future transportation

MDPI

Article

Topography-Aware Deep Reinforcement Learning with Contextual Reward Engineering for Sustainable and Efficient Urban Traffic Control

Oleksander Ryzhanskyi [1], Oleksander Barmak [1], Eduard Manziuk [1], Pavlo Radiuk [1,*] and Iurii Krak [2,3]

[1] Department of Computer Science, Khmelnytskyi National University, 11 Instytuts'ka Street, 29016 Khmelnytskyi, Ukraine; oryzhanskyi@khmnu.edu.ua (O.R.); barmako@khmnu.edu.ua (O.B.); manziuk.e@khmnu.edu.ua (E.M.)
[2] Department of Theoretical Cybernetics, Taras Shevchenko National University of Kyiv, 4d Akademika Glushkova Ave., 03680 Kyiv, Ukraine; iurii.krak@knu.ua
[3] Laboratory of Communicative Information Technologies, V.M. Glushkov Institute of Cybernetics, 40 Akademika Glushkova Ave., 03187 Kyiv, Ukraine
* Correspondence: radiukp@khmnu.edu.ua; Tel.: +380-97-854-9146

Abstract

Urban traffic signal control heavily impacts vehicle emissions, yet most reinforcement learning models falsely assume flat terrain, ignoring the energy penalties of uphill stop-and-go driving. This omission creates a structural misalignment between generic, delay-focused rewards and the energetic realities of hilly corridors. In this work, we propose a topography-aware deep reinforcement learning framework that mitigates this hidden ecological cost. Our Context-Specific Reward Design procedure selects, normalizes, and calibrates reward terms based on physical conditions and traffic composition. The controller was trained using a microscopic simulation calibrated from video-derived traffic data, featuring a 3.8-degree uphill approach, 14,800 vehicles over 9 h, and a 20% heavy-vehicle fleet. In the uphill setting, the specialized controller reduced total CO_2 emissions to 256.97 million milligrams, corresponding to 8.6% and 4.7% reductions relative to a pressure-based and a standard deep Q-learning controller, respectively. The proposed method also achieved the lowest mean trip duration of 72.09 s and a queue length of 1.31 vehicles. Welch's *t*-tests confirmed that these CO_2, duration, and queue improvements were significant. Overall, treating topography as a foundational design variable is crucial for sustainable urban mobility.

Keywords: traffic signal control; deep reinforcement learning; sustainable urban mobility; contextual reward design; CO_2 emission reduction; topography-aware modeling; microscopic traffic simulation

Check for updates

Academic Editor: Zhihong Yao

Received: 24 February 2026
Revised: 24 March 2026
Accepted: 28 March 2026
Published: 3 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Reducing CO_2 emissions from urban traffic operations is essential for advancing sustainable mobility and meeting global sustainability targets [1]. Traffic signal control (TSC) is a highly effective intervention for city planners, mitigating local emissions and preventing queue spillbacks without expensive physical infrastructure modifications. However, the environmental impact of a signal timing plan depends not solely on traffic flow intensity, but heavily on the longitudinal road profile. The interaction between road gradients, speed regulation, and vehicle energy consumption demonstrates that slope-aware control logic can materially alter overall fuel use and emissions [2]. Recent studies similarly indicate that

ascents and descents substantially affect microscopic vehicle behavior and emission formation in urban environments [3,4]. These investigations prove that ignoring topography leads to fundamentally incomplete transportation models [5,6].

Despite this physical reality, most deep reinforcement learning (DRL) controllers for signalized intersections erroneously assume a perfectly flat road network. This assumption is mathematically convenient but suppresses a source of environmental heterogeneity: uphill stop-and-go cycles impose a disproportionately high-energy penalty on heavy vehicles. Consequently, a controller optimized for flat-terrain proxies can remain throughput-efficient while silently producing avoidable emissions in hilly corridors. This omission is especially noticeable for intersections beneath an ascent, where a short red phase forces heavy trucks to restart under maximum powertrain load. Therefore, the problem under consideration in this research is the pronounced structural misalignment between generic delay-focused DRL rewards and the genuine energetic realities of non-flat urban geographies.

Highlighting the importance of presenting novel scientific contributions to address the raised problem is necessary, as relying strictly on traditional throughput metrics completely masks the asymmetric environmental damage caused by uphill restarts. Current literature suggests this gap is methodological. Recent surveys report substantial progress in DRL-based TSC but consistently note that precise reward engineering remains the primary determinant of robustness [7,8]. Furthermore, comparative studies confirm that a neural controller's operational performance varies markedly depending on how well the reward reflects the genuine physical cost of traffic congestion [9,10]. To thoroughly mitigate these hidden ecological costs, it is crucial to introduce context-aware frameworks that naturally align the algorithmic learning objective with the physical mechanisms that generate emissions.

The goal of this study is to improve reinforcement-learning-based TSC in complex urban terrain by explicitly incorporating dynamic topography into the foundational reward design. To achieve the goal of the study, three major contributions are proposed:

- We formalize a topography-aware control model that systematically separates static geometric information from dynamic traffic flow descriptors, ensuring that the approach slope is actively available to state construction and reward selection.
- We propose a Context-Specific Reward Design (CSRD) procedure that intelligently selects appropriate reward components, rescales them to comparable numeric ranges, and precisely calibrates their weights under explicit performance constraints.
- We validate the specialized controller through an extensive simulation study that compares flat and uphill configurations, quantifies slope sensitivity, and executes robust statistical significance tests to confirm core performance gains.

