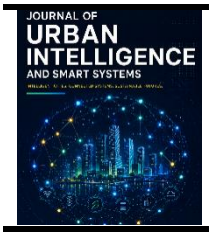

Journal of Urban Intelligence and Smart Systems

www.juiss.org
eISSN: 3153 – 4023

Event-Triggered Preference Adaptation with Large Language Model-Guided Conditional Reinforcement Learning for Pedestrian-Aware Traffic Signal Control

Yunpeng Ma[1,*], Xinwei Zhang[2], Ferenc Mészáros[1]

[1] Department of Transport Technology and Economics, Faculty of Transportation Engineering and Vehicle Engineering, Budapest University of Technology and Economics, Budapest, Hungary
[2] Department of Control for Transportation and Vehicle Systems, Faculty of Transportation Engineering and Vehicle Engineering, Budapest University of Technology and Economics, Budapest, Hungary

ARTICLE INFO

Article history:
Received 15 July 2026
Received in revised form 25 August 2026
Accepted 2 September 2026
Available online 6 September 2026

Keywords:
Urban access control; Traffic signal control; Reinforcement learning; Large language model; Pedestrian fairness; Event-triggered control

ABSTRACT

Traffic signal control at mixed vehicle-pedestrian intersections requires balancing vehicle efficiency, pedestrian service, and operational stability. Existing reinforcement learning methods often rely on fixed reward preferences or unconstrained signal actions, which may limit their adaptability and interpretability under dynamic conflict conditions. To address this issue, the paper proposes an event-triggered large language model-guided conditional reinforcement learning framework. In this framework, the reinforcement learning agent optimizes green duration under a fixed cyclic phase sequence, while the large language model acts as a high-level preference adapter rather than a direct signal controller. A conflict-aware event trigger is designed using pedestrian queue, pedestrian waiting time, and vehicle queue to activate the large language model only under critical states. The generated preference vector guides a preference-conditioned dueling double deep Q-network to balance vehicle efficiency, pedestrian-oriented fairness, and stop-and-go stability. Microscopic simulation experiments in Vissim show that the proposed framework achieves more balanced performance across reward, average delay, stop frequency, pedestrian waiting, and preference-adaptation behavior. Compared with fixed-time control, standard reinforcement learning, and pressure-based control, the proposed method provides stronger adaptability and more interpretable multi-objective decision-making. The results indicate that event-triggered LLM-guided preference adaptation is an effective strategy for pedestrian-aware traffic signal control.

1. Introduction

Signalized intersections are critical nodes in urban transport systems, where vehicle movements and pedestrian crossings must be coordinated under limited spatial and temporal resources. Pedestrians are particularly important in this context because they are vulnerable road users, and their crossing behavior is highly sensitive to signal delay, perceived risk, and interactions with

* *Corresponding author.*
E-mail address: mayunpeng@edu.bme.hu
https://doi.org/10.59543/110jfp29

surrounding vehicles [1, 2]. Excessive pedestrian waiting time may reduce compliance with traffic signals and increase the probability of risky crossing behavior, while vehicle-oriented signal timing may cause long queues, frequent stop-and-go movements, and inefficient intersection operation [1, 2]. Therefore, modern traffic signal control should not only optimize vehicle throughput and average delay, but also explicitly consider pedestrian service fairness and operational stability.

Deep reinforcement learning (DRL) has been widely investigated for adaptive traffic signal control because it can learn signal timing policies through repeated interaction with dynamic traffic environments [3-8]. Recent reviews further confirm that RL has become one of the main methodological directions for intelligent traffic signal control, but also point out challenges related to computational cost, data requirements, and generalization across diverse traffic scenarios [9]. Compared with fixed-time or rule-based control strategies, DRL-based controllers can adapt their decisions according to real-time traffic states. However, existing DRL-based signal control methods still face several limitations when applied to pedestrian-aware intersection control. First, many methods mainly optimize vehicle-related indicators, such as queue length, travel time, or throughput, while pedestrian fairness and extreme waiting time are often treated as secondary indicators rather than core optimization objectives [3]. Second, unconstrained action spaces may allow agents to freely select signal phases, which can lead to unrealistic behaviors such as repeated service of dominant approaches, phase starvation, or unsafe exploration. Such behaviors are inconsistent with practical traffic engineering requirements, in which phase sequences and clearance intervals must meet operational constraints. Third, conventional scalar reward functions usually rely on fixed objective weights. This makes it difficult for the controller to dynamically adjust its optimization priority as traffic conditions change, for example, when pedestrian waiting becomes critical or when the risk of vehicle queue spillback increases.

Recent studies have started to explore multi-objective reinforcement learning, fairness-aware signal control, and Large Language Model (LLM)-assisted decision-making for traffic signal control [10-16]. Existing methods can generally be grouped into LLM-augmented reinforcement learning, LLM-centric reasoning agents, and knowledge-enhanced hybrid systems, indicating that LLMs are increasingly being used to support high-level reasoning and decision assistance in traffic signal control [17]. For example, recent studies have investigated LLM-enabled universal traffic signal control across different intersection layouts and traffic-flow conditions, as well as LLM-based actor–critic signal optimization for complex intersection forms such as displaced left-turn intersections [18,19]. LLMs have strong contextual reasoning capabilities and can provide interpretable, high-level guidance in complex traffic situations. Nevertheless, directly using an LLM as a low-level signal controller is unsuitable for real-time intersection operations. Frequent LLM invocations may incur high computational costs, introduce communication latency, and lead to unstable changes in preferences. Moreover, if the reward preference is updated too frequently during reinforcement learning, the learning target may become non-stationary, slowing convergence and reducing policy stability. These issues indicate the need for a framework that combines the contextual reasoning capabilities of LLMs with the fast execution and learning capabilities of reinforcement learning, while avoiding excessive LLM calls and maintaining realistic signal-control constraints.

However, a research gap still exists. Existing pedestrian-aware reinforcement learning methods often consider pedestrian delay as an additional performance indicator rather than as an explicit fairness-related control objective. Many reinforcement-learning-based signal controllers use fixed reward preferences or unconstrained phase-selection actions, which limit their ability to adapt to dynamic pedestrian–vehicle conflicts while satisfying practical constraints on signal operation. Although recent LLM-assisted methods offer new opportunities for contextual reasoning and cross-

scenario signal control, many still use LLMs as direct or frequent decision-making components in the signal control process [17-19]. Directly using an LLM as a real-time low-level signal controller may introduce latency, computational cost, and unstable decision outputs.

The significance of this research lies in its effort to bridge the gap between adaptive artificial intelligence-based decision-making and the practical requirements of traffic engineering. For mixed vehicle–pedestrian intersections, signal control should not only improve vehicle efficiency, but also prevent excessive pedestrian waiting and reduce stop-and-go-related operational instability. By using the LLM as a high-level preference adapter rather than as a direct signal controller, the proposed approach can exploit the contextual reasoning capability of LLMs while preserving the fast execution, safety constraints, and learning stability of reinforcement learning-based signal control.

The objective of this study is to develop an Event-Triggered Large Language Model-Guided Conditional Reinforcement Learning (ET-LCRL) framework for pedestrian-aware, multi-objective traffic signal control. The key idea is to decouple high-level preference adaptation from low-level signal timing control. The LLM is not used to directly select signal phases or green times. It acts as a low-frequency meta-controller, generating a multi-objective preference vector only when the traffic state indicates a critical pedestrian–vehicle service conflict. The low-level controller is implemented as a preference-conditioned reinforcement learning agent that selects only the green duration for the currently served phase based on the physical traffic state, the current signal phase, and the preference vector.

In the low-level controller, the agent follows a fixed cyclic phase sequence and is allowed to optimize only the green duration of the current phase. A fixed all-red clearance interval is inserted before each green phase. This constrained action-space design reduces unsafe exploration, prevents phase starvation, and improves the engineering feasibility of the learned policy. The main contributions of this study are summarized as follows:

i. A pedestrian-aware event-triggered LLM-guided conditional reinforcement learning framework is proposed for multi-objective traffic signal control. The framework uses the LLM as a high-level preference generator rather than a direct low-level signal controller.
ii. A conflict-aware preference adaptation mechanism with hysteresis thresholds is developed. The mechanism invokes the LLM only under critical traffic states, thereby reducing computational overhead and mitigating reward non-stationarity.
iii. A deployable constrained action-space design is introduced. The agent selects only green duration, while the environment enforces a fixed cyclic phase sequence and all-red clearance, which improves learning stability and prevents phase starvation.
iv. A fairness-aware multi-objective reward structure is designed to jointly consider vehicle efficiency, pedestrian service fairness, and stop-and-go-related operational stability.

The remainder of this paper is organized as follows. Section 2 reviews related studies on pedestrian-vehicle interaction, reinforcement learning-based traffic signal control, multi-objective optimization, and LLM-assisted traffic decision-making. Section 3 presents the proposed ET-LCRL methodology. Section 4 describes the simulation environment, experimental settings, and comparison controllers. Section 5 reports and discusses the experimental results. Section 6 concludes the paper and outlines future research directions.

2. Related Work

2.1 Pedestrian–Vehicle Interaction and Pedestrian-Aware Signal Control

Pedestrian-vehicle interaction is a fundamental issue at urban intersections because pedestrians and vehicles compete for limited right-of-way within the same signal cycle. Pedestrians are vulnerable road users, and their crossing decisions are affected by signal timing, waiting tolerance, vehicle movement, perceived risk, and surrounding social information [1, 2]. Previous studies have shown that long pedestrian waiting times and long signal cycles may increase the likelihood of red-light violations or non-compliant crossing behavior [1]. Therefore, pedestrian delay is not only a service-level indicator but also an important factor related to traffic safety and signal compliance.

Existing studies on pedestrian-aware traffic control have mainly focused on pedestrian delay estimation, crossing behavior analysis, and safety-oriented signal timing. For example, pedestrian delay models have been developed for signalized crosswalks under mixed-traffic conditions, indicating that pedestrian waiting time should be explicitly considered in the evaluation of signal performance [1]. Recent reinforcement learning studies have also attempted to incorporate pedestrian states into traffic signal control, aiming to reduce both vehicle and pedestrian waiting times [3]. These studies demonstrate the importance of considering pedestrians in adaptive signal control.

However, most existing pedestrian-aware signal control studies still treat pedestrians as an additional constraint or an auxiliary performance indicator. Vehicle-oriented metrics, such as throughput, average delay, and queue length, remain the dominant optimization targets. As a result, extreme pedestrian waiting time and fairness among different road users may be overlooked. In particular, a controller that performs well in terms of average vehicle delay may still produce unacceptable pedestrian service outcomes under unbalanced demand. This motivates the need for a signal control framework that explicitly accounts for pedestrian fairness alongside vehicle efficiency and operational stability.

