Received: September 19, 2025. Revised: October 15, 2025. 515

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

A Multi-objective Deep Reinforcement Learning for Adaptive Traffic Signal Control with Curriculum Reward Shaping

Hendra Marcos[1,3]* **Rahmat Gernowo[1]** **Adi Wibowo[2]** **Imam Tahyudin[4]**

[1]*Doctoral Program of Information System, Postgraduate School, Universitas Diponegoro, Indonesia*
[2]*Department of Informatics, Faculty of Science and Mathematics, Universitas Diponegoro, Semarang, Indonesia*
[3]*Department of Informatics, Universitas Amikom Purwokerto, Indonesia*
[4]*Magister of Computer Science, Universitas Amikom Purwokerto, Indonesia*
* Corresponding author"s Email: hendra.marcos@amikompurwokerto.ac.id

Abstract: This research proposes a multi-objective deep reinforcement learning (MODRL) framework for adaptive traffic signal control to address the limitations of conventional systems. The framework uses a Dueling Double Deep Q-Network (D3QN) agent to manage volatile urban traffic. A key contribution is the integration of a multi-objective reward function that balances traffic efficiency (delay reduction) and environmental sustainability (fuel consumption minimization). We also incorporate a curriculum learning strategy to mitigate the cold-start problem and prevent convergence to suboptimal policies. The agent first learns foundational traffic management before tackling complex multi-objective trade-offs. Evaluated in the SUMO simulation environment under a peak-hour scenario, the proposed framework, named MO-CL-D3QN, demonstrates significant performance improvements over conventional and standard DRL baselines. Specifically, it reduced the average vehicle delay by up to 48% compared to a fixed-time controller and simultaneously lowered fuel consumption by over 21% compared to a single-objective DRL agent, validating its effectiveness as a holistic and robust solution for intelligent traffic management.

Keywords: Traffic signal control, Deep reinforcement learning, Dueling DDQN, Multi-objective learning, Curriculum-based reward shaping

1. Introduction

The efficacy of modern urban transportation networks is critically challenged by the unabated growth in traffic volume, precipitating severe congestion, air quality degradation, and inefficient energy consumption [1, 2]. A primary contributor to this issue is the inadequacy of the conventional traffic signal control paradigm, which predominantly relies on fixed-time strategies ill-suited to the dynamic and stochastic nature of traffic flow. Consequently, the simultaneous optimization of multiple objectives-such as efficiency, safety, and operational stability-at intersections remains a formidable challenge [1-3]. In response, Adaptive Traffic Signal Control (ATSC) powered by Deep Reinforcement Learning (DRL) has emerged as a dominant paradigm, offering a data-driven approach to learn adaptive control policies directly from environmental interactions [6, 7].

Nevertheless, the practical application of DRL in this domain is confronted with significant, well-documented challenges [6-8]. A principal limitation of existing research is a myopic focus on singular performance metrics, where control strategies primarily target operational efficiency (e.g., waiting time or queue length) while often neglecting crucial secondary objectives like traffic safety [3, 9]. Furthermore, even when environmental goals such as minimizing CO_2 emissions are considered, the formulation of a suitable reward function is a non-trivial and open research problem known as reward design [6]. Recent studies demonstrate that using a direct CO_2 emission metric as a reward signal can be inefficient for agent training, often leading to suboptimal policies due to numerous factors beyond

the agent's control [10, 11]. This underscores the necessity of careful proxy reward engineering [6]. Compounding this conceptual gap is a technical one: fundamental DRL algorithms like the standard Deep Q-Network (DQN) are known to be susceptible to convergence instability and value overestimation bias, which hinders their performance in complex, real-world traffic scenarios [12, 13].

This study addresses these shortcomings by proposing a holistic framework that targets the dual objectives of operational performance and ecological sustainability [14]. To overcome the instability of traditional DRL algorithms, we employ the Dueling Double Deep Q-Network (D3QN) architecture, an advanced method recognized for achieving stable and robust performance under complex traffic conditions [13]. Furthermore, to navigate the intricate reward landscape, we introduce a curriculum-based training strategy. This approach is specifically designed to handle complex reward functions with multiple, potentially adversarial terms by guiding the agent to first master a simpler, foundational objective before optimizing the full, composite reward [15, 16]. While these components have been explored independently, their synergistic integration to solve the dual challenge of multi-objective reward complexity and algorithmic instability represents a novel contribution. By integrating a robust DRL architecture with a principled training curriculum, our framework is designed to effectively learn a policy that balances conflicting objectives without succumbing to reward hacking or local optima [8, 16].

Accordingly, this research aims to design and evaluate an ATSC framework based on Multi-Objective Deep Reinforcement Learning that simultaneously optimizes two primary pillars: operational performance (measured by travel time, queue length, and throughput) and ecological sustainability (measured via the proxy metric of fuel consumption). The fundamental contribution of this research is not merely the application of existing methods, but the proposal of a cohesive framework where these components work in concert to produce qualitatively superior outcomes. The contributions are threefold: (1) It proposes a holistic dual-objective optimization model that explicitly harmonizes operational and ecological goals, directly addressing the multi-objective challenge identified in recent literature [4], and navigating the complex reward design space highlighted by Schumacher et al. (2023). (2) It implements the Dueling Double DQN architecture, leveraging its proven superior stability to overcome the known convergence issues of standard DQN within the traffic control domain [13]. (3) It introduces an innovative curriculum learning approach, framed as a solution for complex reward functions, demonstrating that such a curriculum is a critical enabler for stable convergence in a multi-objective DRL context, a synergy that has not been extensively explored in prior ATSC research [16].

The remainder of this paper is organized as follows. Section 2 presents a review of the relevant literature on DRL for ATSC and reward design methodologies. Section 3 details the research methodology, including the model architecture and the multi-objective reward formulation. Section 4 reports the experimental results and provides a comparative analysis against established baselines. Finally, Section 5 concludes with the key findings and outlines future research directions, including the exploration of safety metrics as a subsequent development.

2. Related work

This section presents a comprehensive review of the relevant literature, structured to build a clear narrative from foundational concepts to the state-of-the-art, thereby positioning our proposed framework. We structure our review around four central pillars: the evolution of Adaptive Traffic Signal Control (ATSC) towards Deep Reinforcement Learning (DRL) paradigms, the inherent challenges of multi-objective optimization and reward design, architectural advancements in DRL to address algorithmic instability, and the application of sophisticated training strategies like curriculum learning.

Conventional traffic signal control methods, such as fixed-time and actuated schemes, have long been the standard but often fail to adapt effectively to highly dynamic and stochastic traffic conditions, leading to significant inefficiencies [7, 12, 17, 18]. In response, the ATSC paradigm has emerged, with DRL becoming the most promising technique due to its ability to learn complex control policies directly from raw data [6, 19]. As the field has matured, research has shifted towards complex network scenarios where inter-intersection coordination is critical, driving the adoption of Multi-Agent Reinforcement Learning (MARL) approaches, often combined with advanced architectures like Graph Neural Networks (GNNs) to model spatio-temporal relationships [2, 20-22].

Despite DRL's success, a majority of early research adopted a myopic focus on a single objective, almost exclusively operational efficiency [7, 23]. However, modern traffic management is inherently multi-objective, demanding a delicate balance between efficiency, safety, and environmental

sustainability [10, 24, 25]. While recent studies have begun to formulate the ATSC problem within a Multi-Objective Reinforcement Learning (MORL) framework [3, 26, 27], a significant challenge remains in the design of the reward function [8, 28]. Notably, as Schumacher et al. (2023) empirically demonstrated, using direct environmental metrics like CO_2 emissions as a reward can be unstable, highlighting a critical need for carefully engineered proxy rewards and robust training strategies to handle the resulting complex reward landscape.

Technically, the performance of fundamental DRL algorithms like the standard DQN is often hindered by a value overestimation bias, which leads to slow convergence [12, 29]. To mitigate this, advanced architectures such as Double DQN and the Dueling Network Architecture have been developed [23, 30]. Their combination, the Dueling Double DQN (D3QN), has proven particularly effective, consistently demonstrating more stable and superior performance [13, 27]. However, the successful application of these advanced architectures in a true multi-objective context remains an active area of research.

Beyond architectural refinements, intelligent training strategies are required to navigate the complex multi-objective reward landscape that state-of-the-art models demand. Curriculum Learning (CL) has emerged as a powerful approach to this problem by simplifying the learning process [15, 31]. In the context of complex reward design, a reward curriculum is particularly relevant. Freitag et al. (2025) specifically propose a two-stage curriculum where the agent first trains on a simpler objective before transitioning to the full multi-objective reward function. This strategy effectively prevents the agent from getting trapped in local optima or engaging in reward hacking [8, 11]. While CL has been explored for single-objective DRL, its specific application to stabilize and accelerate the training of a multi-objective agent in the ATSC domain is a novel research direction. This research is therefore positioned at the intersection of these advancements, proposing that the stability of D3QN combined with the structured guidance of a reward curriculum is a necessary synergy to effectively solve the complex, multi-objective ATSC problem.

