different, it can be demonstrated that dynamic weights are essential.

Scenario 1: peak time. There are too many vehicles at the intersection. If the weight of sustainability is too large, it will increase congestion, contrary to traffic management objectives. Therefore, the weight should be appropriately reduced to alleviate congestion.

Scenario 2: off-peak time. There is very little traffic at the intersection. The traffic efficiency is very high. If the weight of traffic efficiency is too large, the objective becomes ineffective because the traffic velocity is already at its maximum. Therefore, the weight of sustainable mobility should be increased.

In summary, the above examples illustrate the limitations of static weights. Adaptive dynamic weights based on traffic states are necessary.

4. Models and Training Method

The TSC changes with the traffic state. DQN can recognize dynamic and complex traffic states, so this method is often applied to the TSC. In this paper, the agent for TSC is called the TSC agent, and its reward represents the long-term optimization of traffic efficiency and sustainability.

The weights of traffic efficiency and per capita carbon emissions are dynamic. For example, improving traffic efficiency in traffic congestion is more important, and when road loads are low, traffic sustainability is more important. The traffic state is the basis of dynamic weights. That is, the dynamic weight is based on the vehicle state at the intersection. It can be expressed by Equation (3).

$$\alpha = f(s) \tag{3}$$

The s is the vehicle state at the intersection, f is a function that translates the vehicle state to dynamic weights, and α is the dynamic weight. This paper designs and trains a neural network model to fit f. The model's input is the s, and the output is α. A is a critical component of the reward in the TSC agent and guides the training direction. Therefore, the TSC agent and f should be collaboratively trained to optimize the reward of the TSC agent. Since f is also trained by interacting with the environment, reinforcement learning algorithms should also be used as training. This agent is called a weight agent. Therefore, the model is a multi-agent model comprising the TSC agent and the weight agent. Two agents train together with the goal of an optimal TSC agent.

4.1. Model

4.1.1. TSC Agent

The TSC agent uses the DQN method to control traffic signals. The agent can control the traffic signal phases to maximize efficiency and per capita carbon emissions. The TSC agent is a DQN-based agent responsible for selecting the optimal traffic signal phase at each time step. The traffic signal is at a typical four-way intersection, where each road has four lanes: one left-turn lane, one right-turn lane, and two through lanes. The critical components of the TSC agent are as follows:

State: The state captures the vehicle position distribution and driving state at the intersection. Divide each lane into N small segments of fixed length.

For each segment i and each lane j, we record the following information:

- $v_{i,j}$: the average velocity of vehicles within the segment, normalized to [0, 1].
- $p_{i,j}$: the number of passengers in the vehicles within the segment.
- $c_{i,j}$: the one-hot encoded vehicle type distribution within the segment, where the number of vehicle types is C.

With M lanes and N segments, the state at time step t is a tensor s_t with shape (M, N, C + 2). Equation (4) represents the expression for s_t

$$s_t = \begin{bmatrix} [v_{0,0}, p_{0,0}, c_{0,0}] & \cdots \\ \ldots & [v_{M,N}, p_{M,N}, c_{M,N}] \end{bmatrix} \tag{4}$$

To incorporate temporal information, we stack the states from the past T time steps, resulting in a final state tensor s with shape (T, M, N, C + 2). Equation (5) represents the expression for s.

$$s = [s_1, s_2, \ldots, s_T] \tag{5}$$

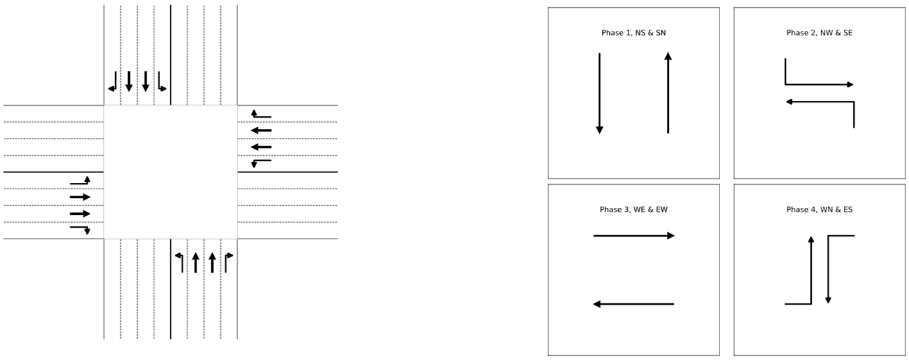
Action: The action space is discrete, representing the selection of traffic signal phases as Figure 2 shows.

- Phase 1: No—h–south through traffic.
- Phase 2: E—t–west through traffic.
- Phase 3: No—h–south left-turn traffic.
- Phase 4: E—t–west left-turn traffic.

At each time step, the TSC agent selects the optimal phase a_t based on the state s_t using an ϵ-greedy policy. The calculation method for a_t is given by Equation (6).

$$a_t = \begin{cases} argmax_a Q(s_t, a; \theta), & with\ probability\ 1 - \epsilon \\ random\ action, & with\ probability\ \epsilon \end{cases} \tag{6}$$

where $Q(s_t, a; \theta)$ is the Q-value of action in a given state s_t, parameterized by θ, and ϵ is the exploration rate.



(**a**)

(**b**)

Figure 2. Intersection layout and signal phase settings. (**a**) The intersection layout with through and left-turn lanes: One left-turn lane, two through lanes, and one right-turn lane. (**b**) Four signal phases: Phase 1 (NS and SN through traffic), Phase 2 (NW and SE left-turn traffic), Phase 3 (WE and EW left-turn traffic), and Phase 4 (WN and ES through traffic).