2.2. Reinforcement Learning-Based Traffic Signal Control and Practical Action-Space Constraints

Reinforcement learning has been widely used in adaptive traffic signal control because it can learn control policies through interaction with dynamic traffic environments. Representative studies such as IntelliLight, PressLight, and CoLight have demonstrated the potential of deep reinforcement learning in reducing delay, queue length, and network-level congestion [4-6]. Compared with fixed-time or manually designed actuated controllers, reinforcement learning-based methods can adapt signal decisions according to observed traffic states. Despite these advantages, the practical deployment of reinforcement learning-based traffic signal control remains challenging. Many existing methods allow the agent to select signal phases directly. Although this design increases decision flexibility, it may also lead to unrealistic or unsafe behaviors, such as repeatedly selecting a dominant phase, skipping minor phases, or causing phase starvation. In real-world signal systems, the signal phase sequence, intergreen time, yellow time, and all-red clearance are subject to strict engineering constraints. Therefore, a reinforcement learning controller that ignores these constraints may obtain good simulation performance but have limited deployability.

Recent studies have increasingly emphasized the importance of realistic action-space design. Some reinforcement-learning-based controllers formulate traffic signal control through duration or cycle adjustment, while others define action spaces around realistic phase-based signal decisions [7, 8]. Compared with fully unconstrained signal selection, a more practical strategy is to let the agent optimize parameters of an existing signal program, such as phase duration or green duration, while

keeping required safety constraints unchanged. This design reduces unsafe exploration and makes the learned policy more compatible with real-world traffic control logic. In this study, a preference-conditioned reinforcement learning structure is adopted.

2.3. Multi-Objective and Fairness-Aware Reinforcement Learning for Traffic Signal Control

Traffic signal control is inherently a multi-objective problem. Vehicle throughput, vehicle delay, pedestrian waiting time, stop-and-go frequency, safety, and environmental impact may conflict with one another. Traditional reinforcement learning methods typically combine multiple objectives into a single scalar reward function using fixed weights. Although this approach is simple and computationally efficient, it has two important limitations. First, fixed reward weights cannot reflect changing traffic priorities under different operating conditions. Second, a scalar reward may mask unfair outcomes, especially when average performance improves at the cost of extreme delays for a specific group of road users.

Multi-objective reinforcement learning has received increasing attention in traffic signal control. Recent studies have explored multi-objective reinforcement learning frameworks that jointly optimize traffic efficiency, safety, and environmental indicators such as carbon emissions [10, 11]. Fairness-aware traffic signal control has also been studied to reduce extreme waiting time and to balance service among competing flows [12-14]. These works show that signal control should not be evaluated only by average delay or throughput but should also consider the distribution of service quality across users and objectives.

Nevertheless, many multi-objective traffic signal control methods still rely on predefined or slowly changing preference structures. Once the reward weights are fixed before training, the learned policy may not respond properly to sudden changes in pedestrian demand or vehicle queue pressure. Training separate policies for different objective preferences is also inefficient and difficult to generalize. Therefore, a more flexible approach is to condition the value function or policy on a preference vector, allowing a single controller to adapt its behavior to different optimization priorities.

2.4. LLM-Assisted Traffic Signal Control and Event-Triggered Preference Guidance

Large language models have recently attracted attention in intelligent transportation systems due to their reasoning capabilities and potential for generalization. For example, LLMLight uses large language models as traffic signal control agents, while LLM-Assisted Light integrates LLM reasoning with traffic signal control to improve human-mimetic decision-making in complex urban environments [15, 16]. These studies indicate that LLMs can provide high-level reasoning beyond conventional numerical reward functions. However, directly using an LLM as a low-level signal controller also introduces important challenges. First, real-time signal control requires fast, stable decisions, whereas online LLM inference may introduce latency and computational overhead. Second, LLM-generated decisions may be sensitive to prompt design and may contain hallucinated or inconsistent reasoning. Third, if LLM outputs are used to frequently adjust reward preferences during reinforcement learning, the learning objective becomes non-stationary, which may reduce training stability and policy convergence.

To address these issues, the present study does not use the LLM to directly select signal phases or green durations. Instead, the LLM is used as a low-frequency high-level preference generator. LLM

is invoked only when a conflict-aware event trigger indicates that the current traffic state has entered a critical condition.

This event-triggered design provides a compromise between adaptive reasoning and learning stability. The LLM contributes contextual preference adaptation, while the reinforcement learning agent remains responsible for fast, low-level control execution.

Finally, the low-level learning module of the present framework is grounded in established reinforcement learning formulations and value-based deep reinforcement learning algorithms. Reinforcement learning control problems are commonly formulated using Markov decision processes [20, 21], while DQN and its stabilized variants, including dueling network architectures, Double DQN, and prioritized experience replay, provide useful foundations for stable value-based policy learning [22-25].

3. Proposed Methodology

3.1. Problem Formulation and Constrained Action Space

This study considers a signalized intersection control problem under mixed vehicle–pedestrian traffic conditions. The objective is to dynamically allocate green time to signal phases while balancing vehicle efficiency, pedestrian service fairness, and operational stability. Unlike unconstrained reinforcement learning methods that allow an agent to freely select both the signal phase and the green duration, the proposed framework follows a fixed phase sequence and allows the agent to determine only the green duration of the currently served phase. This design is introduced to improve deployment feasibility and prevent unrealistic signal behaviors such as phase starvation and reward hacking. The signal control process is formulated as a Markov Decision Process (MDP), defined as:

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma) \tag{1}$$

where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ denotes the constrained action space, $\mathcal{P}$ represents the transition probability of the traffic environment, r is the immediate reward, and $\gamma \in [0,1]$ is the discount factor. At each decision epoch t, the controller observes the current traffic state and signal phase, and then selects a green duration according to the learned policy.

The physical traffic state is represented by a five-dimensional vector:

$$s_t = \left[N_t^v, Q_t^v, C_t^s, N_t^p, Q_t^p\right] \tag{2}$$

where N_t^v denotes the total number of vehicles in the controlled area, Q_t^v denotes the number of queued vehicles, C_t^s denotes the cumulative number of vehicle stop-and-go events, N_t^p denotes the total number of pedestrians near the crossing area, and Q_t^p denotes the number of waiting pedestrians. This state representation jointly captures vehicle demand, pedestrian demand, and operational instability caused by frequent stop-and-go movements.

The current signal phase is denoted by $p_t \in \{1,2,\dots,P\}$, where P is the number of predefined signal phases. To satisfy practical traffic engineering constraints, the phase transition follows a fixed cyclic sequence:

$$p_{t+1} = (p_t \bmod P) + 1 \tag{3}$$

Therefore, the agent is not allowed to skip phases, repeatedly serve a preferred phase, or directly change the signal phase order. The effective action is defined only as the green duration assigned to the current phase:

$$a_t = d_t, \quad d_t \in \mathcal{D} \tag{4}$$

where $\mathcal{D}$ is the discrete set of feasible green durations. In this study, the duration set is defined as:

$$\mathcal{D} = \{10,20,30\}\,\text{s} \tag{5}$$

After each green interval, a fixed all-red clearance time τ_r is inserted before the next phase is activated:

$$\tau_t = d_t + \tau_r \tag{6}$$

where τ_t denotes the total execution time of one decision step. In the experimental implementation, τ_r is set to 3 s.

To support multi-objective adaptive control, a preference vector is introduced:

$$\boldsymbol{\omega}_t = \left[\omega_t^v, \omega_t^p, \omega_t^e\right] \tag{7}$$

where ω_t^v, ω_t^p, and ω_t^e represent the relative importance of vehicle efficiency, pedestrian fairness, and emission-related operational stability, respectively. The preference weights satisfy:

$$\omega_t^v + \omega_t^p + \omega_t^e = 1 \tag{8}$$
$$\omega_t^v \geq 0, \quad \omega_t^p \geq 0, \quad \omega_t^e \geq 0$$

The control objective is to learn a preference-conditioned policy $\pi(d_t \mid s_t, p_t, \boldsymbol{\omega}_t)$ that maximizes the expected discounted return:

$$\max_{\pi} \mathbb{E}_{\pi}\left[\sum_{t=0}^{T} \gamma^t r(s_t, p_t, d_t, \boldsymbol{\omega}_t)\right] \tag{9}$$

Subject to the fixed phase sequence constraint and the feasible green duration set. In this way, the controller can adapt its green-time allocation strategy to both the observed traffic state and the high-level multi-objective preference.

3.2. Preference-Conditioned Action-Value Learning

Based on the constrained signal control formulation in Section 3.1, the low-level controller is designed to learn a preference-conditioned action-value function within a DQN-style value-learning framework. The key idea is that the same physical traffic state may require different control responses depending on the high-level preferences. For example, when pedestrian waiting becomes critical, the controller should allocate green time to improve pedestrian service fairness; when vehicle queues dominate, the controller should prioritize vehicle discharge efficiency. Therefore, the

proposed framework explicitly conditions the action-value function on the multi-objective preference vector.

Before being fed into the neural network, the physical state is normalized to reduce the scale difference among heterogeneous traffic variables:

$$\bar{s}_t = \mathrm{clip}\left(\frac{s_t}{s^{\max}}, 0{,}5\right) \tag{10}$$

where $s^{\max}$ is the normalization vector for the physical state variables. The current phase p_t is encoded as a one-hot vector $e(p_t)$. The complete input to the conditional Q-network is then defined as:

$$z_t = [\bar{s}_t, e(p_t), \omega_t] \tag{11}$$

where $\bar{s}_t$ denotes the normalized physical traffic state, $e(p_t)$ denotes the encoded current phase, and ω_t denotes the high-level preference vector. The preference-conditioned action-value function is expressed as:

$$Q_\theta(s_t, p_t, d_t \mid \omega_t) = Q_\theta(z_t, d_t) \tag{12}$$

where θ represents the parameters of the Q-network, and $d_t \in \mathcal{D}$ denotes a feasible green duration. In this study, the action-value network outputs the Q-values of all candidate green durations simultaneously:

$$Q_\theta(z_t) = [Q_\theta(z_t, d_1), Q_\theta(z_t, d_2), \dots, Q_\theta(z_t, d_K)] \tag{13}$$

where $K = |\mathcal{D}|$ is the number of duration options.

To improve learning stability, a dueling network structure is adopted. The network decomposes the action-value function into a state-value term and an advantage term:

$$Q_\theta(z_t, d_t) = V_\theta(z_t) + A_\theta(z_t, d_t) - \frac{1}{K}\sum_{j=1}^{K} A_\theta\left(z_t, d_j\right) \tag{14}$$

where $V_\theta(z_t)$ estimates the value of the current traffic condition under the given preference, and $A_\theta(z_t, d_t)$ estimates the relative advantage of selecting a specific green duration. This decomposition helps the controller distinguish the overall urgency of the traffic state from the relative benefit of each duration choice.

During training, the controller follows an ε-greedy exploration strategy. The selected duration is given by:

$$d_t = \begin{cases} \textit{random duration from } \mathcal{D}, & \textit{with probability } \varepsilon \\ \underset{d\in\mathcal{D}}{argmax}\, Q_\theta(z_t, d), & \textit{with probability } 1-\varepsilon \end{cases} \tag{15}$$

After the action is executed in the microscopic simulation environment, the transition is stored in the replay buffer as:

$$(s_t, p_t, \omega_t, d_t, r_t, s_{t+1}, p_{t+1}, \omega_{t+1}, \delta_t) \tag{16}$$

where δ_t is the terminal indicator.