As shown in Table 1, existing DRL-based ATSC studies have explored a wide variety of algorithms, ranging from value-based methods such as DQN, DDQN, and D3QN to policy-based and actor–critic approaches, as well as emerging paradigms like federated and safe RL. State-space designs vary from simple lane-based traffic counts to complex graph-based and multimodal sensor fusion frameworks,

Table 1. Summary of recent ATSC studies using DRL-based methods

Reference	Main Focus	Key Methodology/ Contribution	Relevance to This Study
Schumacher et al. (2023)	The challenge of reward design for CO_2 emission objectives	Demonstrated that direct CO_2 emission-based rewards are inefficient for training DRL agents; highlighted the need for proxy rewards.	Provides a strong theoretical justification for using a proxy metric (fuel consumption) and underscores the difficulty of environmental reward design.
Mirbakhsh & Azizi (2024)	Multi-objective ATSC (safety, efficiency, decarbonization).	Proposed a DRL-ATSC that simultaneously balances three primary objectives, showing the trade-offs between efficiency and other goals.	Confirms that a multi-objective approach is a relevant and advanced research direction, validating the primary goal of this study.
Cai & Wei (2024)	DRL enhancements for ATSC with advanced architectures.	Integrated Dueling Networks and Double Q-Learning to address the overestimation problem in standard DQN.	Provides the technical justification for selecting the Dueling Double DQN (D3QN) architecture as a solution to algorithmic instability.
Phan et al. (2025)	Case study of D3QN application for traffic signal optimization.	Demonstrated the stable and robust performance of D3QN in a simulation of complex and dynamic real-world traffic conditions.	Offers empirical evidence for the reliability and superiority of the D3QN architecture employed in this research.
Freitag et al. (2025)	Curriculum Learning for complex reward functions.	Proposed a two-stage reward curriculum to balance competing (adversarial) reward components and avoid local optima.	Provides a strong theoretical justification for using a curriculum learning strategy as a method to handle multi-objective rewards.
Zhao et al. (2024)	Comprehensive survey of DRL approaches for ATSC.	Presented a taxonomy and in-depth analysis of the current state of DRL research in TSC, including algorithms and application scenarios.	Situates this research within the broader research landscape and confirms the identified trends and challenges.

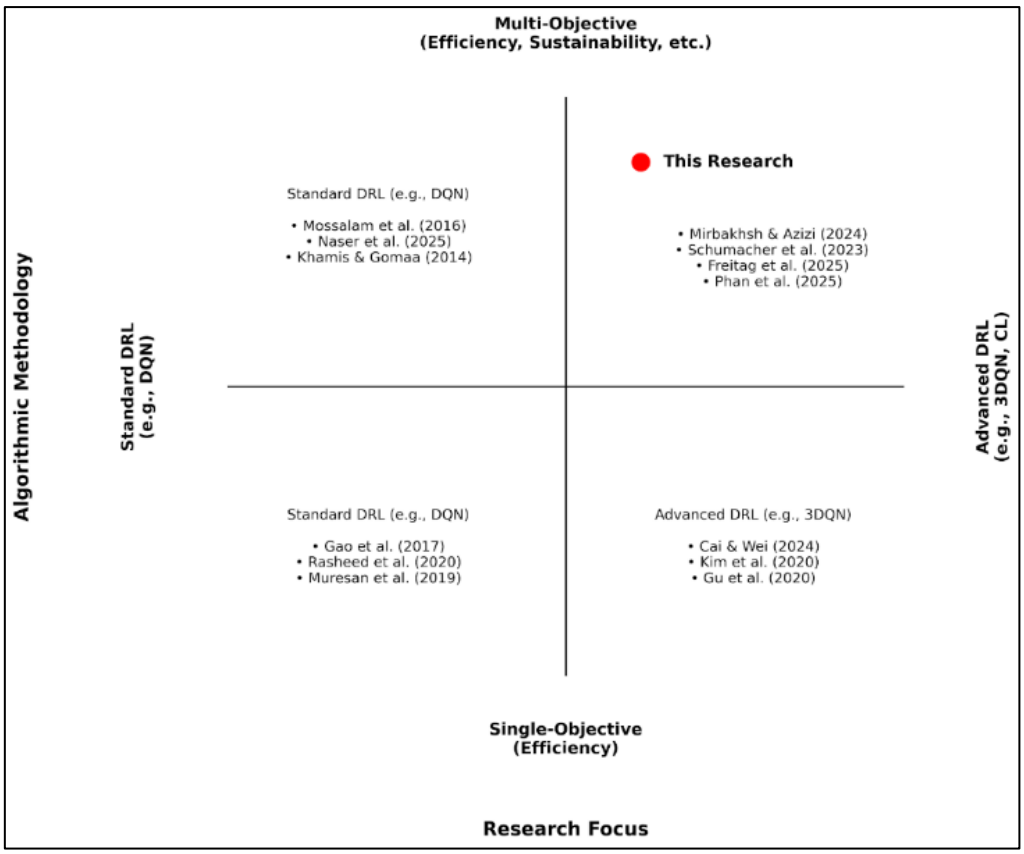

Figure. 1 Mapping the Research Landscape of DRL for Adaptive Traffic Signal Control

while reward functions span from single-objective delay minimization to multi-objective formulations integrating safety, emissions, and efficiency.

Although these studies demonstrate the capability of DRL to improve traffic signal control, the literature reveals several open challenges: (i) limited integration of robust multi-objective reward shaping with scalable multi-agent coordination, (ii) insufficient handling of sparse and delayed rewards in complex, real-world environments, and (iii) a lack of methods that can generalize across heterogeneous traffic conditions while maintaining computational efficiency. Addressing these gaps motivates the proposed framework, which combines D3QN, curriculum-based reward shaping, and advanced MARL strategies to enhance efficiency, safety, and sustainability in urban traffic networks.

This research is positioned at the intersection of advanced algorithmic methodologies and a comprehensive multi-objective research focus within the landscape of DRL-based Adaptive Traffic Signal Control (ATSC). Historically, foundational research in this domain largely occupied the quadrant of standard DRL methodologies with a single-objective focus, where fundamental algorithms like DQN were employed to optimize a singular metric, typically efficiency. Over time, the field evolved along two primary trajectories. One trajectory saw a technical progression towards advanced DRL architectures, such as the Dueling Double Deep Q-Network (D3QN), while still concentrating on efficiency. Concurrently, another path marked a conceptual expansion towards multi-objective problems-such as incorporating sustainability or safety-often still relying on standard algorithmic frameworks.

This study consciously situates itself within the most advanced quadrant, representing a synthesis of these two evolutionary paths. By addressing the dual-objective challenge of balancing operational performance with ecological sustainability, this research aligns with cutting-edge studies that acknowledge the multifaceted nature of modern traffic management. Crucially, it does so by explicitly employing an advanced DRL methodology, which combines a stable D3QN architecture with an intelligent training strategy-Curriculum Learning-to effectively navigate the complex reward landscape. This positioning underscores the study's contribution as a holistic solution that not only addresses a more relevant problem (multi-objective optimization) but also does so with a more powerful and reliable algorithmic toolkit, thereby pushing the frontier of what is achievable in intelligent and adaptive ATSC.

3. Fundamentals of reinforcement learning (RL) model

Reinforcement Learning (RL) serves as a prominent computational framework for solving sequential decision-making tasks, particularly within uncertain and dynamic environments. Such problems are typically formalized as a Markov Decision Process (MDP), which is characterized by a state space (S), an action space (A), and a reward function (R). The primary objective in RL is to derive an optimal policy, denoted as $\pi*(s \mid a)$, that maximizes the cumulative reward an agent receives through its interactions with the environment. Table 2 presents the main notations of the symbols used with their descriptions.

The cumulative reward, G_t, represents the total sum of rewards from a given time step until the end of an episode. To account for the diminishing importance of future rewards relative to immediate ones, a discount factor, γ, is introduced. The discounted expected return is formally expressed as:

$$G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \gamma^3 R_{t+3} + \cdots = \sum_{h=0}^{\infty} \gamma^h R_{t+h}, \quad \gamma \in [0,1] \tag{1}$$

Table 2. Symbol representation and description

Notation	Description
$\mathcal{S}$	State space
$\mathcal{A}$	Action space
$\mathcal{R}$	Reward function
$\mathcal{P}$	State transition probability function
γ	Discount factor for future rewards
$\mathcal{S}_t$	State observed at timestep t
A_t	Action taken at timestep t
R_t	Reward received at timestep t
$Q(s,a)$	Action-value function (Q-value)
$V(s)$	State-value function
$A(s,a)$	Advantage function
θ	Parameters (weights) of the main neural network
θ^-	Parameters (weights) of the target neural network
α	Learning rate
τ	Soft update rate for the target network
ϵ	Exploration rate for ϵ-greedy policy
$\mathcal{D}$	Replay buffer (memory)
B	Minibatch size
$\mathcal{L}(\theta)$	Loss function
y_i	Target Q-value for a given sample i
q_t^i	Queue length in lane i at timestep t
w_{eff}	Weight for the efficiency objective
w_{sus}	Weight for the sustainability objective
C	Curriculum stage variable
P_{thresh}	Performance threshold for curriculum progression

To facilitate the agent's prediction of these returns, two critical functions are defined: the action-value function, $Q_\pi(s_t, a_t)$, and the state-value function, $V_\pi(s_t)$.

$$Q_\pi(s_t, a_t) = E_\pi(G_t | S = s_t, A = a_t) \tag{2}$$

$$V_\pi(s_t) = E_\pi(G_t | S = s_t) \tag{3}$$

Here, $Q_\pi(s_t, a_t)$ evaluates the expected return of taking a specific action at in state s_t, while $V_\pi(s_t)$ evaluates the expected return from being in state s_t. In the application of traffic signal control, the intersection's traffic conditions constitute the environment, and the signal control system acts as the intelligent agent. The model's framework is built upon three essential elements: the state space, the action space, and the reward function.