Reward: The reward function represents the long-term optimization of traffic efficiency and sustainability. Traffic efficiency is measured by the queue length, which is the ratio of vehicles with zero velocity at the intersection to the number of segments. A smaller queue length indicates higher traffic efficiency. Sustainability is measured by the average carbon emissions per person, reflecting fairness and low-carbon transportation. A smaller value indicates better sustainability. The reward r_t at time step t is defined as Equation (7).

$$r_t = -((1 - \alpha_t) \cdot queueLength_t + \alpha_t \cdot per\_capita\_carbon\_emissions_t) \tag{7}$$

where α_t is the dynamic weight for sustainability generated by the weight agent at time step t.

4.1.2. Weight Agent

The weight agent can generate the dynamic weight for balancing traffic efficiency and per capita carbon emissions in the TSC agent's reward function. The reason for choosing to use two agents instead of a single agent to complete this task is that the optimal balance between traffic efficiency and per capita carbon emissions depends on the current traffic state. Because the states are dynamic and complex, building another agent can focus on learning this mapping from traffic states to weights without interfering with the TSC agent's learning of the optimal control policy. By using different reinforcement learning algorithms that are more suitable for each task, weight generation is decoupled from control policy learning.

The collaborative training of the TSC agent and weight agent is a form of meta-learning. The weight agent learns to generate an adaptive reward function for the TSC agent based on the traffic state, helping the TSC agent learn a more effective control policy. The weight agent follows the actor–critic architecture. The critical components of the weight agent are as follows:

State: The state for the weight agent is the same as for the TSC agent, capturing the vehicle position, velocity, number of passengers, and vehicle type information at the intersection.

Action: The action of the weight agent is the dynamic weight α, which is a continuous value between zero and one.

Reward: The intuition behind the weight agent's reward design is that it should learn to generate weights that maximize the TSC agent's long-term reward. Therefore, we define the weight agent's reward as the TSC agent's reward.

This formulation encourages the weight agent to find the optimal balance between traffic efficiency and sustainability, minimizing the weighted sum of queue length and average carbon emissions. Figure 3 illustrates the computational process of the model. Both agents interact directly with the environment, and the weight agent provides the dynamic weight to the TSC agent.

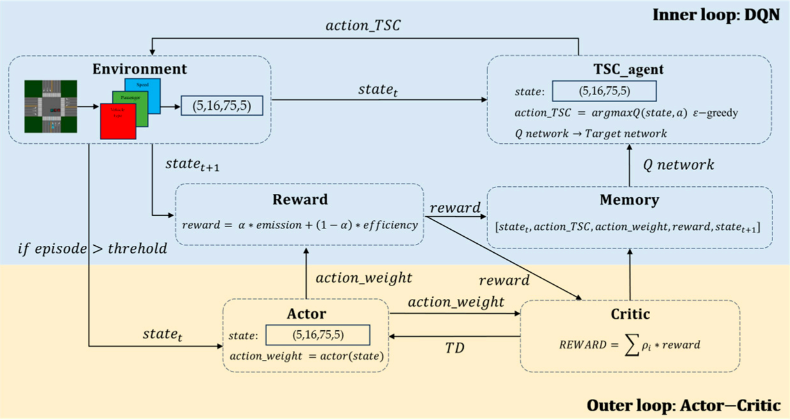


Figure 3. MDD-TSC model diagram. The model consists of two agents, the inner loop is the TSC agent and the outer loop is the weight agent. Two agents collaborate to calculate actions and rewards, and share memory information.

4.2. Training Process

The MMD-TSC model introduces a meta-learning approach to coordinate the training process of the TSC agent and the weight agent. The weight agent computes weights to guide the TSC agent towards a better control policy.

The training process involves the collaborative training of the TSC agent and the weight agent. At each time step, the TSC agent selects an action (traffic signal phase) based on the current state, and the weight agent generates the weight based on the same state. The environment transitions to the next state based on the selected action and returns the reward calculated using Equation (7). The transition tuple $(s_t, a_t, \alpha_t, r_t, s_{t+1})$ is stored in the replay buffer for both agents.

4.2.1. Training Algorithms

For the TSC agent, the training follows the DQN algorithm. The Q-network is updated by minimizing the temporal difference error between the predicted Q-values and the target Q-values. The target Q-values are calculated using the Bellman equation as Equation (8) shows.

$$Q_{target} = r_t + \gamma \max_{a'} Q(s_{t+1}, a'; \theta) \tag{8}$$

where γ is the discount factor and θ represents the parameters of the target Q-network.

For the weight agent, the training follows the Proximal Policy Optimization (PPO) algorithm. The actor network is updated by maximizing the clipped surrogate objective as Equation (9) shows.

$$L^{CLIP}(\phi) = \hat{E}_t\left[min\left(r_t(\phi)\hat{A}_t, clip(r_t(\phi), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right] \tag{9}$$

where $r_t(\phi) = \frac{\pi_\phi(a_t|s_t)}{\pi_{\phi_{old}}(a_t|s_t)}$ is the probability ratio, $\hat{A}_t$ is the estimated advantage, and ε is the clipping threshold. The critic network is updated by minimizing the mean squared error (MSE) between the predicted state values and the target values.

The critic network $V_\psi(s)$ parameterized by ψ estimates the state-value function, which predicts the expected cumulative reward starting from state s. The critic is trained to minimize the mean squared error between the predicted state values and the target values as Equation (10) shows.

$$L(\psi) = \frac{1}{B}\sum_{i=1}^{B}\left(V_\psi(s_i) - V_i^{target}\right)^2 \tag{10}$$

where B is the batch size, and V_i^{target} is the target value calculated using the Generalized Advantage Estimation (GAE). The GAE is a method for estimating the advantage function in policy gradient algorithms, which balances the trade-off between bias and variance by combining Monte Carlo estimation and Temporal Difference (TD) estimation. The GAE introduces a parameter λ to control the interpolation between the two estimation methods. In the proposed approach, the GAE is used to calculate the target values for updating the critic network, as shown in Equation (11):

$$V_i^{target} = \sum(\gamma\lambda)^t \delta_{i+t}^V \tag{11}$$

where $\delta_t^V = r_t + \gamma V_\psi(s_{t+1}) - V_\psi(s_t)$ is the TD residual, γ is the discount factor, and λ is the GAE parameter.