The Q-network is updated using a Double DQN target. The online network first selects the next duration:

$$d_{t+1}^{\star} = \underset{d' \in \mathcal{D}}{argmax} Q_\theta(z_{t+1}, d') \tag{17}$$

Then the target network evaluates the selected duration:

$$y_t = r_t + \gamma(1 - \delta_t) Q_{\bar{\theta}}(z_{t+1}, d_{t+1}^{\star}) \tag{18}$$

where $\bar{\theta}$ denotes the parameters of the target network. The training loss is defined as the priority-weighted Huber loss:

$$\mathcal{L}(\theta) = \frac{1}{B} \sum_{i=1}^{B} w_i \cdot Huber\big(y_i - Q_\theta(z_i, d_i)\big) \tag{19}$$

where $Bw_i$is the mini-batch size.The target network is softly updated by:

$$\bar{\theta} \leftarrow \tau\theta + (1 - \tau)\bar{\theta} \tag{20}$$

where τ is the soft update coefficient.

Through this preference-conditioned learning mechanism, the proposed controller does not need to train a separate policy for each combination of objective weights. Instead, a single conditional Q-network can adapt its green-duration decision based on the current traffic state, signal phase, and a dynamically updated preference vector. This design provides the learning foundation for the event-triggered LLM-based preference adaptation introduced in the next section.

3.3. Event-Triggered LLM-Based Preference Adaptation

The preference vector introduced in Section 3.1 determines the relative importance of vehicle efficiency, pedestrian fairness, and emission-related operational stability. A straightforward strategy is to update this preference vector at every decision epoch. However, frequent changes in preferences may lead to a non-stationary reward landscape for the low-level reinforcement learning agent. In addition, continuous invocation of an LLM introduces unnecessary latency and computational cost. Therefore, this study adopts an event-triggered preference adaptation mechanism, in which the LLM is used only as a low-frequency high-level preference generator rather than a direct traffic signal controller.

At each decision epoch, a human-vehicle conflict index is calculated from the current traffic state. The index considers three normalized components: pedestrian queue pressure, pedestrian waiting pressure, and vehicle queue pressure. These components are defined as:

$$c_t^{pq} = min\left(\frac{Q_t^p}{Q_{max}^p}, c_{max}\right) \tag{21}$$

$$c_t^{pw} = min\left(\frac{W_t^p}{W_{max}^p}, c_{max}\right) \quad (22)$$

$$c_t^{vq} = min\left(\frac{Q_t^v}{Q_{max}^v}, c_{max}\right) \quad (23)$$

where Q_t^p is the number of waiting pedestrians, W_t^p is the maximum pedestrian waiting time at decision epoch t, and Q_t^v is the number of queued vehicles. $Q_{\max}^p$, $W_{\max}^p$, and $Q_{\max}^v$ are normalization constants, while $c_{\max}$ is an upper clipping bound used to prevent extreme values from dominating the trigger judgment.

The overall conflict index is then defined as a weighted aggregation of the three components:

$$h_t = \alpha_{pq} c_t^{pq} + \alpha_{pw} c_t^{pw} + \alpha_{vq} c_t^{vq} \quad (24)$$

where α_{pq}, α_{pw}, and α_{vq} are non-negative coefficients representing the relative importance of pedestrian queue pressure, pedestrian waiting pressure, and vehicle queue pressure, respectively. In the experimental implementation, the three coefficients are set to 0.35, 0.40, and 0.25, respectively.

To avoid frequent switching between normal and emergency states, a hysteresis-based event trigger is introduced. Let $e_t \in \{0,1\}$ denote whether the system is in an emergency state. Two thresholds are used: an activation threshold η_{on} and a recovery threshold η_{off}, where $\eta_{on} > \eta_{off}$. The emergency state is updated as follows:

$$\begin{aligned} e_t &= 1, \quad \textit{if } e_{t-1} = 0 \textit{ and } h_t \geq \eta_{on} \\ e_t &= 0, \quad \textit{if } e_{t-1} = 1 \textit{ and } h_t \leq \eta_{off} \\ e_t &= e_{t-1}, \quad \textit{otherwise} \end{aligned} \quad (25)$$

In this study, η_{on} and η_{off} are set to 0.40 and 0.20, respectively. The hysteresis band prevents the controller from repeatedly invoking the LLM when the conflict index fluctuates around a single threshold. When the conflict index exceeds the activation threshold, the LLM is awakened to generate a new preference vector. When the conflict index falls below the recovery threshold, the system exits the emergency state and returns to a neutral or rule-based preference.

The LLM receives the current physical traffic state, the maximum pedestrian waiting time, and the previous preference vector as contextual inputs. Its output is constrained to a normalized three-dimensional preference vector:

$$\begin{aligned} \omega_t^{LLM} &= g_{LLM}\left(s_t, W_t^p, \omega_{t-1}\right) \\ \omega_t^{LLM} &= \left[\omega_t^v, \omega_t^p, \omega_t^e\right], \quad \omega_t^v + \omega_t^p + \omega_t^e = 1 \end{aligned} \quad (26)$$

where $g_{LLM}(\cdot)$ denotes the LLM-based preference generator. The prompt instructs the LLM to increase the pedestrian fairness weight when pedestrian backlog or maximum waiting time is high, increase the vehicle efficiency weight when vehicle queue spillback risk is high, and increase the emission-related stability weight under relatively stable under-saturated conditions. The LLM is explicitly prohibited from selecting signal phases or green durations; it only outputs high-level preference weights for the low-level reinforcement learning controller.

When the LLM is unavailable or when the event trigger is not activated, a deterministic rule-based fallback is used to maintain robustness. The fallback preference is defined as:

$$\omega_t^{rule} = \begin{cases} [0.20,0.65,0.15], & \text{if } W_t^p \geq 45 \text{ s or } Q_t^p \geq 6 \\ [0.65,0.20,0.15], & \text{if } Q_t^v \geq 10 \\ [0.35,0.35,0.30], & \text{otherwise} \end{cases} \quad (27)$$

The final preference vector used by the low-level controller is determined by the trigger status:

$$\omega_t = \begin{cases} \omega_t^{LLM}, & \text{if the activation condition is satisfied} \\ \omega_t^{cache}, & \text{if the system remains in the emergency state} \\ \omega_t^{rule}, & \text{otherwise} \end{cases} \quad (28)$$

where ω_t^{cache} denotes the most recently accepted LLM-generated preference. This cache mechanism keeps the preference signal stable during an emergency interval and avoids rapid reward reshaping. As a result, the proposed event-triggered adaptation mechanism enables the controller to respond to critical human-vehicle conflicts while preserving the convergence stability of the low-level reinforcement learning process.

3.4. Multi-Objective Reward Design with Pedestrian Fairness Penalty

The reward function is designed to guide the low-level reinforcement-learning controller toward balanced signal-timing decisions across different high-level preference vectors. Since the proposed framework aims to jointly consider vehicle efficiency, pedestrian service fairness, and operational stability, the immediate reward is decomposed into three interpretable components: a vehicle-efficiency component, a pedestrian-fairness component, and an emission-related stability component. The final reward is obtained by weighting these components with the preference vector generated by the high-level module.

At decision step t, let N_t^{pass} denote the number of vehicles that pass through the controlled area during the current control interval, Q_t^v denote the number of queued vehicles, Q_t^p denote the number of waiting pedestrians, W_t^{max} denote the maximum pedestrian waiting time observed at the current step, and ΔC_t^s denote the newly generated vehicle stop-and-go events. To reduce the sensitivity of learning to the absolute magnitudes of traffic variables, the main reward-related quantities are normalized as follows:

$$N_t^{pass,n} = min\left(\frac{N_t^{pass}}{N_{max}^{pass}}, 1.5\right) \quad (29)$$

$$Q_t^{v,n} = min\left(\frac{Q_t^v}{Q_{max}^v}, 1.5\right) \quad (30)$$

$$Q_t^{p,n} = min\left(\frac{Q_t^p}{Q_{max}^p}, 1.5\right) \quad (31)$$

$$\Delta C_t^{s,n} = min\left(\frac{\Delta C_t^s}{C_{max}^s}, 1.5\right) \quad (32)$$

The vehicle-efficiency component rewards vehicle discharge and penalizes excessive vehicle queues:

$$r_t^v = \rho_{pass} N_t^{pass,n} - \rho_{vq} Q_t^{v,n} \quad (33)$$

where ρ_{pass} controls the positive contribution of vehicle throughput and ρ_{vq} controls the penalty for vehicle queue accumulation.

The pedestrian-fairness component penalizes both pedestrian queue accumulation and extreme pedestrian waiting. It is defined as:

$$r_t^p = -\rho_{pq} Q_t^{p,n} - \rho_w W_t^{max,n} \tag{34}$$

where the nonlinear waiting penalty is calculated as:

$$W_t^{max,n} = min\left[\left(\frac{W_t^{max}}{W_{max}}\right)^2, C_{fair}\right] \tag{35}$$

The quadratic form assigns a stronger penalty to extreme pedestrian waiting than to short temporary waiting. This design is introduced to discourage policies that achieve high vehicle efficiency by repeatedly postponing pedestrian service.

The emission-related operational stability component is represented by newly generated stop-and-go events:

$$r_t^e = -\rho_s \Delta C_t^{s,n} \tag{36}$$

where ρ_s is the penalty coefficient for stop-and-go movements. In this study, this term is used as an operational proxy for emission-related instability because frequent stopping and restarting generally indicate unstable traffic progression and higher energy consumption.

Given the preference vector $\omega_t = \left[\omega_t^v, \omega_t^p, \omega_t^e\right]$, the final immediate reward is defined as:

$$r_t = clip\{\kappa\left(\omega_t^v r_t^v + \omega_t^p r_t^p + \omega_t^e r_t^e\right), -R_{max}, R_{max}\} \tag{37}$$

where κ is a reward scaling factor and R_{max} is the clipping bound used to avoid excessively large temporal-difference errors during training.

This reward design has two important properties. First, the reward remains preference-conditioned: changing ω_t changes the relative importance of vehicle discharge, pedestrian service, and operational stability without changing the structure of the low-level controller. Second, the pedestrian fairness penalty explicitly targets the maximum waiting time rather than only the average waiting level. This is important because average-based pedestrian indicators may hide individual extreme waiting cases, whereas the maximum waiting time directly reflects the service risk faced by vulnerable road users.

3.5. Training and Online Decision Procedure

The preceding subsections define the constrained signal control problem, the preference-conditioned action-value function, the event-triggered LLM-based preference adaptation mechanism, and the multi-objective reward function. This subsection summarizes how these components are integrated during training and online decision-making.