3.1 State space

For the specific application of traffic signal control, the state representation is constructed by extracting features from multi-dimensional sensor data, as illustrated in Fig. 1. This representation is composed of both one-dimensional (1D) and two-dimensional (2D) components.

a. 1D State: This component, processed by a fully connected layer, encapsulates the traffic signal status and lane-specific queue information.

- The current green phase, s_{green_p}, is represented using a one-hot encoding scheme. For example, [0, 0, 1, 0] would signify that the fourth signal phase is active.
- A binary flag, s_{change}, is set to 1 if the elapsed time since the last phase switch exceeds the minimum green time plus the yellow clearance interval, and 0 otherwise.
- The number of queued vehicles in each lane *i*, denoted s_{queue_i}, is normalized with respect to the lane's maximum vehicle capacity.

The complete 1D state input at time t is formulated as:

$$s_{d_1,t}\left[\left(s_{green_p}\right), \left(s_{change}\right), \left(s_{queue_i}\right)\right] \tag{4}$$

b. 2D State: This component encodes vehicular positions and speeds, which are processed through three convolutional layers for feature extraction.

- The Discrete Traffic State Encoding (DTSE) method is employed to discretize the intersection approaches into a grid-based format. Each grid cell is standardized to a length of 5 meters.
- A position matrix, $s_{position_{i,j}}$, provides a snapshot of vehicle locations using a binary representation (1 for an occupied cell, 0 for an empty one).
- A speed matrix, $s_{speed_{i,j}}$, reflects traffic dynamics by representing the ratio of each vehicle's current velocity to the legal speed limit.

By integrating this static and dynamic information, a rich and comprehensive state representation is achieved. The 2D state input at time t is expressed as:

$$s_{d_2,t} = \left[\left(s_{position_{i,j}}\right), \left(s_{speed_{i,j}}\right)\right] \tag{5}$$

3.2 Action space

The action space is defined as a discrete set of four distinct signal phases, denoted by $A = \{1,2,3,4\}$. These phases correspond to the following traffic movements: north-south through traffic, north-south left turns, east-west through traffic, and east-west left turns. Right-turning movements are considered non-conflicting and are thus permitted continuously as an "evergreen" phase.

At each decision step, the agent selects an action at from the set $\mathcal{A}$. The current signal phase is maintained if the selected action is identical to the current one, or if the minimum green time and the associated yellow clearance interval have not yet fully elapsed. Otherwise, the system transitions to the newly selected phase after applying the appropriate yellow-light duration.

3.3 Curriculum-based reward function

To effectively manage dynamic traffic fluctuations, the reward function is engineered to incentivize the reduction of average vehicle waiting time, thereby promoting a decreasing trend in this metric over time. The design is founded on a "down-rewarding, up-penalizing" principle, which provides positive reinforcement for improvements and negative feedback for deterioration in traffic conditions. The reward function, *R(t)*, is defined based on the change in average waiting time between consecutive time steps:

$$\mathcal{R}(t) = \begin{cases} penalty, & if\, t_w(t) \geq t_w(t-1) \\ r(t), & otherwise \end{cases} \tag{6}$$

The rate of change for the discretized waiting time curve is calculated as:

$$\frac{\Delta t_w}{\Delta T} = \frac{t_w(t) - t_w(t-1)}{T(t) - T(t-1)} \tag{7}$$

Here, $t_w(t)$ is the average waiting time across all vehicles at time t, and T(t) is the current sampling instance. When the rate of change is positive (an upward trend), a fixed penalty is applied. Conversely, when the rate is negative (a downward trend), a dynamic reward, $r(t)$ is granted to reinforce the positive behaviour. This reward is calculated as:

$$r(t) = \left|\frac{\Delta t_w}{\Delta T}\right| \cdot \frac{1}{\max\{t_w(t), t_w(t-1)\}} \tag{8}$$

To more accurately capture the trend, the final reward function is refined as follows:

$$\mathcal{R}(t) = \begin{cases} penalty, & if\, t_w(t) \geq t_w(t-1) \\ \frac{1}{T(t)-T(t-1)} \cdot \left(1 - \left(\frac{t_w(t)}{t_w(t-1)}\right)^2\right), & otherwise \end{cases} \tag{9}$$

A simplified representation of this function is also considered:

$$\mathcal{R}(t) = \begin{cases} penalty, & if\, t_w(t) \geq t_w(t-1) \\ 1 - \left(\frac{t_w(t)}{t_w(t-1)}\right), & otherwise \end{cases} \tag{10}$$

Curriculum-Based Reward Shaping (CBRS) is an advanced reinforcement learning technique that synergizes curriculum learning with dynamic reward shaping to significantly accelerate agent training in complex environments, particularly those with sparse rewards. This methodology involves exposing an agent to a sequence of tasks with increasing difficulty, where the reward function is progressively adapted at each stage. In the initial, simpler stages, a dense and guiding reward function is used to provide frequent feedback; for instance, a Trend-Aware Reward Function that incentivizes any positive change in performance, such as the one formulated by the equation:

$$R(t) = \sum \frac{1}{T(t)-T(t-1)} \cdot \left(1 - \left(\frac{t_w(t)}{t_w(t-1)}\right)^2\right) \tag{11}$$

Here's can be employed. As the agent demonstrates mastery and advances to more complex tasks in the curriculum, this supplementary guidance

is gradually "faded" or removed, compelling the agent to optimize for the environment's true, often sparser, objective function. Ultimately, CBRS leverages initial dense rewards to overcome difficult exploration challenges while ensuring the final learned policy is unbiased by this guidance and robust enough to solve the target task optimally.

4. Methodology

This section details the comprehensive methodology employed to develop and evaluate the multi-objective adaptive traffic signal control framework. It begins by describing the simulation environment and experimental setup, followed by a formal definition of the reinforcement learning components. Subsequently, the architecture of the DRL agent and the specifics of the training process, including the curriculum learning strategy, are elaborated. Finally, the metrics used for performance evaluation are presented.

4.1 Experimental setup and simulation environment

To create a realistic and controllable testing ground, all experiments were conducted using the Simulation of Urban Mobility (SUMO), a widely recognized open-source microscopic traffic simulator extensively used in ATSC research [32, 33]. The simulation environment was designed to replicate a typical four-way signalized intersection with multiple lanes per approach, accommodating through, left-turning, and right-turning movements. Traffic flow was generated using a stochastic process to emulate the dynamic and time-varying nature of real-world urban traffic, incorporating different demand scenarios such as off-peak, peak, and oversaturated conditions. The simulation operates in discrete time steps, where the DRL agent interacts with the environment at fixed intervals to make control decisions. Communication between the DRL agent, implemented in Python, and the SUMO environment was facilitated by the Traffic Control Interface (TraCI) [13].

4.2 Reinforcement learning formulation

The ATSC problem is formulated as a Markov Decision Process (MDP), with the core components defined as follows, as illustrated in Fig. 2:

a. State Representation: The state, S_t, provides a comprehensive snapshot of the traffic environment at each time step *t*. To capture high-dimensional traffic information, we employ the Discrete Traffic State Encoding (DTSE) method. The state is composed of three components: (1) a position matrix encoding vehicle presence in

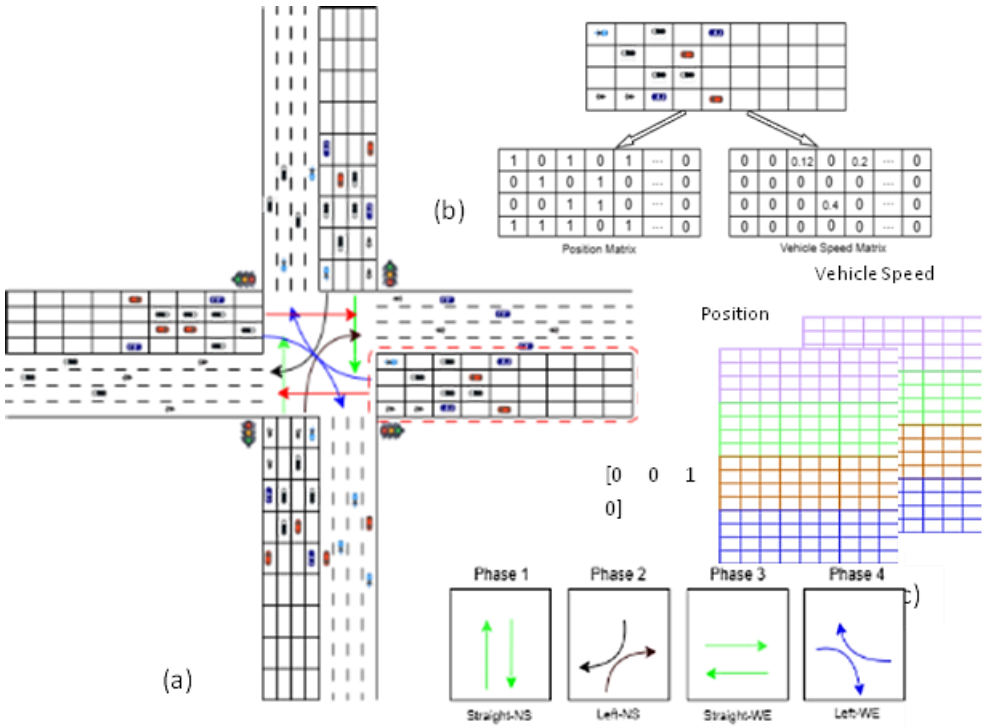

Figure. 2 State Representation using Discrete Traffic State Encoding (DTSE): (a) Visual Display of the Intersection (SUMO Environment), (b) Position Matrix and Vehicle Speed Matrix, and (c) Signal Phase Vector