In summary, the weight agent learns to generate dynamic weights for balancing traffic efficiency and sustainability in the TSC agent's reward function. The actor–critic architecture and the PPO algorithm enable the weight agent to adapt to the dynamic traffic state and provide informative rewards for the TSC agent's learning.

4.2.2. Multi-Agent Collaborative Training

(1) Information Exchange and Shared Replay Buffer

The collaborative training of the TSC agent and the weight agent involves a continuous exchange of information between the two agents. The TSC agent relies on the weights generated by the weight agent to calculate its reward, while the weight agent's reward is based on the performance of the TSC agent. The transition tuples $(s_t, a_t, \alpha_t, r_t, s_{t+1})$ are stored in a shared replay buffer, allowing both agents to learn from the same experiments.

(2) Training Trigger

To ensure stable training, a training trigger is implemented for the weight agent. In the current implementation, the weight agent is trained once for every *threshold* step of TSC agent training. This allows the TSC agent to adapt to the new weights before the weight agent updates its policy. The training trigger can be adjusted based on the convergence of the TSC agent's reward. For example, if the TSC agent's reward converges, the frequency of the weight agent's training can be increased to further fine-tune the weights.

The appropriate value of the threshold depends on the specific problem and should be determined through empirical experimentation. If the *threshold* is too small, the weight agent will update its policy too frequently, causing instability in the TSC agent's learning process as it struggles to adapt to the rapidly changing weights. On the other hand, if the *threshold* is too large, the weight agent's policy updates will be too infrequent, leading to slow convergence and suboptimal performance.

Algorithm 1 presents the pseudocode for the collaborative training process of the TSC agent and the weight agent. The training process consists of M episodes, each with T time steps. At each time step, the TSC agent selects an action a_t using an ε-greedy policy based on its Q-network, while the weight agent generates a weight α_t using its actor network. The transition tuple $(s_t,\ a_t,\ \alpha_t,\ r_t,\ s_{t+1})$ is then stored in the shared replay buffer D. Every threshold step, the weight agent is trained for K epochs using the PPO algorithm, updating its actor and critic networks. Meanwhile, the TSC agent is trained at each time step by sampling a batch of transitions from D, computing the target Q-values, and updating its Q-network parameters. This collaborative training process enables both agents to learn from each other and adapt to the dynamic traffic environment.

Algorithm 1 Collaborative training of the TSC agent and the weight agent.

```
Initialize Q-network parameters $\theta$ for TSC agent
Initialize target Q-network parameters $\theta_t arget = \theta$ for TSC agent
Initialize actor network parameters $\varphi$ for weight agent
Initialize critic network parameters $\psi$ for weight agent
Initialize shared replay buffer $D$
for episode = 1 to M do the following
    Initialize state s
    for t = 1 to T do the following
        Select action  $a_t$ for TSC agent using $\varepsilon -$ greedy policy based on $Q(s, a; \theta)$
        Generate weight $\alpha_t$ for weight agent using actor network $\pi(a|s; \varphi)$
        Execute action $a_t,$ observe reward r_t and next state $s_{t+1}$
        Store transition $(s_t, a_t, \alpha_t, r_t, s_{t+1})$ in the shared replay buffer $D$
        if t % threshold == 0 then
            for epoch = 1 to K do the following
                Sample a batch of transitions $(s_i, a_i, \alpha_i, r_i, s_{i+1})$ from D
                Compute probability ratio $r_t(\varphi)$ and estimated advantage $_t$
                Compute clipped surrogate objective $L^{CLIP}(\phi)$ for the Weight agent
                Compute critic loss $L(\psi)$ for weight agent
                Update actor network parameters $\varphi$ for weight agent
                Update critic network parameters $\psi$ for weight agent
            end for
        end if
        Sample a batch of transitions $(s_i, a_i, \alpha_i, r_i, s_{i+1})$ from D
        Compute target Q-values using Equation (8) for TSC agent
        Compute Q-network loss for TSC agent Update Q-network parameters $\theta$ for TSC agent
        Update target Q-network parameters $\theta_{target} = \theta$ for TSC agent
        $s_t = s_{t+1}$
    end for
end for
```

5. Experiments

This section shows the results of training the MMD-TSC model and controlling the traffic flow at intersections, and evaluates the traffic conditions in the given traffic flow using the Simulation of Urban Mobility (SUMO) v.1.20.0.

5.1. Experiments Goals

In this section, we present an experimental evaluation of the model proposed in this paper. The goals of our study are as follows:

(1) To demonstrate the improvement of the sustainable transition, and to compare the performance of MAM-TSC with other weight design methods, such as single-objective TSC and multi-objective TSC with static weights.

(2) To evaluate the influence of different traffic state segments on dynamic weights.

5.2. Experimental Setting

5.2.1. Baseline

In this section, MMD-TSC is compared with four other TSC methods. One comparison is with single-objective TSC methods, including maximum traffic efficiency and minimum per capita carbon emissions. The other comparison is with state weighting methods.

- FT-TSC (fixed time): Traffic lights control traffic flow by setting periodic signal changes. This is a common baseline.
- CEP-TSC (carbon emissions priority): The reward function is to achieve maximum cumulative traffic efficiency. Research aimed at energy saving generally targets the minimization of harmful gas emissions [2,8]. In this paper, minimizing total carbon dioxide emissions represents this research.
- CEQB-TSC (carbon emissions and queue length balance): Traffic efficiency and per capita carbon emissions weights are static parameters in the reward function. Set this parameter to equal, that is, 1:1. Research on the multi-objective optimization of traffic efficiency and energy saving typically uses fixed weights [25,26,29,30]. In this paper, a 1:1 goal ratio serves as a representation of this research.
- QLP-TSC (queue length priority): The reward function is to achieve the minimum queue length. Research targeting maximum traffic efficiency often focuses on minimizing queue length or travel time [1,7]. In this paper, minimizing queue length is used as a representative goal for this research.