Input: traffic simulation environment, duration set, replay buffer, Q-networks, trigger thresholds, and all-red clearance time. Output: trained preference-conditioned control policy.

i. Initialize the online Q-network Q_θ and the target Q-network Q_{θ^-}.
ii. Initialize the replay buffer $\mathcal{R}$ and the cached preference vector ω_0.
iii. For each training episode, reset the simulation environment and obtain the initial state s_0.
iv. At each decision epoch t, observe the physical state s_t, the current phase p_t, and the maximum pedestrian waiting time.
v. Compute the human-vehicle conflict index and update ω_t using the event-triggered LLM-based preference adaptation module.
vi. Select a green duration d_t from $\mathcal{D}$ using the epsilon-greedy strategy.
vii. Execute d_t under the fixed phase sequence constraint and insert the all-red clearance time τ_r.
viii. Observe the reward r_t, the next state s_{t+1}, the next phase p_{t+1}, and the terminal indicator δ_t.
ix. Store the transition in the replay buffer.
x. If the replay buffer is sufficiently large, sample a mini-batch and update Q_θ by minimizing the temporal-difference loss.
xi. Update the target network Q_{θ^-} using soft parameter replacement.
xii. Repeat steps 4-11 until the episode terminates.

During online testing or deployment, the learning-related operations are removed. The trained Q-network is fixed, and the exploration rate is set to zero. At each decision epoch, the controller observes the current state, updates or retains the preference vector according to the event-triggered mechanism, and selects the green duration with the maximum Q-value:

$$d_t^* = \underset{d\in\mathcal{D}}{argmax} Q_\theta(s_t, p_t, \omega_t, d) \tag{38}$$

The selected green duration is then executed under the same fixed phase sequence and all-red clearance constraints.

4. Experimental Design

4.1. Simulation Environment and Study Scenarios

The experiments are conducted in a microscopic traffic simulation environment based on PTV Vissim. A Python-based control interface communicates with the simulator through the COM automation interface, enabling the reinforcement learning controller to reset the simulation, assign random seeds, obtain real-time traffic states, execute green-time decisions, and record performance indicators. This closed-loop simulation design provides a controllable and repeatable environment for evaluating adaptive signal control strategies under mixed vehicle-pedestrian traffic conditions.

Each simulation episode represents a 900s operational period. Before control decisions are evaluated, the simulator is advanced through an initial warm-up period to initialize traffic states and reduce the influence of empty-network conditions. During the controlled period, the signal controller follows the fixed cyclic phase sequence defined in Section 3.1. At each decision epoch, the controller determines only the green duration of the current phase, while the simulator enforces the phase

transition and the all-red clearance interval. This ensures that all compared methods are evaluated under the same signal safety constraints.

The study uses three signalized intersection scenarios developed from Budapest intersection cases. To evaluate both policy learning and cross-scenario generalization, two scenarios are used for training, and one unseen scenario is reserved for zero-shot testing. Specifically, C2GQ+GC S3 and F393+MC4 S3 are used as training scenarios, while G22W+7C S3 is used as the test scenario (see Table 1). The zero-shot testing setting is designed to assess whether the learned policy can maintain robust performance when transferred to a scenario not observed during training.

Table 1
Scenario split used in the experiments

Scenario	File name	Role	Purpose
C2GQ+GC S3	C2GQ+GC S3	Training	Policy learning under one observed traffic scenario
F393+MC4 S3	F393+MC4 S3	Training	Policy learning under another observed traffic scenario
G22W+7C S3	G22W+7C S3	Testing	Zero-shot evaluation on an unseen traffic scenario

For a fair comparison, the same simulation networks, demand settings, random seed protocol, phase sequence constraints, and evaluation horizon are applied to all baseline and ablation methods. Episode-level logs are used to report aggregate indicators, including throughput, average delay, average stops, maximum pedestrian waiting time, and total stops.

4.2. Training and Testing Protocol

A two-stage experimental protocol is adopted to evaluate the proposed ET-LCRL framework. The first stage trains the reinforcement-learning-based controllers on two signalized intersection scenarios, while the second stage evaluates the trained policies on an unseen scenario without any further fine-tuning. This setting is designed to assess both learning performance and cross-scenario generalization capability.

Table 2 summarizes the complete training, evaluation, testing, and baseline comparison protocol used in the experiments. It clarifies the purpose of each experimental stage and the corresponding implementation setting. For non-learning baselines, such as fixed-time control and max-pressure control, the same simulation horizon and random seed setting are used to ensure fair comparison. For RL-based variants, the same training and testing protocol is applied wherever applicable. All episode-level and step-level trajectories are recorded for subsequent statistical analysis, including vehicle efficiency, pedestrian fairness, operational stability, and LLM invocation characteristics.

To improve stability, the target network is updated via soft-update, and the best model is selected based on the mean evaluation reward from periodic evaluations.

Periodic evaluation is conducted during training without selecting exploratory actions. In the evaluation episodes, the exploration rate is set to zero, and the agent always selects the green duration with the highest estimated action value. The evaluation results are used only for model selection and monitoring training progress; they are not used for gradient updates. After training, the best-performing checkpoint of each trainable variant is loaded for final testing. The final testing stage is conducted only on the unseen scenario. No policy updates, replay buffer updates, or parameter fine-tuning is performed during testing. This zero-shot testing protocol ensures that the reported performance reflects each controller's generalization ability rather than adaptation to the test scenario.

For non-learning baselines, such as fixed-time control and max-pressure control, the same simulation horizon and random seed setting are used to ensure fair comparison. For RL-based variants, the same training and testing protocol is applied wherever applicable.

Table 2
Training and testing protocol used in the experiments

Stage	Purpose	Main setting
Training	Learn the preference-conditioned control policy	Trainable RL variants interact with two training scenarios; exploratory action selection and replay-buffer updates are enabled.
Periodic evaluation	Monitor convergence and select the best checkpoint	Evaluation is conducted at fixed intervals with exploration disabled; evaluation episodes are not used for gradient updates.
Final testing	Assess zero-shot generalization	The best checkpoint is loaded and evaluated on the unseen test scenario; no additional learning or fine-tuning is allowed.
Baseline comparison	Provide reference performance	Fixed-time, max-pressure, vanilla DQN, and other comparison controllers are evaluated under the same simulation horizon and seed protocol.

The key protocol settings used in the current implementation are summarized in Table 3. These settings define the experimental procedure and do not depend on the final numerical results.

Table 3
Key settings for training and zero-shot testing

Item	Setting
Simulation horizon per episode	900 s
Warm-up period	30 s
Training scenarios	C2GQ+GC S3 and F393+MC4 S3
Unseen test scenario	G22W+7C S3
Training episodes	1,200 episodes in the current implementation
Evaluation frequency during training	Every 25 episodes
Evaluation episodes per evaluation round	3 episodes
Final test episodes	20 episodes per method
Base random seed	42
Checkpoint strategy	Latest checkpoint saved periodically; the best checkpoint is selected according to evaluation reward
Testing policy	Greedy action selection with exploration disabled

4.3. Baseline Methods

To evaluate the effectiveness of the proposed ET-LCRL framework, this study compares it with both conventional signal control methods and reinforcement learning-based baselines (see Table 4). The comparison is designed to examine whether the proposed framework can improve pedestrian-aware multi-objective signal control beyond fixed timing, pressure-based heuristic control, and standard deep reinforcement learning. In addition, several ablation variants are included to isolate the effects of event-triggered LLM preference adaptation, preference conditioning, and the pedestrian fairness penalty.

All compared methods are implemented on the same simulation networks, with a fixed phase-sequence constraint, a feasible green-duration set, an episode length, and a random-seed protocol. For reinforcement learning-based methods, the same training and testing protocol described in Section 4.2 is adopted unless otherwise stated. For non-learning methods, the controller is directly

evaluated in the same test scenario without gradient-based training. This design ensures that observed performance differences are primarily due to the control strategy rather than to differences in traffic demand, simulation horizon, or signal safety constraints.

Table 4
Baseline methods used for performance comparison

Method	Type	Main control logic	Purpose of comparison
Fixed-Time	Conventional baseline	Uses a predefined timing plan and fixed cyclic phase sequence.	Provides a simple engineering reference without adaptive response.
Max-Pressure	Heuristic adaptive baseline	Selects or favors service according to traffic pressure or queue imbalance.	Tests whether the proposed method can balance efficiency with pedestrian service beyond pressure-based control.
Vanilla-DQN	RL baseline	Learns green-duration decisions from traffic states without LLM guidance or preference conditioning.	Evaluates the benefit of preference-conditioned multi-objective learning over standard DQN.
ET-LCRL	Proposed method	Uses event-triggered LLM-based preference adaptation and preference-conditioned Q-learning under constrained phase sequencing.	Serves as the proposed method for comparison with other control methods. Fixed-time, max-pressure, vanilla DQN, and ablation variants are evaluated under the same simulation horizon and seed protocol.

The fixed-time controller represents a commonly used non-adaptive signal control strategy. It follows the predefined phase sequence and applies fixed green durations throughout the simulation. Although this method is easy to deploy and interpret, it cannot respond to real-time fluctuations in vehicle and pedestrian demand. The max-pressure controller is included as an adaptive heuristic baseline. It is designed to prioritize traffic movements with higher pressure or queue imbalance, and therefore usually favors vehicle throughput and queue dissipation. However, because the pressure objective is primarily vehicle-oriented, it may not sufficiently account for pedestrian waiting fairness under conflicting demand conditions.

The Vanilla-DQN baseline removes both LLM-based preference adaptation and preference conditioning. It uses the physical traffic state and current phase information to learn green-duration decisions, but it does not receive the multi-objective preference vector. Therefore, the comparison between Vanilla-DQN and ET-LCRL evaluates whether preference-conditioned learning and high-level preference adaptation improve multi-objective control performance.

4.4. Evaluation Metrics

To provide a balanced assessment of the proposed controller, the evaluation metrics are designed to cover vehicle efficiency, pedestrian service fairness, operational stability, and LLM invocation cost. Since the objective of this study is not to optimize a single traffic performance indicator, relying only on vehicle throughput or average delay may lead to a biased interpretation. Therefore, the final comparison considers both conventional traffic efficiency indicators and pedestrian-oriented fairness indicators.

The primary vehicle-efficiency metric is throughput, defined as the number of vehicles that successfully depart from the controlled area during one episode:

$$T_{veh} = N_{dep} \tag{39}$$

where N_dep denotes the number of departed vehicles. A higher throughput indicates that the controller is able to discharge more vehicles within the same simulation horizon.

Average delay is used to quantify the travel efficiency of vehicles in the network. It is obtained from the network performance measurement module of the microscopic simulation environment. A lower average delay indicates smoother vehicle progression and less congestion.

To evaluate operational stability, the average number of vehicle stop-and-go events is calculated as:

$$S_{avg} = \frac{S_{total}}{max(1,T_{veh})} \tag{40}$$

where S_total denotes the cumulative number of vehicle stop-and-go events during an episode. This metric reflects the degree of instability in traffic operations. A lower value indicates fewer repeated stop-and-start movements, which is beneficial for both traffic smoothness and emission-related operational stability.