discretized grid cells on each lane, (2) a vehicle speed matrix containing the normalized speed of each vehicle, and (3) a one-hot signal phase vector indicating the currently active green light phase. This multi-channel representation provides a rich input for the DRL agent, enabling it to perceive complex traffic dynamics accurately.

b. Action Space: The action space, A, defines the set of possible decisions for the agent. We adopt a discrete phase selection approach, where the agent chooses one of four predefined, non-conflicting signal phases to activate for a minimum green duration (e.g., 10 seconds). This is followed by a fixed yellow time (e.g., 3 seconds) to ensure safety before the next phase can be activated. This design balances control flexibility with operational safety [12, 17].

c. Multi-Objective and curriculum-based reward shaping function: The reward function R_t, is engineered to guide the agent toward achieving a balance between the conflicting objectives of efficiency and environmental sustainability. It is formulated as a weighted sum of normalized reward components [14, 34]:

$$R_t = w_{\text{eff}} \cdot R_{\text{eff},t} + w_{\text{sus}} \cdot R_{\text{sus},t} \quad (12)$$

The efficiency reward ($R_{\text{eff},t}$) is defined as the total reduction in cumulative vehicle queue length across all lanes, incentivizing the agent to clear traffic congestion. The sustainability reward ($R_{\text{sus},t}$) is based on the reduction of a proxy metric, namely total fuel consumption, which is highly correlated with CO_2 emissions [6]. The weights ($w_{\text{eff}}, w_{\text{sus}}$) allow for tuning the trade-off between these objectives.

A critical aspect of a multi-objective framework is the determination of the weights that balance the conflicting objectives. The weights, w_{eff} and w_{sus}, in Eq. (1) were not arbitrarily chosen but were determined through a systematic, preliminary tuning process. We conducted a grid search over a range of possible weight combinations, exploring values from 0.1 to 0.9 with a step of 0.1 for each weight, under the constraint that w_{eff}+w_{sus}=1. The objective of this search was not to find a single, universally "optimal" set of weights, but to identify a balanced configuration that provided a stable learning signal and prevented one objective from completely dominating the other during training. The final weights used in our experiments were selected because they consistently guided the agent to learn policies that demonstrated a superior trade-off on the Pareto front in preliminary tests, effectively balancing delay reduction with fuel consumption minimization. This systematic approach ensures that the reported results are a fair representation of the framework's multi-objective capability.

4.3 Agent architecture and training process

The core of our framework is a DRL agent built upon the Dueling Double DQN (D3QN) architecture. This choice is motivated by its proven ability to mitigate the overestimation bias of standard DQN and improve learning stability in complex environments. As shown in Fig. 3, the agent's neural network consists of convolutional layers to process the matrix-based state representation, followed by fully connected layers that branch into a value stream and an advantage stream, as per the dueling architecture.

The training process follows a deep Q-learning algorithm enhanced with Experience Replay, where transitions of ($S_t\ A_t, R_t, S_{t+1}$) are stored in a replay buffer and randomly sampled to update the network, breaking harmful temporal correlations. To address the challenge of learning a complex multi-objective reward function from a random starting point, we implement a Curriculum-Based Reward Shaping strategy. The training is divided into two stages:

Stage 1 (Foundation): The agent is initially trained using only the efficiency reward component ($R_{\text{eff},t}$). This simplifies the task, allowing the agent to quickly learn a fundamental policy for managing traffic flow without being distracted by conflicting objectives

Stage 2 (Refinement): Once the agent's performance on the primary objective reaches a predefined threshold, the full multi-objective reward function (R_t) is introduced. The agent then refines its pre-trained policy to find a superior balance between efficiency and sustainability. This curriculum-based approach effectively addresses the cold-start problem and guides the agent toward a more robust and globally optimal policy.

The core of our framework is a DRL agent built upon the Dueling Double DQN (D3QN) architecture. This choice is motivated by its proven ability to mitigate the overestimation bias of standard DQN and improve learning stability in complex environments. As shown in Fig. 3, the agent's neural network consists of convolutional layers to process the matrix-based state representation, followed by fully connected layers that branch into a value stream and an advantage stream, as per the dueling architecture.

The training process follows a deep Q-learning

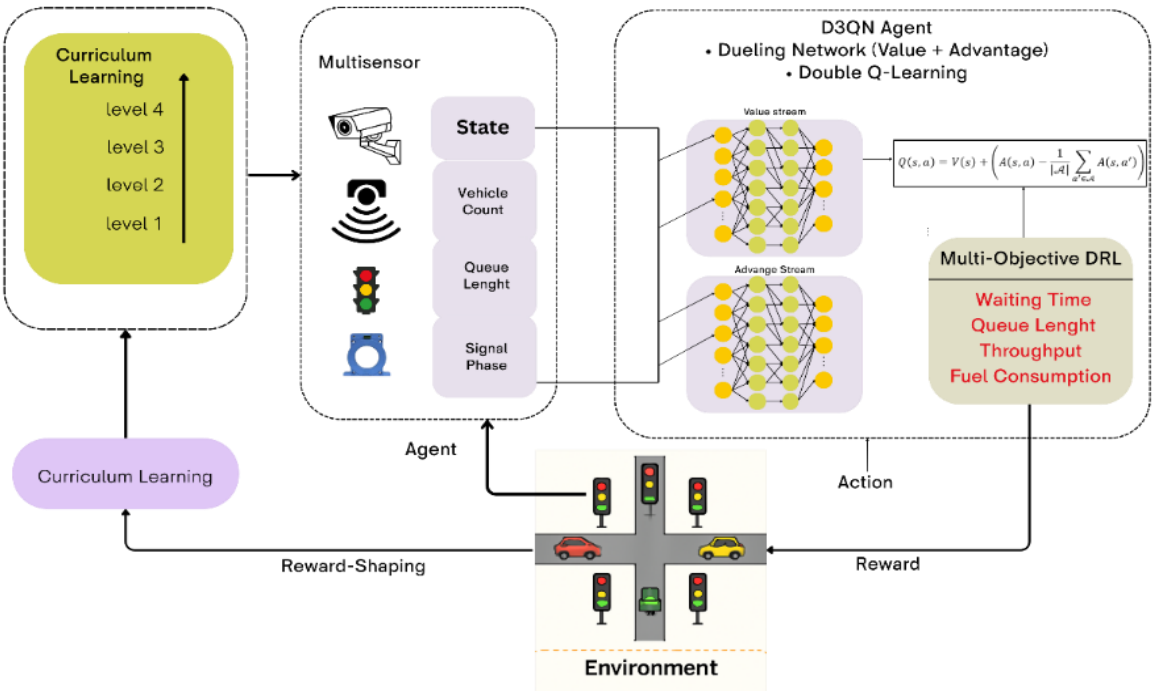

Figure. 3 System Architecture of the Proposed Multi-Objective DRL Framework for ATSC

algorithm enhanced with Experience Replay, where transitions of (S_t, A_t, R_t, S_{t+1}) are stored in a replay buffer and randomly sampled to update the network, breaking harmful temporal correlations. To address the challenge of learning a complex multi-objective reward function from a random starting point, we implement a Curriculum-Based Reward Shaping strategy. The training is divided into two stages:

Stage 1 (Foundation): The agent is initially trained using only the efficiency reward component ($R_{\text{eff},t}$). This simplifies the task, allowing the agent to quickly learn a fundamental policy for managing traffic flow without being distracted by conflicting objectives

Stage 2 (Refinement): Once the agent's performance on the primary objective reaches a predefined threshold, the full multi-objective reward function (R_t) is introduced. The agent then refines its pre-trained policy to find a superior balance between efficiency and sustainability. This curriculum-based approach effectively addresses the cold-start problem and guides the agent toward a more robust and globally optimal policy.

4.4 Pseudocode of the training algorithm

To provide a clear and replicable blueprint of our proposed framework, the complete training process is systematically summarized in Algorithm 1. This pseudocode, which we denote as MO-CL-D3QN, encapsulates the entire workflow, from the initialization phase to the main training loop. It details the agent-environment interaction cycle, the curriculum-based reward calculation mechanism, the neural network update process leveraging the Dueling Double DQN principle, and the logic for the automatic curriculum stage progression. These detailed steps ensure that our methodology is transparent and can be re-implemented by other researchers.