5.2.2. Simulation Parameters

The experiments in this study are conducted using the SUMO, an open-source, microscopic, and multi-modal traffic simulation tool. The SUMO allows for the creation of realistic traffic scenarios, including road networks, traffic demand, and vehicle behavior. It supports the definition of traffic signal systems, which can be controlled using the TraCI (Traffic Control Interface) API. This API enables real-time interaction between the simulation and external control algorithms, making the SUMO suitable for evaluating the proposed multi-agent reinforcement learning approach for adaptive traffic signal control.

The corresponding SUMO simulation parameters are shown in Table 1.

Table 1. Sumo simulation parameters in experiments.

Title 1	Title 2
Steps	800
The number of vehicles	1000
Green light duration	10
Yellow light duration	4

These parameters constitute a representative set of simulation settings that effectively capture the dynamic evolution of traffic conditions and the formation of congestion. The simulation is conducted on an empty road network with a continuous arrival of vehicles, resulting in a gradual increase in traffic flow throughout the simulation period.

The simulation duration is set to 800 steps, which provides sufficient time to observe the transition of traffic conditions from free-flow to congested states. This extended simulation period allows for a comprehensive analysis of the traffic dynamics and the impact of increasing traffic volume on road performance. With a total of 1000 vehicles, the simulation scale is representative of a realistic traffic scenario, enabling the evaluation of traffic management strategies under conditions that closely resemble real-world traffic patterns. The 1000 vehicles consist of cars, buses, and EVs. The number of passengers in each vehicle is a random number up to its capacity, with cars carrying 1–4 people and buses carrying 1–10 people. In addition to the difference in passenger capacity, the type of vehicle also affects its speed and carbon emissions.

The traffic light settings, with a green light duration of ten and a yellow light duration of four, are consistent with commonly observed values in real-world traffic signal configurations. These settings play a crucial role in the simulation, as they contribute to the accumulation of vehicles at intersections, leading to the formation of congestion in the latter half of the simulation period.

5.3. Results

5.3.1. Experiment Results

Figure 4 illustrates the relationship between crucial traffic indicators and time during the traffic simulation process. The critical traffic indicators are average queue length and vehicle waiting time. Shorter queue lengths and waiting times indicate that vehicles can pass through intersections more quickly, improving overall traffic efficiency and reflecting better traffic signal regulation.

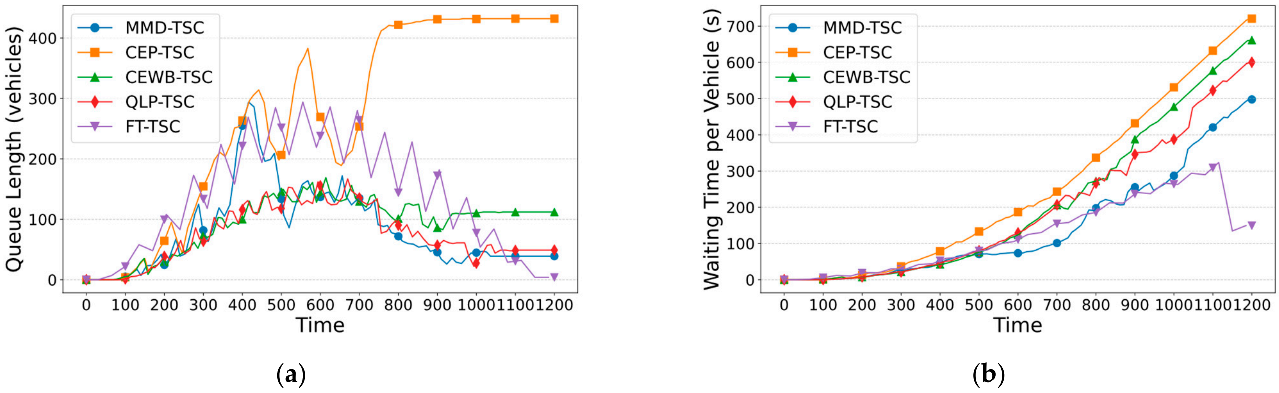


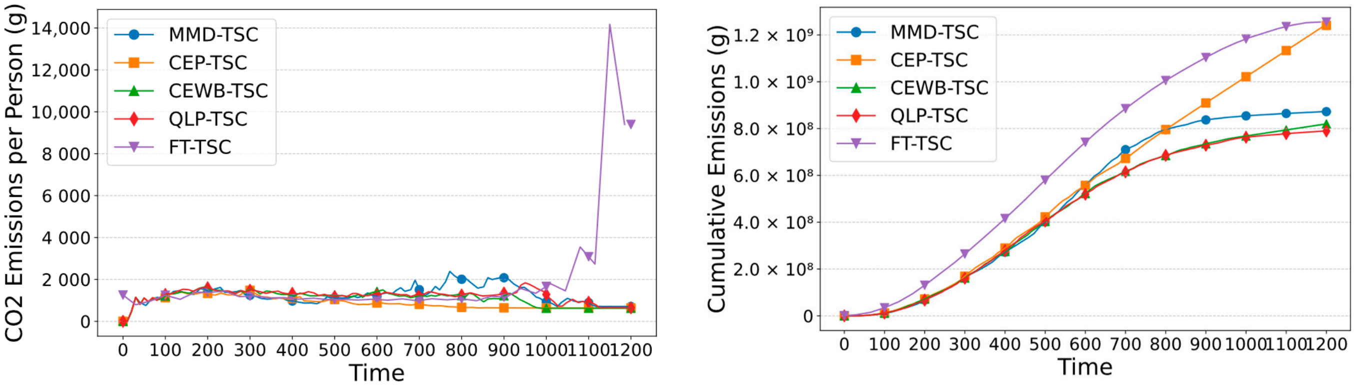
Figure 4. Comparison of sustainability indicators among different TSC models over time. (**a**) Relationship between per capita carbon emissions and simulation time for each model. (**b**) Relationship between cumulative carbon emissions and simulation time for each model.