Pedestrian fairness is evaluated using the maximum pedestrian waiting time during an episode:

$$W_{ped}^{max} = \underset{k\in P}{max} W_k \tag{41}$$

where W_k denotes the waiting time of pedestrian k, and P denotes the set of pedestrians observed in the controlled area. Different from average waiting time, this indicator focuses on the most underserved pedestrians and is therefore more suitable for evaluating extreme unfairness. A lower maximum pedestrian waiting time indicates that the controller can more effectively prevent long pedestrian delays.

In addition to the final traffic performance metrics, several diagnostic indicators are recorded to analyze the proposed framework's internal behavior. The first diagnostic indicator is the mean human-vehicle conflict index over an episode:

$$H_{mean} = \frac{1}{K}\sum_{t=1}^{K} h_t \tag{42}$$

where h_t is the conflict index at decision step t, and K is the total number of decision steps in an episode. The maximum conflict index is also recorded to identify whether a controller produces severe transient conflict states. The second diagnostic indicator is the number of LLM queries per episode. This metric measures how frequently the high-level LLM preference generator is activated. It is particularly important for evaluating the computational practicality of the proposed event-triggered mechanism. A lower number of LLM queries indicates lower online inference cost and lower communication latency. The third diagnostic indicator is the number of preference switches. It records how often the preference vector changes during an episode. This metric is used to evaluate whether the controller produces stable high-level optimization priorities. Excessive preference switching may indicate non-stationary control objectives and can destabilize low-level reinforcement learning.

Among these indicators, throughput, average delay, average stops, and total stops evaluate vehicle-oriented operational performance; maximum pedestrian waiting time evaluates fairness for vulnerable road users; and LLM queries, preference switches, and conflict-related metrics are used

to explain the internal behavior of the proposed ET-LCRL framework. This metric design enables the experimental analysis to assess whether the proposed controller achieves a balanced trade-off among efficiency, fairness, stability, and computational practicality.

4.5. Implementation Details and Parameter Settings

Table 5 indicate the experimental configurations. All methods were implemented in Python and interacted with the microscopic simulation environment through the Vissim COM interface. The reinforcement learning modules were implemented with PyTorch. To ensure a fair comparison, the same simulation duration, state definition, phase-sequence constraint, feasible green-duration set, and random-seed protocol were used for all compared methods, except for the intentionally modified modules in the ablation variants. Each episode corresponds to a complete 900 s simulation run. During training, the agent periodically performs evaluation episodes without exploratory actions, and the checkpoint with the best evaluation reward is saved for final testing. During testing, the learned policy is fixed, and no further gradient update is performed on the unseen scenario.

Table 5
Main experimental configuration

Item	Setting
Simulation software	PTV Vissim microscopic traffic simulation
Python interface	Vissim COM interface
RL implementation	PyTorch-based DQN implementation
Simulation period per episode	900 s
Warm-up period	30 s
Simulation resolution	5 simulation steps per second
Training episodes	1200 in the current experimental configuration
Evaluation frequency	Every 25 training episodes
Evaluation episodes	3 episodes per evaluation round
Testing episodes	20 episodes on the unseen scenario
Random seed	Base seed = 42; episode-level seeds are generated by offsetting the base seed
Device	CUDA if available; otherwise CPU

The reinforcement learning hyperparameters are reported in Table 6. A prioritized replay buffer is used for the proposed ET-LCRL and other conditioned RL variants.

Table 6
Reinforcement learning hyperparameters

Parameter	Value
Discount factor gamma	0.99
Learning rate	1 x 10^-3
Batch size	64
Replay buffer capacity	20,000 transitions
Minimum buffer size before learning	512 transitions
Initial exploration rate epsilon	1.00
Minimum exploration rate	0.05
Epsilon decay factor	0.995
Soft target update coefficient tau	0.01
Prioritized replay alpha	0.60
Initial importance-sampling beta	0.40
PER beta annealing frames	100,000

The loss function is the Huber loss, and gradient clipping is applied to improve numerical stability. The target network is updated using soft updates, which helps reduce abrupt changes in the bootstrapped target values.

The state, action, and preference-related settings are summarized in Table 7. The physical traffic state is normalized before entering the neural network. The action space contains three feasible green duration options, and the phase sequence is enforced by the environment. Therefore, the agent is trained to allocate green time under practical constraints of signal operation rather than to freely select arbitrary phases.

Table 7
Main experimental configuration

Item	Setting
Physical state dimension	5
Phase encoding dimension	2
Preference vector dimension	3
Number of signal phases	2
Feasible green durations	{10 s, 20 s, 30 s}
Minimum green time	10 s
Maximum green time	30 s
All-red clearance time	3 s
State normalization scales	[60, 30, 300, 40, 20]
Preference dimensions	Vehicle efficiency, pedestrian fairness, emission-related operational stability

The event-triggered LLM adaptation settings are given in Table 8. The conflict index is computed from pedestrian queue pressure, pedestrian waiting pressure, and vehicle queue pressure. A hysteresis band is used to avoid frequent preference switching. When the conflict index exceeds the upper threshold, the LLM is activated to update the preference vector. When the conflict index falls below the lower threshold, the system exits the emergency state and returns to a balanced preference unless a new emergency is detected. If the LLM is unavailable or returns invalid output, a deterministic, rule-based fallback is used to ensure a valid, normalized preference vector.

Table 8
Event-triggered LLM adaptation settings

Parameter	Value
LLM role	High-level preference generator only
LLM output	Normalized weights for vehicle efficiency, pedestrian fairness, and operational stability
LLM model in implementation	tongyi-xiaomi-analysis-pro
API interface	OpenAI-compatible DashScope interface
Trigger threshold eta_on	0.40
Recovery threshold eta_off	0.20
LLM cooldown steps	0 in the current configuration
Pedestrian queue normalization constant	10.0
Pedestrian waiting normalization constant	60.0 s
Vehicle queue normalization constant	20.0
Pedestrian queue coefficient	0.35
Pedestrian waiting coefficient	0.40
Vehicle queue coefficient	0.25
Rule-based fallback	Enabled

5. Results and Analysis

This section evaluates the proposed ET-LCRL framework from three complementary perspectives: overall traffic performance, pedestrian fairness and operational stability, and the behavior of the event-triggered LLM-guided preference-adaptation mechanism. All reported test results are averaged over ten runs on the unseen test scenario under the same simulation horizon, traffic-demand setting, signal constraints, and random-seed protocol. Because the proposed controller is explicitly multi-objective, the results are interpreted jointly rather than by ranking the methods according to a single traffic indicator.

5.1. Overall Comparison of Control Performance

The overall comparison first considers episode reward, throughput, and average vehicle delay. These indicators characterize aggregate control performance and vehicle-oriented efficiency, but they do not by themselves measure pedestrian fairness or stop-and-go stability. Consequently, a method with the highest reward or throughput should not automatically be regarded as the best overall controller if that gain is accompanied by unfavorable service or stability outcomes.

The numerical results used in the subsequent comparisons are summarized in Table 9, which reports both performance outcomes and mechanism-level diagnostic indicators for the seven controllers.

Table 9
Summary of performance and diagnostic indicators on the unseen test scenario

Metric	Vanilla DQN	ET-LCRL	Fixed-time	Max-pressure	ET-LCRL w/o fairness	Fixed preference	Rule-only preference
Reward	-27.65	**22.11**	-45.00	87.60	10.80	-32.09	35.48
Throughput	366.7	**317.7**	355.3	410.5	325.1	339.6	314.8
Avg. delay (s)	8.099	**7.468**	17.798	11.293	12.683	7.304	16.043
Avg. stops	0.434	**0.432**	0.720	0.555	0.525	0.414	0.607
Max ped. wait (s)	35.9	**32.2**	39.2	37.3	33.6	38.5	36.3
Total stops	159.4	**137.0**	255.7	228.1	140.7	170.5	190.8
LLM queries	0.0	**3.3**	0.0	0.0	5.0	0.0	0.0
Preference switches	0.0	**20.3**	0.0	0.0	21.0	0.0	26.5
Conflict mean	0.2489	**0.2031**	0.2856	0.2970	0.2225	0.2183	0.2627
Conflict max	0.5116	**0.5500**	0.5846	0.5689	0.5223	0.5460	0.5010

Note: Higher values are better for reward and throughput; lower values are better for delay, stop, pedestrian-waiting, and conflict measures. LLM queries and preference switches are diagnostic indicators.

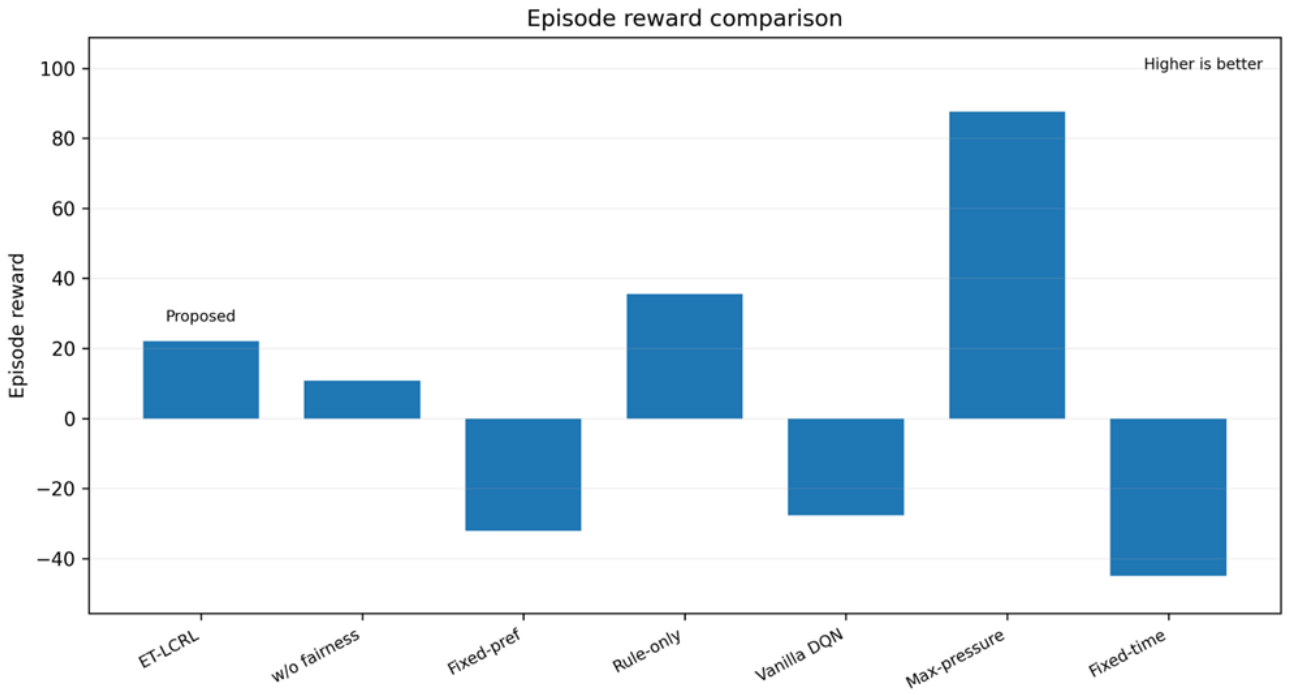

Fig. 1. Episode Reward Comparison Across Controllers

Fig. 1 compares the mean episode reward of the seven controllers. Max-pressure achieves the highest reward, with a mean value of 87.60, followed by Rule-only preference at 35.48. ET-LCRL obtains a positive mean reward of 22.11 and ranks third, outperforming ET-LCRL w/o fairness (10.80), Vanilla DQN (-27.65), Fixed preference (-32.09), and Fixed-time control (-45.00). The strong Max-pressure result is consistent with its pressure-responsive logic, which aggressively serves movements with high queue pressure. However, episode reward is an aggregate optimization signal and does not independently describe pedestrian service quality or the stability of vehicle progression.