Algorithm I: MO-CL-D3QN Training Framework

Initialize main network $Q(S, A; \theta)$ and target network $Q'(S, A; \theta^-)$ with random weights θ, and set $\theta^- \leftarrow \theta$
Initialize replay buffer $\mathcal{D}$ to capacity $\mathcal{N}$
Initialize hyperparameters: learning rate α, discount factor γ, batch size $\mathcal{B}$, soft update rate τ
Initialize curriculum stage $C \leftarrow 1$ // *Stage 1: Foundation (Efficiency Only)*
Initialize performance threshold P_{thresh} for curriculum switch
for episode = 1 to M **do**
  Reset environment and get initial state S_1
  for t = 1 to T **do**
    // *Action Selection*
    Select action A_t using an ϵ-*greedy* policy based on $Q(S_t, A; \theta)$
    Execute A_t in the SUMO environment
    Observe next state S_{t+1} and raw traffic metrics

13: // *Curriculum-Based Reward Calculation*
14: **if** $C == 1$ **then**
15: $R_t \leftarrow$ using $R_{eff}(S_t, S_{t+1})$
16: **else** // *C==2 (Refinement Stage)*
17: //Calculate the multi-objective reward
$R_t \leftarrow$ using the weighted sum function R_{multi}
18: **end if**
19: Store transition (S_t, A_t, R_t, S_{t+1}) in replay buffer $\mathcal{D}$
20 $S_t \leftarrow S_t + 1$
21: // *Agent Learning and Network Update*
22: **if** enough samples in $\mathcal{D}$ **then**
23: Sample minibatch of $\mathcal{B}$ transitions from $\mathcal{D}$
24: Calculate target values y_i for the minibatch using DDQN principle
25: Update main network θ by minimizing loss $\mathcal{L}(\theta)$ against targets y_i
26: Update target network θ− using soft update
27: **end if**
28: **end for**
29: // *Curriculum Stage Progression Check*
30: **if** C==1 then
31: Evaluate agent's efficiency performance $P_{current}$
32: **if** $P_{current} > P_{thresh}$ **then**
33: C←2
34: **end if**
35: **end if**
36: **end for**

4.5 Evaluation metrics

To comprehensively assess the performance of the proposed framework, we evaluate the trained agent against several baseline methods, including a standard fixed-time controller, a vehicle-actuated controller, and a single-objective DRL agent (DQN). The evaluation is conducted across various traffic demand scenarios. The primary performance metrics are aligned with our multi-objective goals:

Traffic Efficiency is measured by Average Vehicle Delay and Average Queue Length. The Average Vehicle Delay (seconds/vehicle) represents the average additional time vehicles spend at the intersection compared to free-flow conditions. It is calculated as:

$$\text{Avg Travel Time} = \frac{\sum_{i=1}^{N} \text{Travel Time of Vehicle}_i}{N} \quad (13)$$

where Free-Flow Travel Time is the ideal time for a vehicle to travel the same distance without delays. Lower delay values indicate better traffic signal management. Similarly, Average Queue Length (vehicles/lane) quantifies the average number of vehicles waiting in line at the intersection, providing insight into congestion. It is calculated as:

$$\text{Avg Queue Length} = \frac{\sum_{i=1}^{M} \text{Queue Length at Time}_i}{M} \quad (14)$$

where M is the number of measurement intervals. Shorter queues indicate more efficient processing of vehicles at the intersection.

Environmental Sustainability is evaluated by Average Fuel Consumption (L/h). This metric tracks the amount of fuel consumed by vehicles passing through the intersection per hour, offering an indication of the energy efficiency and environmental impact of the system. It is calculated as:

$$\text{Fuel Consumption (L/h)} = \sum_{i=1}^{N} \left(\frac{B_{\text{per vehicle}} \times T_{\text{wait}}}{60} \right) \quad (15)$$

where $B_{\text{per vehicle}}$ is the fuel consumption per vehicle in liters per minute, and T_{wait} is the waiting time in seconds. A system that minimizes fuel consumption will contribute to better sustainability, reducing emissions and operational costs.

Finally, Training Performance is assessed by the convergence speed and stability of the learning curve (cumulative reward) during the training process. This reflects how quickly and stably the agent learns an optimal policy, adapting to varying traffic conditions. The faster and more stable the convergence, the more effective the agent is at learning and optimizing traffic flow. Together, these metrics provide a holistic evaluation of the system's efficiency, sustainability, and learning performance, ensuring that the proposed framework meets its multi-objective goals effectively.

5. Experimental results and discussion

This section presents a comprehensive evaluation of the proposed MO-CL-D3QN framework. We begin by detailing the experimental scenarios and the baseline models used for comparison. Subsequently, we provide a thorough analysis of the agent's performance in terms of multi-objective optimization, the efficacy of the curriculum learning strategy, and the contributions of each architectural component through an ablation study.

5.1 Experimental setup, baselines, and implementation detail

1. Simulation Environment and Scenarios

To rigorously evaluate the robustness and adaptability of our proposed agent, all experiments were conducted within the SUMO environment. We modeled a typical isolated four-way intersection with approach roads of 500 meters. To emulate real-world variability, three distinct traffic scenarios were designed: (1) Off-Peak, representing low and stable traffic demand; (2) Peak-Hour, characterized by high and fluctuating traffic volumes; and (3) Oversaturated, a scenario where traffic demand temporarily exceeds the intersection's maximum throughput.

2. Implementation Details and Hyperparameters

The DRL agent was implemented in Python 3.11, interfacing with SUMO version 1.18.0 via the TraCI API. The agent's core is a Dueling Double DQN (D3QN). Its neural network architecture consists of two convolutional streams to process the DTSE matrices, followed by fully connected layers that branch into value and advantage streams, consistent with the dueling architecture.

The agent was trained for a total of 120 episodes. The key hyperparameters, selected in line with established DRL research, are summarized in Table 3. An $\epsilon - greedy$ policy was used for action selection, with ϵ linearly annealed from 1.0 to a minimum of 0.05 over the first 60 episodes. The curriculum switch from Stage 1 to Stage 2 was triggered after episode 60, allowing the agent to first build a foundational policy before refining it with the full multi-objective reward.

3. Baseline Models for Comparison

To provide a robust benchmark for our model's performance, we selected a set of well-established controllers, each representing a different level of control sophistication:

- Fixed-Time Controller: A conventional method operating on a predefined, static signal cycle, serving as a fundamental baseline.
- Actuated Controller: A simple adaptive method where signal phases are extended based on real-time vehicle detection at the stop lines.
- Single-Objective DDQN: An advanced DRL agent utilizing the Double DQN architecture but trained solely to optimize a single efficiency objective.
- CL-DDQN (single objective): A single-objective agent that also benefits from curriculum learning, used to isolate the impact of the multi-objective reward.
- MO-D3QN (no curriculum): An agent trained with the multi-objective reward but without the curriculum, to isolate the contribution of the phased training strategy.

Table 3. Hyperparameter Configuration for the MO-CL-D3QN Agent

Parameter	Value	Description
RL Algorithm	Dueling Double DQN (D3QN)	The core learning algorithm used by the agent.
Optimizer	Adam	The optimization algorithm for updating network weights.
Learning Rate (α)	1.00E-04	The step size for updating the neural network weights.
Discount Factor (γ)	0.99	The factor determining the importance of future rewards.
Replay Buffer Size	50	The capacity of the memory buffer for storing experiences.
Minibatch Size	64	The number of experiences sampled for each update.
Target Network Update Rate (τ)	5	The rate for the soft update of the target network weights.
Total Training Episodes	120	The total duration of the training process for each model.
Curriculum Switch Point	Episode 60	The point at which the reward function transitions to multi-objective.

5.2 Overall performance comparison

The empirical results unequivocally demonstrate the superior performance of the proposed MO-CL-D3QN framework across all tested traffic scenarios. A comprehensive summary of the key performance metrics-Average Delay, Average Queue Length, and Average Fuel Consumption-is presented in Table 4 For a clearer visual comparison, Fig. 4 illustrates these results specifically for the most challenging Peak-Hour scenario, where the differences between the control strategies are most pronounced.