Figure 4a shows that MMD-TSC performs the best at most time points. At the end of the simulation, the queue length of MMD-TSC can be reduced to a minimal value, indicating that this signal light regulation method can effectively manage vehicle traffic. Compared to QLP-TSC and FT-TSC, MMD-TSC outperforms QLP-TSC at most time points, except for the time range of 300–500. Compared to CEWB-TSC, MMD-TSC performs better in the later stages because dynamic weights can better adapt to various traffic states to efficiently control traffic flow through intersections.

In Figure 4b, the performance of other models varies based on the weight assigned to traffic efficiency, with higher weights generally leading to greater efficiency. FT-TSC outperforms QLP-TSC. Although QLP-TSC aims to maximize traffic efficiency, the waiting time indicator may not be the best since the model's objective is queue length. MMD-TSC has the best overall performance, with its line at the bottom. Adaptive weight adjustment enables MMD-TSC to improve not only on a single objective but also to comprehensively enhance traffic efficiency.

In summary, MMD-TSC performs better than other models in reducing queue lengths and waiting times throughout the simulation process, achieving higher overall traffic efficiency. This highlights the advantage of MMD-TSC in dynamically adjusting signal lights to adapt to changes in traffic states.

Figure 5 illustrates the relationship between sustainability indicators and simulation time throughout the traffic simulation process. The critical sustainability indicators are per capita carbon emissions and cumulative carbon emissions. Per capita carbon emissions are a microscopic indicator that reflects an individual's sustainability. In contrast, cumulative carbon emissions are a macroscopic indicator reflecting all vehicles' overall carbon emissions effect.



(**a**) (**b**)

Figure 5. Comparison of traffic indicators among different TSC models over time. (**a**) Relationship between average queue length and simulation time for each model. (**b**) Relationship between average vehicle waiting time and simulation time for each model.

In Figure 5a, the CEP-TSC model performs the best because it aims to minimize per capita carbon emissions. The per capita carbon emissions of the MMD-TSC model are slightly higher than those of the CEP-TSC model only in the time range of 650–800.

In Figure 5b, QLP-TSC and CEWB-TSC are the best models for total carbon emissions. Although CEP is the model aimed at optimal sustainability, per capita carbon emissions are a microscopic indicator and cannot represent the macroscopic total carbon emissions. The cumulative emissions of MMD-TSC are slightly higher in the latter half compared to QLP-TSC and CEWB-TSC models.

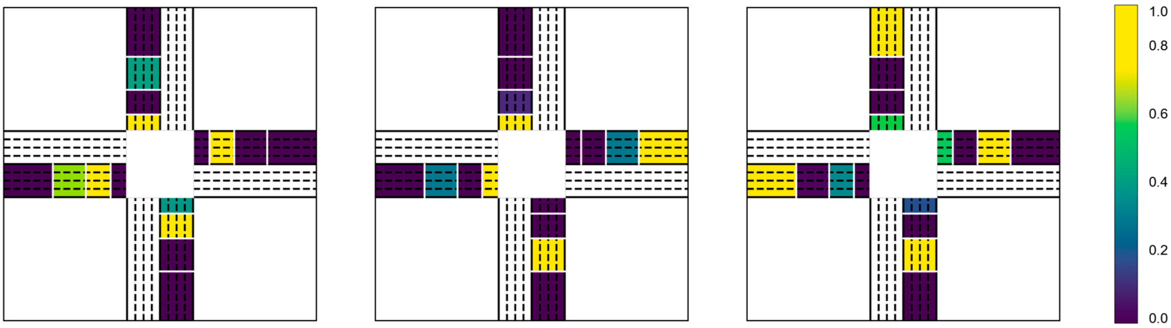
From the above analysis, MMD-TSC can better promote a sustainable transition. In most cases, MMD-TSC outperforms other models in terms of traffic efficiency and is also close to the best model in terms of sustainability.

5.3.2. Spatial Influence Analysis on Actor of Weight Agent

To investigate which spatial locations in the traffic state have a more significant influence on the weights and to better explain the actor model, permutation importance is used to examine the impact of traffic environment values in different regions on the weights.

Divide the road length into proportions of 2:3:4:6, and examine the influence of each section's values on the results when they vary randomly. The data used are a sample of traffic states, including uniformly collected data from the entire simulation process, totaling 100 samples.

This set of figures shows the influence of different regions at the traffic intersection on the weights. Figure 6a–c are graphs for three time periods, with lighter colors indicating a greater degree of influence. For time periods one, two, and three, the regions with a high degree of influence shift backward successively, which aligns with the common sense of vehicle flow approaching the intersection and demonstrates the reasonableness of the model. Furthermore, with each lane approaching the intersection, only one region usually has a significantly high degree of influence. This indicates that, even though vehicles are in motion, only one region has a notable impact.