Vehicle throughput shows a similar efficiency-oriented pattern. Max-pressure achieves the highest throughput, at 410.5 vehicles per episode, followed by Vanilla DQN at 366.7 vehicles per episode. ET-LCRL processes 317.7 vehicles, which is approximately 22.6% lower than Max-pressure and 13.4% lower than Vanilla DQN. This reduction represents the principal efficiency trade-off of the proposed controller. Unlike strongly vehicle-oriented strategies, ET-LCRL allocates control priority among vehicle efficiency, pedestrian fairness, and operational stability rather than maximizing discharge alone. The comparison with ET-LCRL w/o fairness is particularly informative: removing the fairness-oriented component raises throughput only from 317.7 to 325.1 vehicles, an increase of approximately 2.3%.

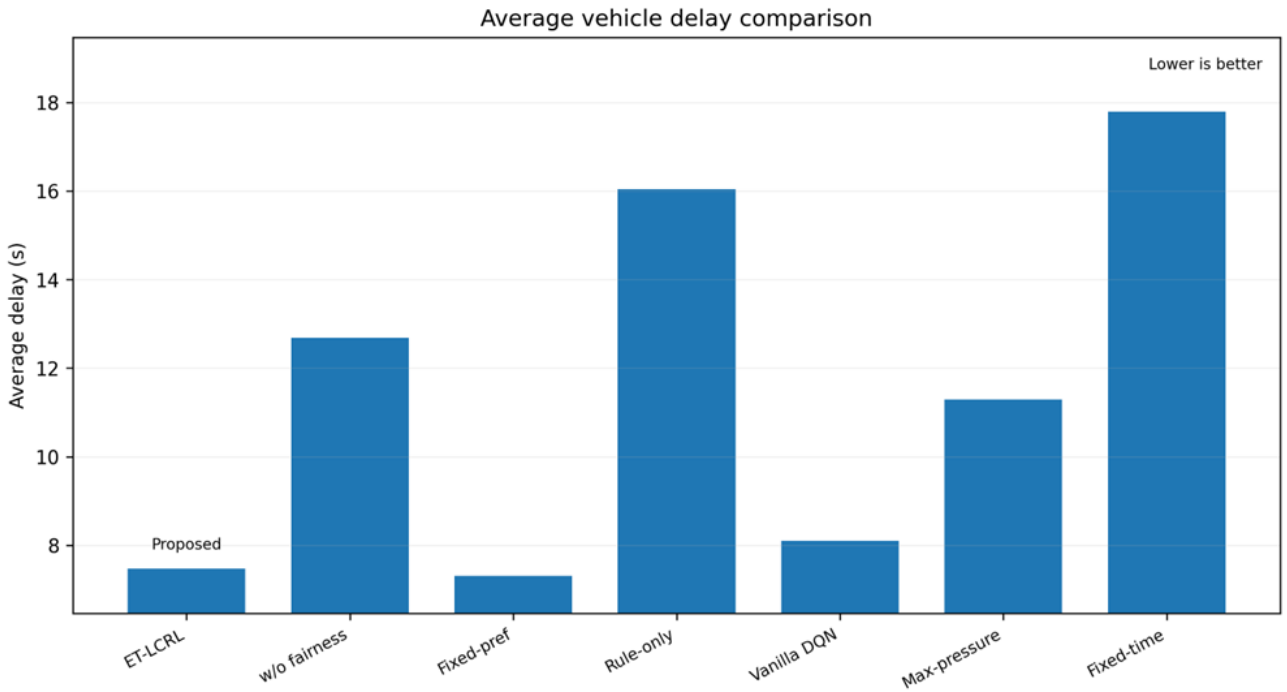

Fig. 2. Average Vehicle Delay Comparison Across Controllers

Fig. 2 presents the average vehicle delay. Fixed preference achieves the lowest mean delay of 7.304 s, while ET-LCRL ranks second at 7.468 s, only 0.164 s higher. Vanilla DQN records 8.099 s, followed by Max-pressure at 11.293 s, ET-LCRL w/o fairness at 12.683 s, Rule-only preference at 16.043 s, and Fixed-time control at 17.798 s. Compared with Vanilla DQN, ET-LCRL reduces average delay by approximately 7.8%, and the reduction relative to Fixed-time reaches approximately 58.0%.

The ablation result further indicates that introducing the complete fairness-aware framework does not simply exchange pedestrian benefits for deteriorating vehicle efficiency. Compared with ET-LCRL w/o fairness, the complete ET-LCRL reduces average delay from 12.683 s to 7.468 s, increases episode reward from 10.80 to 22.11, and sacrifices only approximately 2.3% of throughput. At the same time, the comparison with Fixed preference shows that a static objective weighting can achieve very low delay when it matches the prevailing conditions, but it lacks an explicit mechanism to adjust the optimization priority as pedestrian-vehicle conflicts evolve.

Overall, ET-LCRL does not dominate every vehicle-oriented indicator. Its throughput is lower than that of several efficiency-oriented controllers, and its reward remains below Max-pressure and Rule-only preference. Nevertheless, it preserves near-best vehicle-delay performance while avoiding an exclusively throughput-oriented policy. The value of this trade-off is examined next through the lenses of pedestrian fairness and stop-related operational stability.

5.2. Pedestrian Fairness and Operational Stability

A central objective of ET-LCRL is to prevent improved vehicle efficiency from being achieved at the expense of vulnerable road users or unstable stop-and-go operation. Pedestrian fairness is therefore evaluated using the maximum pedestrian waiting time, which focuses on the most underserved pedestrian rather than the average pedestrian condition. Operational stability is

evaluated using stop-related indicators, because frequent stopping and restarting indicate discontinuous traffic progression and may offset the apparent benefit of high throughput.

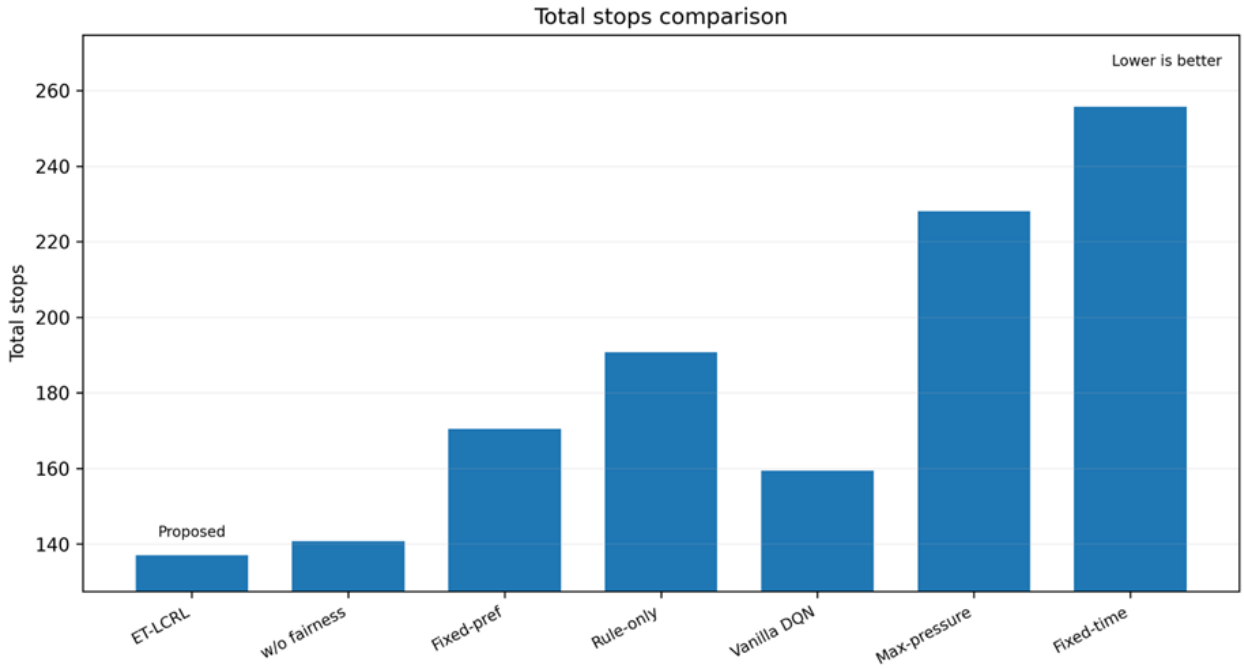

Fig. 3. Total Stop Comparison Across Controllers

Fig. 3 shows the total number of vehicle stops. ET-LCRL achieves the lowest value among all controllers, with 137.0 stops per episode. ET-LCRL w/o fairness follows at 140.7, while Vanilla DQN and Fixed preference record 159.4 and 170.5 stops, respectively. Rule-only preference produces 190.8 stops, whereas Max-pressure and Fixed-time generate substantially higher values of 228.1 and 255.7. Relative to Max-pressure, ET-LCRL reduces total stops by approximately 39.9%, and the reduction relative to Fixed-time reaches approximately 46.4%.

The total-stop comparison is important because it reveals an operational consequence that is not visible from throughput alone. Max-pressure processes substantially more vehicles than ET-LCRL, but it also generates approximately 66.5% more total stopping events. Thus, aggressive pressure-based discharge does not necessarily translate into smoother traffic progression. ET-LCRL also records an average-stop value of 0.432, close to the best-performing Fixed-preference controller at 0.414, indicating that the improved system-level stability is not achieved by increasing the typical stopping burden per vehicle.