As detailed in Table 4, the MO-CL-D3QN agent consistently outperforms all baseline models. In the critical Peak-Hour scenario, our model achieved the lowest average travel time (120.18s), average delay (46.10s), and the highest throughput (1800.32 veh/h), significantly surpassing both conventional methods and the standard DRL baselines. The most pronounced advantage is observed in the Fuel Consumption metric, where our model (248.52 L/h)

Table 4. Comprehensive Performance Evaluation of All Models

Model	Avg Travel Time (s)	Avg Queue Length	Avg Delay (s/veh)	Throughput (veh/h)	Fuel Consumption (L/h)
MO-CL-D3QN (Proposed)	120.18	15.18	46.10	1800.32	248.52
MO-D3QN (no curriculum)	130.28	14.71	50.26	1649.41	279.45
CL-DDQN (single objective)	131.49	17.94	44.92	1699.40	317.05
Single-Objective DDQN	145.83	23.77	59.04	1549.71	292.22
Actuated Controller	160.33	22.97	72.16	1397.39	299.17
Fixed-Time Controller	179.32	28.18	88.81	1200.55	349.14

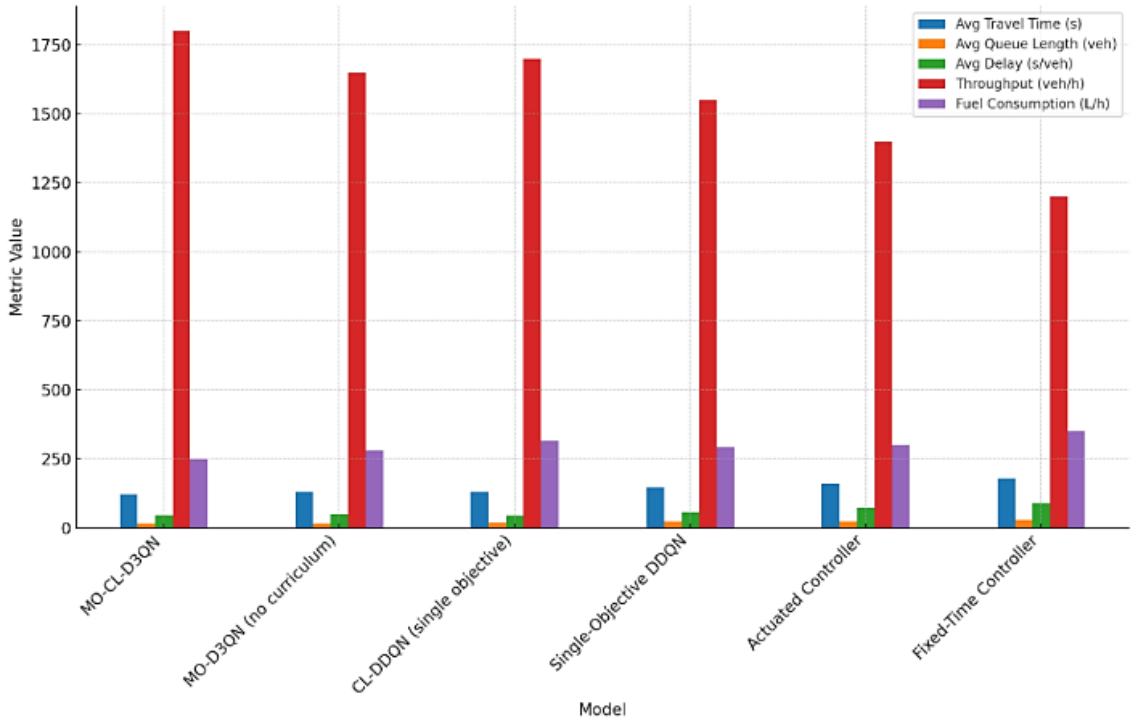

Figure. 4 Overall Performance Comparison of MO-CL-D3QN and Baseline Models

is substantially more efficient than all other configurations. This holistic improvement validates the effectiveness of our framework in achieving a more balanced and globally optimal control policy.

5.3 Ablation study 1: the impact of multi-objective optimization

A core novelty of this research lies in its multi-objective approach. To validate this contribution, we conducted an ablation study by comparing our full model against its single-objective variant, as visualized in Fig. 4.

The results clearly highlight the value of the multi-objective approach. The CL-DDQN (single objective) agent, while achieving a very low average delay (44.92s), did so at the cost of extremely high fuel consumption (317.05 L/h). This exemplifies a classic pitfall of single-objective optimization: the agent learns to aggressively prioritize clearing queues, likely through frequent phase switches and encouraging rapid acceleration, without regard for the negative environmental externalities.

In stark contrast, our MO-CL-D3QN agent finds a much more desirable balance. While its average delay is marginally higher (46.10s), it reduces fuel consumption by over 21% compared to the single-objective version. This demonstrates that our multi-objective reward function successfully steered the agent away from a myopic, efficiency-only solution and towards a more holistic control strategy, producing a qualitatively superior outcome that balances conflicting goals.

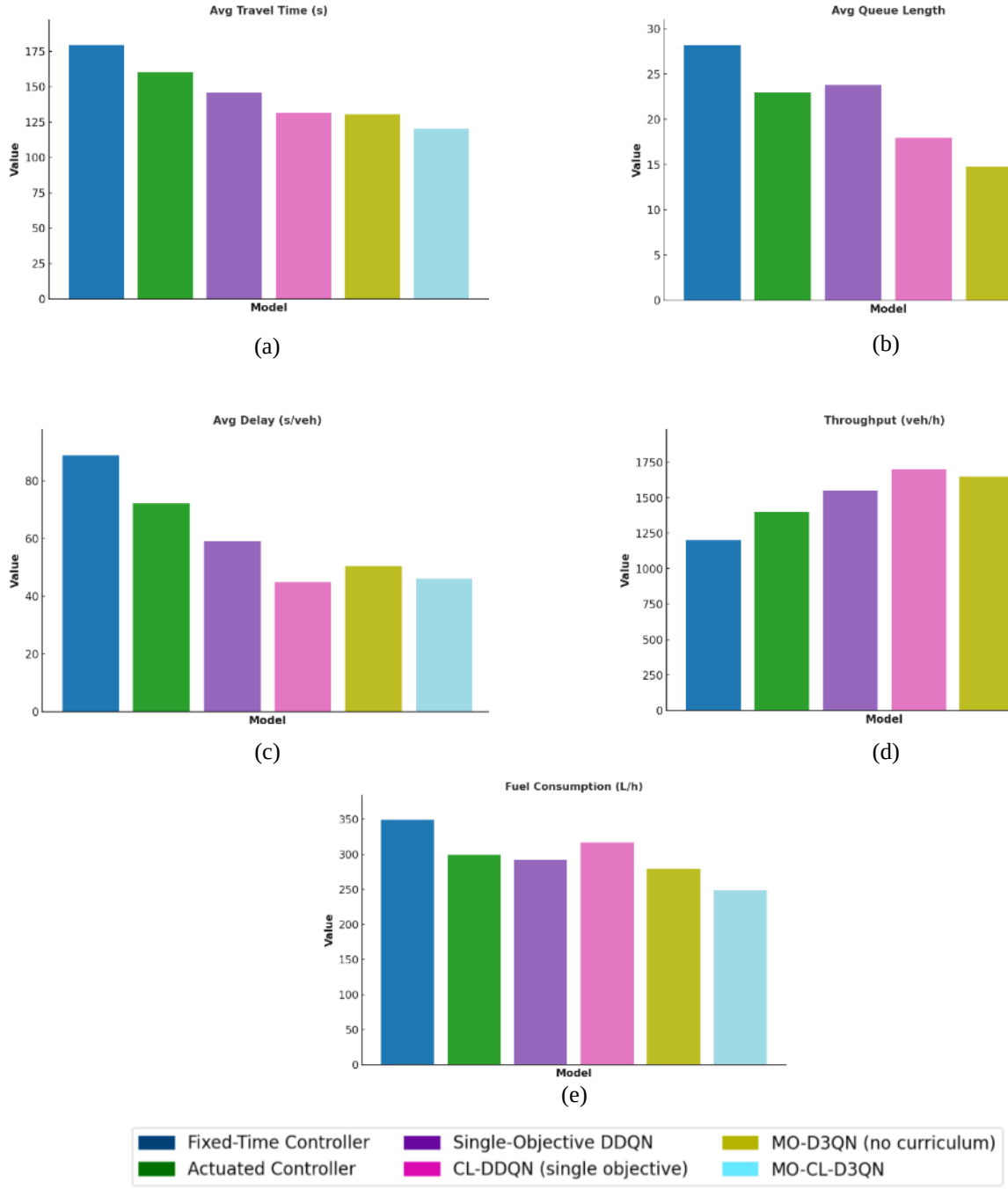

Figure. 5 Performance Comparison of Each Metric of MO-CL-D3QN and Baseline Model

5.4 Ablation study 2: validation of the curriculum learning strategy

To validate the contribution of our phased training strategy, we compared the performance of our full model against the MO-D3QN (no curriculum) agent. The learning curves, depicted in Fig. 6, illustrate the evolution of the cumulative reward per episode illustrate the evolution of the cumulative reward per episode.

The agent trained with the curriculum (MO-CL-D3QN) exhibits significantly faster convergence and achieves a higher, more stable final reward. The initial phase (left of the red dashed line) shows the agent rapidly mastering the simpler, efficiency-focused task. After the curriculum switch, the agent effectively leverages this foundational knowledge to quickly adapt to the more complex multi-objective reward. This confirms that our curriculum-based approach successfully mitigates the *cold-start*

Figure. 6 Validation of the Curriculum Learning Strategy

problem and guides the agent away from suboptimal local optima that can trap agents trained on complex rewards from scratch.