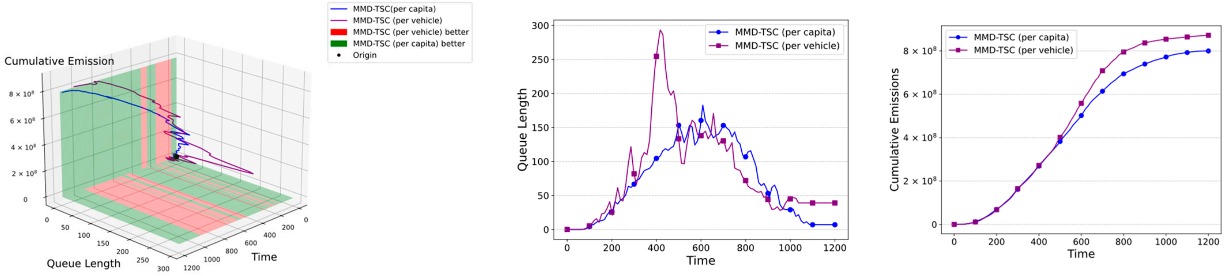
(**a**) (**b**) (**c**)

Figure 6. Heatmaps of the influence of different spatial locations on weights at three time periods: (**a**) Time period 1, (**b**) Time period 2, and (**c**) Time period 3. The influence is standardized to a range of 0–1, and the larger the value, the greater the degree of influence. Yellow has the greatest impact here.

5.3.3. Comparative Analysis of Per Capita and Per Vehicle Carbon Emissions

Most studies focus on vehicle carbon emissions without passengers. To demonstrate the necessity of incorporating vehicle occupancy into the model, compare the goals of per capita carbon emissions with per vehicle carbon emissions. Then, compare the simulation results of the minimum per capita and minimum per vehicle carbon emissions as sustainability indicators. Queue length is the traffic state evaluation indicator, and cumulative carbon emissions are the sustainability evaluation indicator.

Figure 7 is a three-dimensional graph with dimensions of time, cumulative carbon emissions, and queue length. MMD-TSC (per vehicle) represents the model with per vehicle carbon emissions as the sustainability objective, while MMD-TSC (per capita) represents the model with per capita carbon emissions. As simulation time increases, cumulative carbon emissions continuously increase. Along the time axis, the green part indicates that MMD-TSC (per capita) performs better. At the beginning of the simulation, the cumulative carbon emissions indicator is red, but from Figure 7c, it can be seen that there is little difference between the two indicators. Only during the time period of 600–800, MMD-TSC (per vehicle) has slightly better traffic efficiency. Therefore, except for the time when the number of passing vehicles is at its maximum, MMD-TSC (per capita) is Pareto superior to MMD-TSC (per vehicle). This suggests that, except for the case of excessive vehicle congestion, MMD-TSC (per capita) can more effectively regulate traffic intersections for sustainable transition. The reason is that when there are too many vehicles, excessively weakening vehicle attributes will affect the improvement of traffic efficiency.

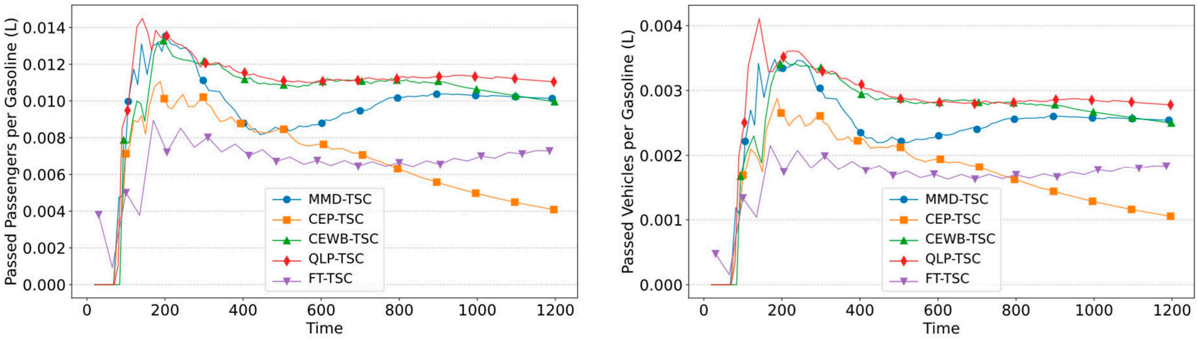


(**a**) (**b**) (**c**)

Figure 7. The degree of influence of different locations in the intersection on the dynamic weights is standardized to a range of 0–1. Larger values indicate a greater degree of influence. Yellow indicates the greatest influence. (**a**) A 3D plot shows the relationship between simulation time, queue length, and cumulative carbon emissions in a single graph. To provide a clearer view of the relationship between the two indicators and simulation time, (**b**,**c**) are line graphs showing the relationship between queue length, cumulative carbon emissions, and simulation time.

5.3.4. Comparative Analysis of Carbon Emission Utilization

Figure 8 illustrates the gasoline consumption utilization of different TSC models. The horizontal axis represents the simulation time, while the vertical axis represents the number of vehicles and passengers that can drive per liter of gasoline. As the simulation time increases, the gasoline consumption utilization of all models shows a trend of rising first and then stabilizing. It can be seen that MMD-TSC outperforms FT-TSC by about 35% in terms of gasoline consumption utilization. This indicates that MMD-TSC is more effective in optimizing traffic signal control strategies, allowing more vehicles and passengers to pass through with the same amount of gasoline consumption.



(**a**) (**b**)

Figure 8. Comparative analysis of energy consumption utilization among different TSC models. (**a**) Relationship between passed passengers per gasoline for each model. (**b**) Relationship between passed vehicles per gasoline for each model.

6. Discussion

Reducing carbon emissions can be achieved while improving traffic efficiency. The coordinated training mode of MDD-TSC has a better effect. In the first half of the simulation, when traffic flow increases, but congestion has not yet occurred, the total carbon emissions of MMD-TSC are lower. In the second half of the simulation, when vehicle congestion occurs, MMD-TSC has better traffic efficiency and can effectively alleviate congestion pressure. Therefore, MMD-TSC can better adapt to complex traffic environments and provide solutions that meet traffic management requirements.