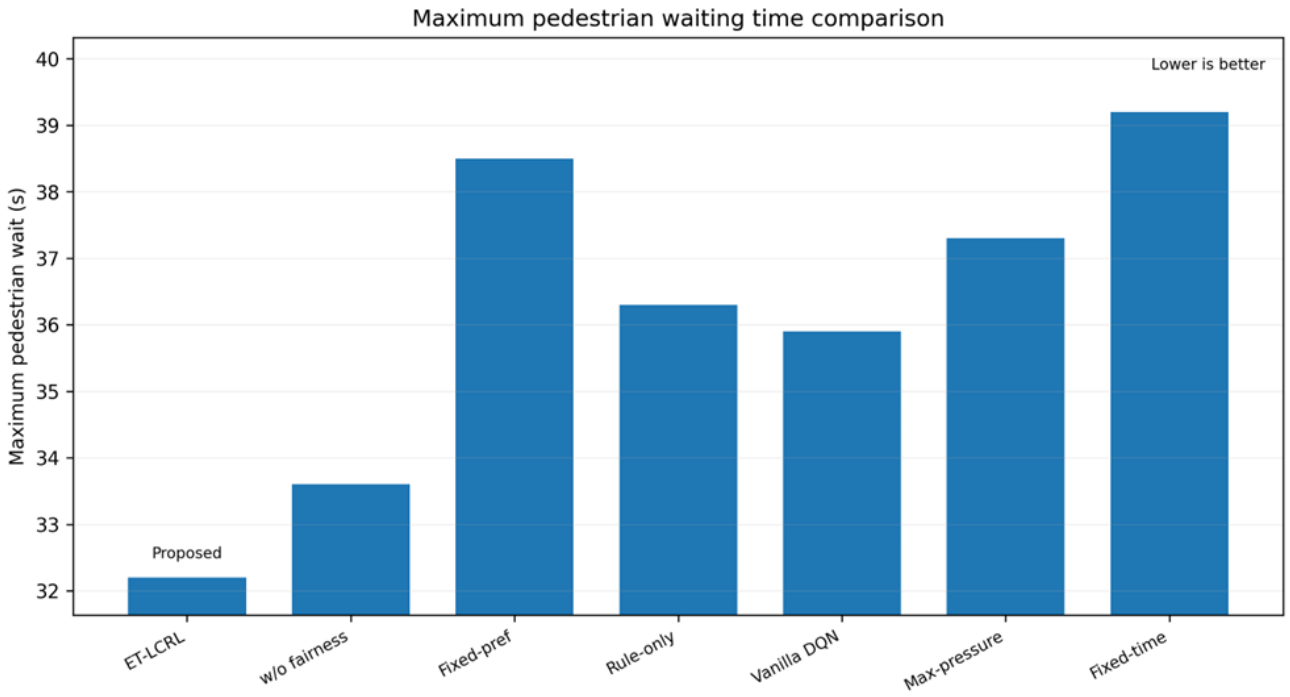

Fig. 4. Maximum Pedestrian Waiting Time Comparison Across Controllers

Fig. 4 presents the maximum pedestrian waiting time. ET-LCRL achieves the lowest value among all seven methods, at 32.2 s. ET-LCRL w/o fairness records 33.6 s, followed by Vanilla DQN at 35.9 s and Rule-only preference at 36.3 s. Max-pressure, Fixed preference, and Fixed-time produce higher maximum waiting times of 37.3, 38.5, and 39.2 s, respectively. Relative to Fixed-time, ET-LCRL reduces the maximum pedestrian waiting time by approximately 17.9%; the corresponding reductions relative to Max-pressure, Fixed preference, and Vanilla DQN are approximately 13.7%, 16.4%, and 10.3%, respectively.

The comparison with Fixed preference is particularly notable. Although fixed preference results in a marginally lower vehicle delay, its maximum pedestrian waiting time is approximately 19.6% higher than that of ET-LCRL. This indicates that strong vehicle-delay performance under a fixed weighting does not necessarily guarantee equitable pedestrian service. The fairness ablation provides additional support: removing the fairness component increases the maximum pedestrian waiting time from 32.2 to 33.6 s, whereas the complete framework improves both the maximum pedestrian waiting time and the total number of stops.

Taken together, ET-LCRL achieves the lowest maximum pedestrian waiting time, the lowest total stops, and the second-lowest average vehicle delay. Its main disadvantage remains its lower throughput compared with strongly vehicle-oriented controllers. The results reveal a clear multi-objective trade-off: ET-LCRL sacrifices part of the maximum discharge capacity in exchange for improved extreme pedestrian service and smoother system-level operation, rather than improving one road-user group by simply transferring delay or instability to another.

5.3 Mechanism Analysis of LLM-guided Preference Adaptation

The traffic-performance results alone do not establish whether the LLM-guided preference mechanism operates in accordance with its design motivation. ET-LCRL is intended to invoke the LLM

only when the pedestrian-vehicle service conflict becomes critical, while the low-level reinforcement-learning controller continues to execute green-duration decisions at every decision epoch. The diagnostic results support this separation of responsibilities. ET-LCRL records a mean conflict index of 0.2031, the lowest value among the seven controllers. The corresponding values are 0.2183 for Fixed preference, 0.2225 for ET-LCRL w/o fairness, 0.2489 for Vanilla DQN, 0.2627 for Rule-only preference, 0.2856 for Fixed-time, and 0.2970 for Max-pressure. Relative to Max-pressure and Rule-only preference, ET-LCRL reduces the mean conflict index by approximately 31.6% and 22.7%, respectively, indicating a lower sustained level of pedestrian-vehicle service conflict during the test episodes.

The maximum conflict index provides a complementary view of short-lived peak conditions. ET-LCRL records a mean maximum conflict index of 0.5500. This value is not the lowest among the comparison methods: Rule-only preference, Vanilla DQN, and ET-LCRL w/o fairness record 0.5010, 0.5116, and 0.5223, respectively. The result, therefore, does not indicate that ET-LCRL eliminates all transient conflict peaks. Instead, the combination of the lowest mean conflict index with competitive overall traffic performance suggests that the proposed mechanism is more effective at limiting sustained conflict exposure than at suppressing every isolated peak. This distinction is important because it avoids interpreting the event trigger as a hard safety constraint; its role is to adapt optimization priorities when the conflict state becomes sufficiently critical.

The LLM-query and preference-switching statistics further clarify how the high-level adaptation operates. ET-LCRL invokes the LLM an average of 3.3 times per test episode and records 20.3 preference switches. The ET-LCRL w/o fairness variant requires 5.0 LLM queries and 21.0 switches, whereas Rule-only preference performs no LLM queries but produces 26.5 preference switches. Fixed preference, Vanilla DQN, Max-pressure, and Fixed-time do not use adaptive LLM guidance and therefore record no LLM queries; Fixed preference also retains a constant preference vector. Compared with the w/o-fairness ablation, the complete ET-LCRL framework uses approximately 34.0% fewer LLM queries. Compared with Rule-only preference, it produces approximately 23.4% fewer preference switches. Because the switch count records changes in the preference vector regardless of whether they originate from an LLM update or a non-LLM preference transition, it should be interpreted separately from the number of LLM calls.

These results place ET-LCRL between two limiting preference strategies. Fixed preference is stable, but cannot explicitly react when the relative urgency of vehicle queues and pedestrian waiting changes. Rule-only preference is responsive but changes its objective weighting more frequently according to predefined rules. ET-LCRL instead uses the conflict-aware trigger to determine when contextual high-level guidance is warranted, while hysteresis and preference caching prevent the LLM from becoming a continuous low-level controller. Overall, the diagnostic evidence is consistent with the intended role of the LLM as a selective preference adapter: high-level reasoning is invoked only a small number of times during an episode, while the conditional Q-network remains responsible for routine signal execution under the fixed phase sequence and clearance constraints.

5.4 Discussion and Summary

To summarize the trade-offs across controllers, Fig. 5 compares five principal performance dimensions after min-max normalization. Episode reward and throughput are normalized, with higher values indicating better performance, while average delay, maximum pedestrian waiting, and total stops are reversed so that a higher normalized score consistently indicates better performance.

The figure is intended to visualize the shape of each controller's trade-off profile rather than to define a new scalar ranking.

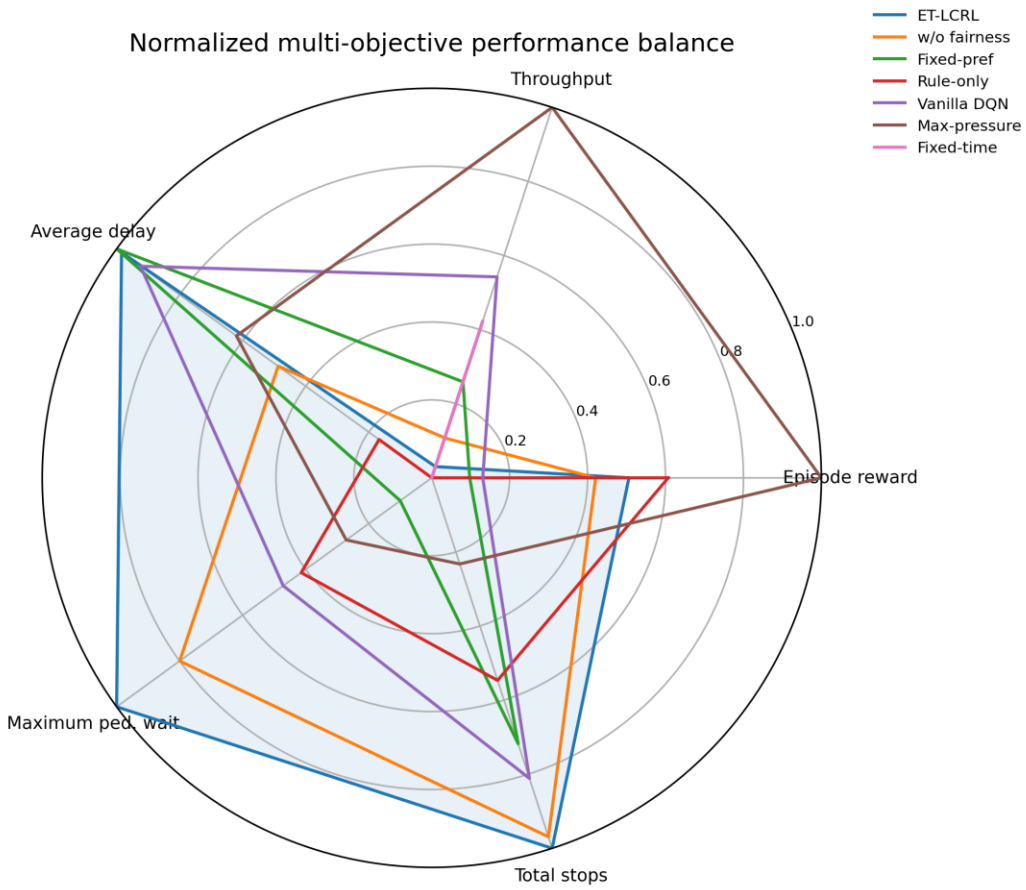

Fig. 5. Normalized Multi-Objective Performance Balance Across Controllers

The normalized profiles confirm that no single controller dominates every dimension. Max-pressure is strongest in reward and throughput, but its advantage is accompanied by higher vehicle delay, substantially more total stops, and longer maximum pedestrian waiting than ET-LCRL. Fixed preference achieves the best vehicle-delay result and slightly lower average stops, but its static weighting produces poorer extreme pedestrian service and more total stopping events. Rule-only preference obtains a relatively high reward but performs poorly on delay and stop-related stability, illustrating that frequent rule-based objective adjustment does not automatically yield a balanced traffic outcome.