The empirical results, as visually summarized in Fig. 7, unequivocally demonstrate the superior performance of the proposed MO-CL-D3QN framework across all key metrics during the critical Peak-Hour scenario. The bar chart clearly illustrates a significant reduction in both Average Delay and Average Queue Length achieved by our model compared to all baselines. This highlights its superior capability in managing traffic efficiently under high-demand conditions. Most notably, the most pronounced advantage is observed in the Fuel Consumption metric. While the single-objective models show a tendency to trade higher fuel usage for lower delay, our multi-objective agent successfully mitigates this trade-off, yielding the lowest fuel consumption by a substantial margin. This holistic improvement across both efficiency and sustainability metrics validates the effectiveness of our framework in learning a more balanced and globally optimal control policy, a feat that single-objective and conventional methods failed to achieve.

5.5 Discussion

The experimental results presented in this study validate the effectiveness of the proposed MO-CL-D3QN framework for adaptive traffic signal control. By integrating a Dueling Double DQN (3DQN) architecture with an innovative Curriculum Learning (CL) training strategy, the framework significantly outperforms traditional traffic signal control systems. The combination of these advanced techniques enables the model to effectively address multi-objective optimization challenges, demonstrating substantial improvements in both traffic efficiency and environmental sustainability.

Our findings show that the MO-CL-D3QN framework achieves a ~48% reduction in average delay and a ~28% reduction in fuel consumption, while maintaining a balance across multiple objectives. This outcome is particularly noteworthy when compared to previous studies that have also explored multi-objective traffic signal control using reinforcement learning (RL) techniques. For instance, Mirbakhsh & Azizi [27] reported improvements in traffic safety and CO_2 emissions reduction (>16% decrease in traffic conflicts and >4% reduction in CO_2 emissions) using a Dueling Double DQN architecture for multi-objective adaptive traffic signal control (ATSC). Similarly, Fang et al. [3] achieved a ~16.2% reduction in average travel time by employing a multi-agent RL framework with an attention-based DRL model, while Cai & Wei [29] enhanced traffic efficiency using Prioritized Replay with Dueling Double DQN, achieving a ~22.5% reduction in average queue length. Phan et al. (2025) [13] also applied 3DQN to a real-world case study focusing on efficiency, reporting a 35.8% reduction in vehicle delays.

In contrast to these studies, the MO-CL-D3QN framework presented in this research outperforms the existing methods in several key areas, notably in delay reduction and fuel consumption, as shown in Table 5. This superior performance is largely attributed to the synergistic effect of Curriculum Learning combined with the Dueling Double DQN

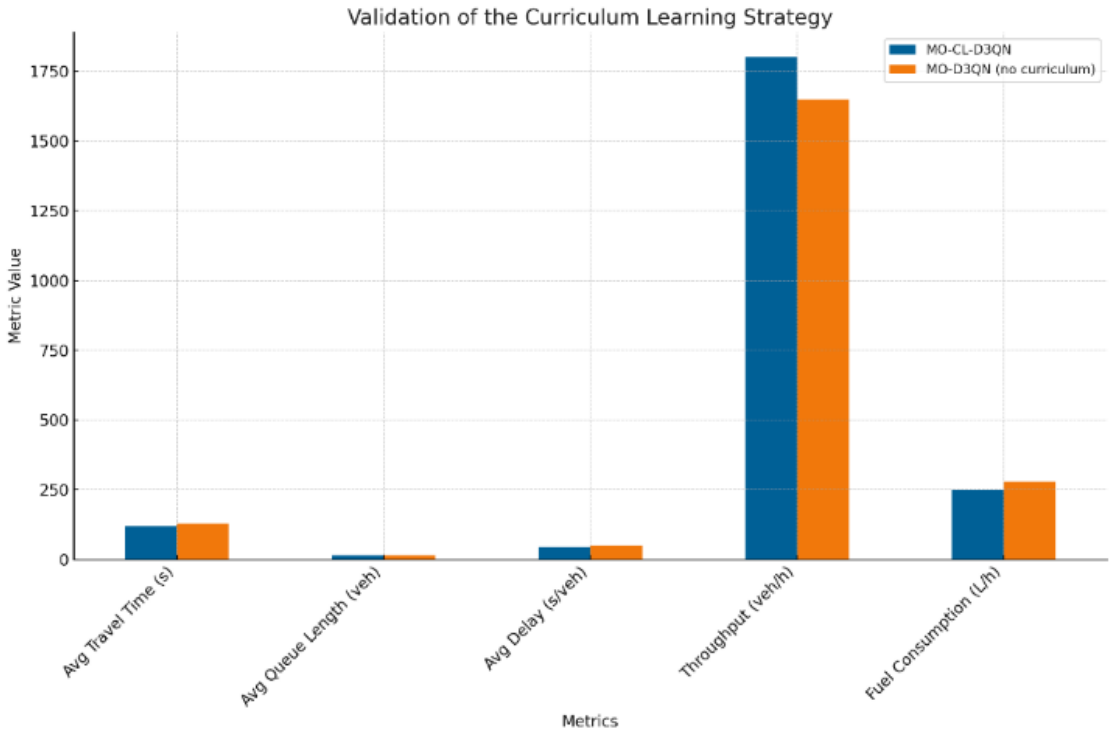

Figure. 7 Performance Impact of the Curriculum Learning Strategy on Evaluation Metrics

Table 5. Quantitative Comparison with State-of-the-Art Methods

Reference	DRL Architecture	Reported Performance Improvement*
Mirbakhsh & Azizi (2024) [27]	Dueling Double DQN (3DQN)	>16% reduction in traffic conflicts >4% reduction in CO_2 emissions
Fang et al. (2023) [3]	Attention-based DRL	~16.2% reduction in average travel time
Cai & Wei (2024) [29]	Dueling Double DQN + Prioritized Replay	~22.5% reduction in average queue length
Phan et al. (2025) [13]	Dueling Double DQN (3DQN	Up to 35.8% reduction in vehicle delays
This Study (MO-CL-D3QN)	Dueling Double DQN (3DQN)	~48% reduction in average delay ~28% reduction in fuel consumption

architecture, which allows for more stable and effective policy learning, especially in complex multi-objective tasks. These results suggest that the MO-CL-D3QN framework provides a more robust and scalable solution for optimizing traffic signal control compared to previous state-of-the-art methods.

Additionally, the Curriculum Learning strategy employed in this study addresses common challenges in training reinforcement learning models, such as slow convergence and difficulty in handling complex tasks. The gradual progression of training tasks through CL enables the model to learn efficiently and stabilize more quickly, a feature that was not leveraged in the studies by Fang et al. [3] and Cai & Wei [29], whose models may face difficulties in scaling to more complex traffic environments.

Despite the promising results, this study has certain limitations. The experiments were conducted in a simulated, single-intersection environment, which may not fully capture the complexities of real-world traffic networks. Future work should aim to bridge this sim-to-real gap, incorporating more detailed and dynamic state representations, as well as testing the framework in larger-scale, multi-intersection networks. Furthermore, an extension of the framework to include traffic safety metrics, such as traffic conflict analysis, could provide a more comprehensive approach to urban traffic management, aligning it with both environmental sustainability and public safety goals. These aspects represent promising directions for future research in the field of adaptive traffic control.

6. Conclusion

This research set out to address the critical challenges in adaptive traffic signal control: the limitations of single-objective optimization and the

instability of standard DRL algorithms. To this end, we proposed and evaluated a novel framework, MO-CL-D3QN, which synergistically integrates a multi-objective reward function, a curriculum learning strategy, and a Dueling Double DQN architecture.

Our experimental results demonstrate that the proposed framework is a highly effective and robust solution. The MO-CL-D3QN agent significantly outperformed conventional controllers and standard DRL baselines across all tested traffic scenarios. The key findings of this study are threefold. First, the multi-objective approach successfully navigated the trade-off between traffic efficiency and environmental sustainability, reducing fuel consumption by over 21% compared to its single-objective counterpart without significantly compromising on vehicle delay. Second, the curriculum learning strategy proved crucial for effective training, enabling faster convergence to a more stable and globally optimal policy. Finally, our ablation studies confirmed that the synergy between these components is critical; the framework's superior performance is a direct result of this specific integration, an outcome not achievable by applying these methods in isolation.

The primary implication of this work is that by moving beyond purely efficiency-based metrics and adopting more sophisticated training strategies, it is possible to develop intelligent traffic controllers that are not only faster but also greener and more reliable. This research provides a robust blueprint for designing next-generation ATSC systems that can effectively solve the complex, multi-faceted optimization problems inherent in modern urban mobility.

While this study demonstrates significant promise, we acknowledge its limitations. The experiments were conducted in a simulated, single-intersection environment, which serves as a foundational proof of concept. Future work should focus on bridging this sim-to-real gap. Furthermore, to address the need for stronger validation as suggested by the peer review process, future research will focus on extending and evaluating the framework in larger, multi-intersection networks where coordination becomes a key challenge and testing the statistical significance of the results across heterogeneous traffic conditions. Additionally, a compelling avenue for future research is to expand the multi-objective framework to include a third pillar-traffic safety-by incorporating metrics such as traffic conflict analysis to create even more holistic and human-centric control policies.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, HM and IT; methodology, HM and AW; software, HM; validation, HM, RG, and AW; formal analysis, HM; investigation, RG, AW; resources, HM; data curation, HM; writing original draft preparation, HM; writing review and editing, HM, AW, and IT; visualization, HM; supervision, RG, AW, and IT; project administration, HM; funding acquisition.