Compared to FT-TSC, the energy consumption utilization of the MMD-TSC model has improved by 35%. With the same amount of carbon emissions, more vehicles and passengers can pass through an intersection. However, MMD-TSC is slightly less efficient in terms of gasoline consumption compared to QLP-TSC. On the one hand, the number of passengers, vehicle type, and vehicle velocity can enhance the analysis of traffic states. This enhancement positively impacts traffic efficiency. On the other hand, the QLP-TSC model focuses on reducing queue length to improve traffic efficiency. This approach will result in frequent changes to traffic signals, which can slightly increase gasoline consumption in vehicles. Furthermore, the waiting time under MMD-TSC is consistently better than that of other models. The conclusion is that, compared to QLP-TSC, the slightly lower gasoline consumption of MMD-TSC contributes to achieving more effective traffic management.

7. Conclusions

This paper investigates the problem of current TSC models having single objectives and multi-objective models only having static weights, making it difficult for the models to achieve a sustainable transition. The challenge lies in constructing a multi-objective learning model with dynamic weights and how to train the parameters. We propose the MMD-TSC model, which realizes its functionality through the interaction of the TSC agent and the weight agent, and alternately trains the two agents using the meta-learning approach. Experimental results show that the MMD-agent model achieves higher traffic efficiency; in non-congested states, it has lower per capita carbon emission, and in congested states, it can better ensure traffic efficiency and dynamically meet traffic management requirements.

TSC models that aim to optimize traffic efficiency can reduce carbon emissions while improving traffic efficiency, but they cannot lower emissions to the lowest possible level. Models with static parameters as multi-objective reward functions can better reduce carbon emissions, but the improvement in traffic efficiency is insufficient. Moreover, they cannot identify traffic states, making traffic efficiency not the primary objective even when the road is congested, which does not align with traffic management requirements. The MMD-TSC model proposed in this paper can dynamically identify traffic states, ensuring both traffic efficiency and energy saving, thereby achieving sustainable transition. Additionally, we find that the main factor influencing the weights is a specific region of the road.

Furthermore, this paper analyzes the gasoline consumption utilization of different TSC models. The results show that the MMD-TSC model reduces gasoline consumption by approximately 35% compared to the FT-TSC model. Although the gasoline consumption of the MMD-TSC model is slightly greater than that of the QLP-TSC model, it achieves more effective traffic management objectives.

Our work can be extended in multiple directions. On one hand, the TSC optimization problem for single intersections can be extended to multi-intersection problems and can be refined to lane-level navigation. The influence of pedestrians or bicyclists would also be important. On the other hand, the application of meta-learning in TSC can be explored to improve control precision. Future research can further investigate optimal TSC under various variable traffic parameters, such as green light duration and signal light sequence.

Author Contributions: Conceptualization, B.W. and J.S.; Methodology, Y.Z. (Yuqi Zhang), Y.Z. (Yingying Zhou), and J.S.; Software, Y.Z. (Yuqi Zhang) and Y.Z. (Yingying Zhou); Validation, J.S. and Y.Z. (Yuqi Zhang); Formal analysis, Y.Z. (Yuqi Zhang); Data curation, Y.Z. (Yuqi Zhang); Writing—original draft, Y.Z. (Yuqi Zhang); Writing—review and editing, J.S.; Visualization, Y.Z. (Yingying Zhou); Supervision, B.W. and J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Mora, L.; Deakin, M.; Zhang, X.; Batty, M.; de Jong, M.; Santi, P.; Appio, F.P. Assembling Sustainable Smart City Transitions: An Interdisciplinary Theoretical Perspective. *J. Urban Technol.* **2020**, *28*, 1–27. [CrossRef]
2. Zhang, Y.; Wang, H.; Wang, X. Towards Fairness-Aware Crowd Management System and Surge Prevention in Smart Cities. In Proceedings of the 2024 IEEE Workshop on Design Automation for CPS and IoT (DESTION), Hong Kong, China, 13–14 May 2023; pp. 46–54.
3. Chu, K.-F.; Lam, A.Y.S.; Li, V.O.K. Traffic Signal Control Using End-to-End Off-Policy Deep Reinforcement Learning. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 7184–7195. [CrossRef]
4. Zhang, Y.; Wang, H.; Wang, X. Research on the improvement of transportation efficiency of smart city by traffic visualization based on pattern recognition. *Neural Comput. Appl.* **2022**, *35*, 2211–2224. [CrossRef]
5. Farazi, N.P.; Zou, B.; Ahamed, T.; Barua, L. Deep reinforcement learning in transportation research: A review. *Transp. Res. Interdiscip. Perspect.* **2021**, *11*, 100425. [CrossRef]
6. Reddy, R.; Almeida, L.; Gaitán, M.G.; Santos, P.M.; Tovar, E. Synchronous Management of Mixed Traffic at Signalized Intersections Toward Sustainable Road Transportation. *IEEE Access* **2023**, *11*, 64928–64940. [CrossRef]
7. Razack, A.J.; Ajith, V.; Gupta, R. A Deep Reinforcement Learning Approach to Traffic Signal Control. In Proceedings of the 2021 IEEE Conference on Technologies for Sustainability (SusTech), Kollam, India, 8–9 August 2021.
8. Kang, L.; Huang, H.; Lu, W.; Liu, L. A Dueling Deep Q-Network method for low-carbon traffic signal control. *Appl. Soft Comput.* **2023**, *141*, 110304. [CrossRef]
9. Ma, D.; Li, S.; Han, B.; Zhang, Y. A decentralized model predictive traffic signal control method with fixed phase sequence for urban networks. *J. Intell. Transp. Syst.* **2020**, *25*, 455–468. [CrossRef]