The ablation comparisons further support the proposed design. Relative to ET-LCRL w/o fairness, the complete framework achieves lower average delay, lower maximum pedestrian waiting, fewer total stops, and higher episode reward, while throughput decreases by only approximately 2.3%. Relative to Fixed preference, ET-LCRL incurs only a 0.164 s increase in average vehicle delay but substantially improves maximum pedestrian waiting and system-level stopping performance. These results suggest that adaptive preference conditioning contributes to a more favorable allocation of performance across competing objectives rather than merely shifting the optimum toward one fixed metric.

The results should therefore be interpreted as evidence of a balanced operating point rather than universal dominance. ET-LCRL intentionally accepts a moderate throughput penalty compared with strongly vehicle-oriented controllers, but this cost is accompanied by near-best delay, the lowest extreme pedestrian waiting, and the lowest total stops. In a mixed vehicle-pedestrian intersection, such a trade-off is practically meaningful because a control policy that maximizes discharge while producing unstable progression or prolonged pedestrian deprivation may be undesirable despite strong performance on a single efficiency measure.

From a methodological perspective, the experiments also clarify why the LLM is placed above rather than inside the low-level signal-control loop. The proposed architecture uses the LLM only to adapt the relative importance of interpretable objectives under selected conflict states, while the reinforcement-learning controller executes constrained green-duration decisions. This separation preserves fast control execution and engineering constraints while allowing the optimization priority to change when traffic conditions require it.

6. Conclusions

This study proposed an Event-Triggered Large Language Model-Guided Conditional Reinforcement Learning (ET-LCRL) framework for pedestrian-aware traffic signal control. The framework separates high-level multi-objective preference adaptation from low-level signal timing. A conflict-aware event trigger activates the LLM only under critical pedestrian-vehicle service conditions, and the LLM generates a normalized preference vector rather than direct signal actions. A preference-conditioned dueling Double DQN then selects green duration under a fixed cyclic phase sequence and all-red clearance constraints. In this way, the framework combines adaptive high-level reasoning with stable and practically constrained low-level control.

Testing on an unseen scenario demonstrates a favorable trade-off among vehicle efficiency, pedestrian fairness, and operational stability. ET-LCRL achieves the lowest maximum pedestrian waiting time of 32.2 s and the lowest total number of stops of 137.0, while maintaining the second-lowest average vehicle delay of 7.468 s. Its throughput of 317.7 vehicles is lower than that of strongly vehicle-oriented strategies such as Max-pressure, indicating that the proposed method does not maximize discharge. However, the reduced throughput is accompanied by substantially better extreme pedestrian service, fewer stopping events, and lower vehicle delay than Max-pressure. The ablation comparisons further indicate that the complete fairness-aware and adaptive-preference framework yields a more balanced outcome than its reduced variants.

The mechanism analysis also supports the intended role of the LLM. ET-LCRL records the lowest mean conflict index (0.2031) and invokes the LLM only 3.3 times per test episode on average, while maintaining a lower preference-switching frequency than the purely rule-based preference controller. Rather than replacing the reinforcement-learning controller, the LLM therefore acts as a selective high-level preference adapter when conflict-related conditions justify a change in optimization priority. This event-triggered organization reduces the need for continuous interaction with the language model and maintains an interpretable separation between preference reasoning and signal execution.

Several limitations remain. The current evaluation is based on a limited set of microscopic simulation scenarios and has not yet been extended to large-scale coordinated multi-intersection control or real-world deployment. The reported comparisons are primarily descriptive and should be complemented in future work by broader scenario coverage and formal statistical testing of run-level differences. Practical deployment also requires further investigation of LLM inference latency,

communication reliability, model consistency, and implementation cost. Future research will therefore examine more diverse traffic-demand patterns and network configurations, investigate coordinated multi-intersection control, and explore lightweight or locally deployed language models for more reliable real-time applications.

Author Contributions

Conceptualization, Yunpeng Ma and Xinwei Zhang; methodology, Xinwei Zhang.; software, Xinwei Zhang; validation, Yunpeng Ma and Xinwei Zhang.; formal analysis, Xinwei Zhang.; investigation, Xinwei Zhang.; resources, Yunpeng Ma.; data curation, Xinwei Zhang.; writing—original draft preparation, Yunpeng Ma and Xinwei Zhang.; writing—review and editing, Yunpeng Ma, Xinwei Zhang, and Ferenc Mészáros; visualization, Xinwei Zhang.; supervision, Ferenc Mészáros.; project administration, Yunpeng Ma and Ferenc Mészáros. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Data Availability Statement

The data used in this study were collected through field observations of vehicle and pedestrian movements at selected signalized intersections in Budapest. The raw observation data is not publicly available due to privacy considerations and restrictions related to field data collection. However, the processed and aggregated data supporting the findings of this study may be made available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research was not funded by any grant.

References

[1] Li, B. (2013). A model of pedestrians' intended waiting times for street crossings at signalized intersections. Transportation Research Part B: Methodological, 51, 17-28. https://doi.org/10.1016/j.trb.2013.02.002

[2] Vedagiri, P., & Kadali, B. R. (2016). Evaluation of pedestrian-vehicle conflict severity at unprotected midblock crosswalks in India. Transportation Research Record, 2581(1), 48-56. https://doi.org/10.3141/2581-06

[3] Yazdani, M., Sarvi, M., Asadi Bagloee, S., Nassir, N., Price, J., & Parineh, H. (2023). Intelligent vehicle pedestrian light (IVPL): A deep reinforcement learning approach for traffic signal control. Transportation Research Part C: Emerging Technologies, 148, 103991. https://doi.org/10.1016/j.trc.2022.103991

[4] Wei, H., Zheng, G., Yao, H., & Li, Z. (2018). IntelliLight: A reinforcement learning approach for intelligent traffic light control. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2496-2505. https://doi.org/10.1145/3219819.3220096

[5] Wei, H., Chen, C., Zheng, G., Wu, K., Gayah, V. V., Xu, K., & Li, Z. (2019). PressLight: Learning max pressure control to coordinate traffic signals in arterial network. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 1290-1298. https://doi.org/10.1145/3292500.3330949

[6] Wei, H., Xu, N., Zhang, H., Zheng, G., Zang, X., Chen, C., Zhang, W., Zhu, Y., Xu, K., & Li, Z. (2019). CoLight: Learning network-level cooperation for traffic signal control. Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 1913-1922. https://doi.org/10.1145/3357384.3357902

[7] Liang, X., Du, X., Wang, G., & Han, Z. (2019). A deep reinforcement learning network for traffic light cycle control. IEEE Transactions on Vehicular Technology, 68(2), 1243-1253. https://doi.org/10.1109/TVT.2018.2890726

[8] Guo, M., Wang, P., Chan, C. Y., & Askary, S. (2019). A reinforcement learning approach for intelligent traffic signal control at urban intersections. Proceedings of the 2019 IEEE Intelligent Transportation Systems Conference, 4242-4247. https://doi.org/10.1109/ITSC.2019.8917268

[9] Xiao, H., Li, H., Wang, Z., Yang, Z., Qi, S., Zhang, J., ... & Yin, J. (2026). Intelligent traffic signal control based on reinforcement learning: a survey. Artificial Intelligence Review, 59(5), 128. https://doi.org/10.1007/s10462-026-11530-9

[10] Gong, Y., Abdel-Aty, M., Yuan, J., & Cai, Q. (2020). Multi-objective reinforcement learning approach for improving safety at intersections with adaptive traffic signal control. Accident Analysis & Prevention, 144, 105655. https://doi.org/10.1016/j.aap.2020.105655

[11] Elharoun, M., El-Badawy, S. M., Shwaly, E. A.-E., & Shahdah, U. E. (2025). Adaptive traffic signal control using deep reinforcement learning: A multi-objective approach for single and multi-intersection scenarios. IATSS Research, 49(4), 481-492. https://doi.org/10.1016/j.iatssr.2025.10.004

[12] Ye, Y., Ding, J., Wang, T., Zhou, J., Wei, X., & Chen, M. (2023). FairLight: Fairness-aware autonomous traffic signal control with hierarchical action space. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 42(8), 2434-2446. https://doi.org/10.1109/TCAD.2022.3226673

[13] Raeis, M., & Leon-Garcia, A. (2021). A deep reinforcement learning approach for fair traffic signal control. arXiv preprint arXiv:2107.10146. https://doi.org/10.1109/ITSC48978.2021.9564847

[14] Du, X., Li, Z., Long, C., Xing, Y., Yu, P. S., & Chen, H. (2024). FELight: Fairness-aware traffic signal control via sample-efficient reinforcement learning. IEEE Transactions on Knowledge and Data Engineering, 36(9), 4678-4692. https://doi.org/10.1109/TKDE.2024.3376745

[15] Lai, S., Xu, Z., Zhang, W., Liu, H., & Xiong, H. (2024). LLMLight: Large language models as traffic signal control agents. arXiv preprint arXiv:2312.16044. https://doi.org/10.1145/3690624.3709379

[16] Wang, M., Pang, A., Kan, Y., Pun, M.-O., Chen, C. S., & Huang, B. (2024). LLM-Assisted Light: Leveraging large language model capabilities for human-mimetic traffic signal control in complex urban environments. arXiv preprint arXiv:2403.08337.

[17] Wu, Y., Luo, C., Wang, J., Wang, Y., Jiang, Y., & Yao, Z. (2026). A survey of large language models for urban traffic signal control. Expert Systems with Applications, 330, 132998. https://doi.org/10.1016/j.eswa.2026.132998

[18] Fu, X., Yu, H., Jiang, H., Li, M., Cui, Z., & Ren, Y. (2026). LLM-enabled universal traffic signal control across different intersections and traffic flows. *Knowledge-Based Systems*, 115663. https://doi.org/10.1016/j.knosys.2026.115663

[19] Jovanović, A., Uzelac, A., Teodorović, D., & Kukić, K. (2025). Traffic signal control at displaced left-turn intersections using an LLM-based actor-critic algorithm. *Expert Systems with Applications*, 130555. https://doi.org/10.1016/j.eswa.2025.130555

[20] Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT Press. ISBN: 9780262039246.

[21] Puterman, M. L. (1994). Markov decision processes: Discrete stochastic dynamic programming. Wiley. https://doi.org/10.1002/9780470316887

[22] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529-533. https://doi.org/10.1038/nature14236

[23] Wang, Z., Schaul, T., Hessel, M., Van Hasselt, H., Lanctot, M., & De Freitas, N. (2016). Dueling network architectures for deep reinforcement learning. Proceedings of the 33rd International Conference on Machine Learning, 48, 1995-2003. https://doi.org/10.48550/arXiv.1511.06581

[24] Van Hasselt, H., Guez, A., & Silver, D. (2016). Deep reinforcement learning with double Q-learning. Proceedings of the AAAI Conference on Artificial Intelligence, 30(1), 2094-2100. https://doi.org/10.1609/aaai.v30i1.10295

[25] Schaul, T., Quan, J., Antonoglou, I., & Silver, D. (2016). Prioritized experience replay. Proceedings of the 4th International Conference on Learning Representations. https://doi.org/10.48550/arXiv.1511.05952