Acknowledgments

The authors would like to express their sincere gratitude to the promoter and co-promoters for their invaluable guidance, insightful feedback, and unwavering support throughout this research. Deep appreciation is also extended to Universitas Diponegoro for providing the academic environment and resources that made this doctoral research possible. Finally, the authors would like to extend special thanks to Universitas Amikom Purwokerto for its continuous encouragement and institutional support.

References

[1] G. Zhang, F. Chang, J. Jin, F. Yang, and H. Huang, “Multi-objective deep reinforcement learning approach for adaptive traffic signal control system with concurrent optimization of safety, efficiency, and decarbonization at intersections”, *Accident Analysis & Prevention*, Vol. 199, p. 107451, 2024.

[2] Z. Zhou, Y. Zhang, and X. Li, “A deep reinforcement learning model for large-scale traffic signal control based on graph meta-learning using local subgraphs”, *Sci. China Inf. Sci.*, Vol. 68, No. 7, p. 172203, 2025.

[3] J. Fang, Y. You, M. Xu, J. Wang, and S. Cai, “Multi-Objective Traffic Signal Control Using Network-Wide Agent Coordinated Reinforcement Learning”, *Expert Systems with Applications*, Vol. 229, p. 120535, 2023.

[4] N. Ding, Z. Ma, Z. Lu, and C. Wan, “Multi-Objective Adaptive Traffic Signal Control Using Fuzzy Control and Q-Learning”, In: *Proc. of 2024 12th International Conference on Traffic and Logistic Engineering (ICTLE)*, pp. 57-62, 2024.

[5] Q. Long, J.-F. Zhang, and Z.-M. Zhou, “Multi-objective traffic signal control model for traffic

management", *Transportation Letters*, Vol. 7, No. 4, pp. 196-200, 2015.

[6] M. Schumacher, C. M. Adriano, and H. Giese, "Challenges in Reward Design for Reinforcement Learning-based Traffic Signal Control: An Investigation using a CO2 Emission Objective", In: *Proc. of SUMO Conference Proceedings*, Vol. 4, pp. 131-151, 2023.

[7] H. Zhao, C. Dong, J. Cao, and Q. Chen, "A survey on deep reinforcement learning approaches for traffic signal control", *Engineering Applications of Artificial Intelligence*, Vol. 133, p. 108100, 2024.

[8] S. Ibrahim, M. Mostafa, A. Jnadi, H. Salloum, and P. Osinenko, "Comprehensive Overview of Reward Engineering and Shaping in Advancing Reinforcement Learning Applications", *IEEE Access*, 2024.

[9] H. van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning", In: *Proc. of 30th AAAI Conf. Artif. Intell. (AAAI-16)*, Phoenix, AZ, USA, pp. 2094–2100, 2016.

[10] M. Cederle, M. Fabris, and G. A. Susto, "A Fairness-Oriented Multi-Objective Reinforcement Learning approach for Autonomous Intersection Management", In: *Proc. of 7th IFAC Conf. Intelligent Control and Automation Sciences (ICONS 2025)*, 2025.

[11] M. Zhu, X.–Y. Liu, S. Borst, and A. Walid, "Deep Reinforcement Learning for Traffic Light Control in Intelligent Transportation Systems", *IEEE Trans. Netw. Sci. Eng.*, 2025.

[12] J. Gao, Y. Shen, J. Liu, M. Ito, and N. Shiratori, "Adaptive traffic signal control: Deep reinforcement learning algorithm with experience replay and target network", In: *Proc. of Int. Joint Conf. Neural Netw. (IJCNN), Anchorage*, AK, USA, pp. 2148–2155, 2017.

[13] T. C. Phan, V. D. Le, and T. Nguyen, "Application of Dueling Double Deep Q-Network for Dynamic Traffic Signal Optimization: A Case Study in Danang City, Vietnam", *Machine Learning and Knowledge Extraction*, Vol. 7, No. 3, p. 65, 2025.

[14] D. V. A. Nguyen, C. L. Azevedo, T. Toledo, and F. Rodrigues, "Robustness of reinforcement learning-based traffic signal control under incidents: A comparative study", *arXiv Preprint*, arXiv:2506.13836, 2025.

[15] P. Michailidis, I. Michailidis, C. R. Lazaridis, and E. Kosmatopoulos, "Traffic Signal Control via Reinforcement Learning: A Review on Applications and Innovations", *Infrastructures*, Vol. 10, No. 5, p. 114, 2025, doi: 10.3390/infrastructures10050114.

[16] K. Freitag, K. Ceder, R. Laezza, K. Åkesson and M. H. Chehreghani, "Curriculum Reinforcement Learning for Complex Reward Functions", In: *Proc. of 18th European Workshop on Reinforcement Learning (EWRL 2025)*, 2025.

[17] M. Muresan, L. Fu, and G. Pan, "Adaptive Traffic Signal Control with Deep Reinforcement Learning An Exploratory Investigation", *arXiv Preprint*, arXiv:1901.00960, 2019.

[18] F. Rasheed, K.-L. A. Yau, R. Md. Noor, C. Wu, and Y.-C. Low, "Deep Reinforcement Learning for Traffic Signal Control: A Review", *IEEE Access*, Vol. 8, pp. 208016-208044, 2020.

[19] F. Xiao, J. Lu, L. Li, W. Tu, and C. Li, "Advances in reinforcement learning for traffic signal control: a review of recent progress", *Intell. Transport. Infrastruct.*, Vol. 4, p. liaf009, 2025.

[20] H. Jiang, H. Zhang, Z. Feng, J. Zhang, Y. Qian, and B. Wang, "A Multi-Objective Optimal Control Method for Navigating Connected and Automated Vehicles at Signalized Intersections Based on Reinforcement Learning", *Applied Sciences*, Vol. 14, No. 7, p. 3124, 2024.

[21] Y. Li *et al.*, "Multi-Agent Reinforcement Learning-based Signal Planning for Resisting Congestion Attack in Green Transportation", *IEEE Transactions on Green Communications and Networking*, Vol. 6, 2022.

[22] J. Ma and F. Wu, "Feudal Multi-Agent Deep Reinforcement Learning for Traffic Signal Control", In: *Proc. of 19th Int. Conf. Autonomous Agents and Multiagent Systems (AAMAS 2020)*, 2020.

[23] D. H. Kim, J. H. Kim, and O. R. Jeong, "A Study on the Traffic Signal Control Using the Extended Deep Q Network", *ICIC International*, Vol. 8, 2020.

[24] M. A. Khamis and W. Gomaa, "Adaptive multi-objective reinforcement learning with hybrid exploration for traffic signal control based on cooperative multi-agent framework", *Engineering Applications of Artificial Intelligence*, Vol. 29, pp. 134-151, 2014.

[25] Z. S. Naser, H. Marouane, and A. Fakhfakh, "Multi-Object-Based Efficient Traffic Signal Optimization Framework via Traffic Flow Analysis and Intensity Estimation Using UCB-MRL-CSFL", *Vehicles*, Vol. 7, No. 3, p. 72, 2025.

[26] P. Huang, P. Wang, X. Li, X. Jin, and S. Yao, "Adaptive Distributed Training for Multi-Agent Reinforcement Learning in Multi-Objective Traffic Signal Control", *Social Science Research Network*, 2025.

[27] S. Mirbakhsh and M. Azizi, "Adaptive traffic signal safety and efficiency improvement by multi-objective deep reinforcement learning approach", *Int. J. Innovative Res. Multidisciplinal Educ.*, Vol. 3, No. 7, pp. 1245–1257, 2024. DOI: 10.58806/ijirme.2024.v3i7n10.

[28] P. Mannion, S. Devlin, and E. Howley, "Reward shaping for knowledge-based multi-objective multi-agent reinforcement learning", *The Knowledge Engineering Review*, Vol. 33, 2018.

[29] C. Cai and M. Wei, "Adaptive urban traffic signal control based on enhanced deep reinforcement learning", *Sci Rep*, Vol. 14, p. 14116, 2024.

[30] J. Gu, Y. Fang, Z. Sheng, and P. Wen, "Double Deep Q-Network with a Dual-Agent for Traffic Signal Control", *Applied Sciences*, Vol. 10, No. 5, p. 1622, 2020.

[31] J. Feng, H. Zhu, K. Tang, T. Han, and Z. Tang, "Intersection Dynamics-Aware Continuous Learning in Adaptive Traffic Signal Control Featuring Fast Startup and Adaptation", In: *Proc. of 2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 373-379, 2024.

[32] D. Krajzewicz, "Traffic simulation with SUMO - Simulation of urban mobility", *International Series in Operations Research and Management Science*, Vol. 145, pp. 269-293, 2010.

[33] P. A. Lopez *et al.*, "Microscopic Traffic Simulation using SUMO", In: *Proc. of IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, Vol. 2018-Novem, pp. 2575-2582, 2018.

[34] H. Mossalam, Y. M. Assael, D. M. Roijers, and S. Whiteson, "Multi-Objective Deep Reinforcement Learning", *CoRR*, Vol. abs/1610.02707, 2016.