10. Koch, L.; Brinkmann, T.; Wegener, M.; Badalian, K.; Andert, J. Adaptive Traffic Light Control with Deep Reinforcement Learning: An Evaluation of Traffic Flow and Energy Consumption. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 15066–15076. [CrossRef]
11. Huttunen, S.; Tykkyläinen, R.; Kaljonen, M.; Kortetmäki, T.; Paloviita, A. Framing just transition: The case of sustainable food system transition in Finland. *Environ. Policy Gov.* **2024**, *34*, 463–475. [CrossRef]
12. Lin, W.-Y.; Song, Y.-Z.; Ruan, B.-K.; Shuai, H.-H.; Shen, C.-Y.; Wang, L.-C.; Li, Y.-H. Temporal Difference-Aware Graph Convolutional Reinforcement Learning for Multi-Intersection Traffic Signal Control. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 327–337. [CrossRef]
13. Antes, T.d.O.; Bazzan, A.L.; Tavares, A.R. Information upwards, recommendation downwards: Reinforcement learning with hierarchy for traffic signal control. *Procedia Comput. Sci.* **2022**, *201*, 24–31. [CrossRef]
14. Abdoos, M.; Bazzan, A.L. Hierarchical Traffic Signal Optimization Using Reinforcement Learning and Traffic Prediction with Long-Short Term Memory. *Expert Syst. Appl.* **2021**, *171*, 114580. [CrossRef]
15. Alshayeb, S.; Stevanovic, A.; Mitrovic, N.; Espino, E. Traffic Signal Optimization to Improve Sustainability: A Literature Review. *Energies* **2022**, *15*, 8452. [CrossRef]
16. Guo, J.; Cheng, L.; Wang, S. CoTV: Cooperative control for traffic light signals and connected autonomous vehicles using deep reinforcement learning. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 10501–10512. [CrossRef]
17. Górka, A.; Czerepicki, A.; Krukowicz, T. The Impact of Priority in Coordinated Traffic Lights on Tram Energy Consumption. *Energies* **2024**, *17*, 520. [CrossRef]
18. Kolat, M.; Kővári, B.; Bécsi, T.; Aradi, S. Multi-Agent Reinforcement Learning for Traffic Signal Control: A Cooperative Approach. *Sustainability* **2023**, *15*, 3479. [CrossRef]
19. Chen, H.; Wu, F.; Qiu, T.Z. Achieving Energy-Efficient and Travel Time-Optimized Trajectory and Signal Control for CAEVs. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 10246–10259. [CrossRef]
20. Sun, G.; Qi, R.; Liu, Y.; Xu, F. A dynamic traffic signal scheduling system based on improved greedy algorithm. *PLoS ONE* **2024**, *19*, e0298417. [CrossRef]
21. Ye, Y.; Ding, J.; Wang, T.; Zhou, J.; Wei, X.; Chen, M. FairLight: Fairness-Aware Autonomous Traffic Signal Control With Hierarchical Action Space. *IEEE Trans. Comput. Des. Integr. Circuits Syst.* **2023**, *42*, 2434–2446. [CrossRef]
22. Kumar, N.; Rahman, S.S.; Dhakad, N. Fuzzy Inference Enabled Deep Reinforcement Learning-Based Traffic Light Control for Intelligent Transportation System. *IEEE Trans. Intell. Transp. Syst.* **2020**, *22*, 4919–4928. [CrossRef]
23. Cui, S.; Xue, Y.; Gao, K.; Wang, K.; Yu, B.; Qu, X. Delay-throughput tradeoffs for signalized networks with finite queue capacity. *Transp. Res. Part B Methodol.* **2024**, *180*, 102876. [CrossRef]
24. Akyol, G.; Göncü, S.; Silgu, M.A. Multi-objective Optimization Framework for Trade-Off Among Pedestrian Delays and Vehicular Emissions at Signal-Controlled Intersections. *Arab. J. Sci. Eng.* **2024**, *49*, 14117–14130. [CrossRef]
25. Gong, Y.; Abdel-Aty, M.; Yuan, J.; Cai, Q. Multi-Objective reinforcement learning approach for improving safety at intersections with adaptive traffic signal control. *Accid. Anal. Prev.* **2020**, *144*, 105655. [CrossRef] [PubMed]
26. Saiki, T.; Arai, S. Flexible Traffic Signal Control via Multi-Objective Reinforcement Learning. *IEEE Access* **2023**, *11*, 75875–75883. [CrossRef]
27. Zhang, X.; Fan, X.; Yu, S.; Shan, A.; Men, R. Multi-Objective Optimization Method for Signalized Intersections in Intelligent Traffic Network. *Sensors* **2023**, *23*, 6303. [CrossRef]
28. Zhang, G.; Chang, F.; Jin, J.; Yang, F.; Huang, H. Multi-objective deep reinforcement learning approach for adaptive traffic signal control system with concurrent optimization of safety, efficiency, and decarbonization at intersections. *Accid. Anal. Prev.* **2024**, *199*, 107451. [CrossRef]
29. Fang, J.; You, Y.; Xu, M.; Wang, J.; Cai, S. Multi-Objective Traffic Signal Control Using Network-Wide Agent Coordinated Reinforcement Learning. *Expert Syst. Appl.* **2023**, *229*, 120535. [CrossRef]
30. Reyad, P.; Sayed, T. Real-Time multi-objective optimization of safety and mobility at signalized intersections. *Transp. B Transp. Dyn.* **2022**, *11*, 847–868. [CrossRef]
31. Wang, M.; Wu, L.; Li, M.; Wu, D.; Shi, X.; Ma, C. Meta-learning based spatial-temporal graph attention network for traffic signal control. *Knowl.-Based Syst.* **2022**, *250*, 109166. [CrossRef]
32. Zhu, L.; Peng, P.; Lu, Z.; Tian, Y. MetaVIM: Meta Variationally Intrinsic Motivated Reinforcement Learning for Decentralized Traffic Signal Control. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 11570–11584. [CrossRef]
33. Zhang, X.; Zhong, S.; Ling, S.; Jia, N.; Qi, H.; He, Z. How to promote the transition from solo driving to mobility services delivery? An empirical study focusing on ridesharing. *Transp. Policy* **2022**, *129*, 176–187. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.