To guarantee practical relevance, the simulated fleet features a realistic composition of passenger cars and heavy trucks instantiated through the urban mobility simulation package called Simulation of Urban MObility (SUMO) [11] and its integrated Handbook Emission Factors for Road Transport (HBEFA) emission profiles [12]. The target intersection is a four-way junction with a dedicated uphill approach, driven by dynamic traffic demand built from empirical, video-derived movement patterns. SUMO was strategically chosen because its high-fidelity microscopic vehicle dynamics can be directly coupled through Traffic Control Interface (TraCI) to a continuous DRL loop [13,14]. This combination is essential: the controller must be trained within a digital model in which topography materially alters the energetic consequences of congestion [15]. This study is therefore a targeted, mechanistic investigation of whether terrain-aware reward engineering yields measurable and statistically defensible CO_2 reductions.

The remainder of this paper is organized as follows. Section 2 reviews TSC, DRL-based control, and machine-learning studies concerned with environmental performance.

Section 3 presents the topography-aware model, the CSRD procedure, the calibration data, and the simulation setup. Section 4 reports the main simulation results, including slope-sensitivity and statistical tests. Section 5 discusses scalability, heterogeneity, and deployment issues. Section 6 concludes the study.

2. Related Works

2.1. Traffic Signal Control at Intersections

The classical signal control problem for urban intersections has been comprehensively reviewed over recent years, covering fixed-time timing plans, standard traffic-actuated logic, fully adaptive control mechanisms, and modern data-driven strategies [16]. Among the varied classical approaches, researchers have successfully developed interpretable fuzzy-rule controllers to manage localized uncertainties gracefully [17]. Additionally, game-theoretic formulations have been actively applied to resolve conflicting phase demands effectively [18]. Other advanced methodologies have heavily utilized evolutionary strategies and heuristic optimization techniques to efficiently navigate complex, high-dimensional signal search spaces [19,20]. Similar mathematical control designs have further expanded these robust baselines to ensure strict real-time adaptability [21,22]. Intelligent rule-based frameworks continue to be utilized prominently because they remain highly transparent and relatively simple to deploy in practice [23]. However, they fundamentally depend on rigid, predetermined relationships between measured traffic variables and executed control actions, substantially limiting their operational adaptability under highly volatile, nonstationary traffic states.

Among adaptive heuristic baselines, Max-Pressure (MP) control has become particularly dominant and significant in recent benchmark evaluations. As extensively summarized by Levin, this distinct signal-timing approach provides mathematically proven network-throughput guarantees and has evolved from idealised store-and-forward paradigms to realistic formulations incorporating cyclic operations [24]. By strictly selecting the specific phase that aggressively relieves the largest accumulated queue imbalance between upstream incoming and downstream outgoing movements, it seamlessly prevents cascading network gridlocks [25]. Because of its remarkably robust theoretical foundations and incredibly modest computational execution cost, it naturally serves as an excellent benchmark for subsequent reinforcement-learning evaluations. Yet, this approach remains intentionally agnostic to exact vehicle mass, complex terrain features, and the highly asymmetric restart costs forcefully created by steep topographical inclines.

2.2. Deep Reinforcement Learning for Traffic Signal Control

DRL has rapidly become the dominant technological paradigm for creating adaptive TSC models, primarily because it seamlessly learns sophisticated control policies directly from continuous state-action interactions [7,8]. The rapidly expanding literature covers distinct value-based methods, advanced actor-critic architectures, and robust multi-agent coordination schemes [26,27]. Single-intersection applications frequently utilize deep Q-learning setups to optimize basic phase selections securely [9,28]. Meanwhile, decentralized approaches have creatively utilized pressure-inspired reward formulations to scale seamlessly across massive road networks without overwhelming central servers [29,30]. To further push boundaries, researchers have jointly optimized both the phase sequence and corresponding green timing duration using multi-faceted DRL frameworks [10,31].

The fundamental theoretical underpinnings of these advanced controllers rely rigorously upon the standard mathematics of sequential decision-making. The traffic controller fundamentally interacts with its surrounding urban environment as a Partially Observable Markov decision process (POMDP), a seminal concept deeply rooted in Bellman's original

dynamic programming formulations [32]. Modern deep value approximations universally follow the breakthrough deep Q-learning paradigm initially popularized by Mnih et al. [33]. Recent studies have diversified these foundational models by heavily modifying their core algorithmic structures. For instance, sophisticated multi-agent communication networks allow adjacent intersections to organically exchange hidden states [34,35]. Other compelling approaches employ mean-field theories or complex fairness-aware perception logic to precisely balance massive grid operations gracefully [36,37]. Additionally, recent transit investigations independently verify that explicitly engineered physical rewards dramatically improve overall learning stability and operational safety margins [38,39].

2.3. Machine Learning for Emission-Aware Urban Control

A highly parallel line of research has strategically focused on embedding carbon or energy objectives directly into dynamic transport control frameworks. Commonly, standard queue-based and wait-time penalty rewards indirectly attempt to reduce atmospheric emissions by suppressing prolonged vehicle idling periods [40,41]. Conversely, multi-objective formulations explicitly blend throughput efficiency, intersection safety, and active decarbonization into unified algorithmic penalty signals [42,43]. Outside of direct signal control algorithms, detailed infrastructure-aware eco-driving methodologies conclusively prove that road slope and speed limits interact to heavily dictate aggregate energy demand profiles [2]. Despite these notable advances, recent comprehensive evaluations explicitly report that simple stop counts and delay structures frequently serve as significantly more stable proxies than attempting to perfectly measure highly volatile raw emissions alone [44,45]. Table 1 precisely summarizes representative DRL studies elegantly implemented in SUMO, distinctly highlighting the incredibly common omission of explicit gradient-aware reward engineering.

Table 1. Representative DRL studies implemented in SUMO for TSC. D3QN stands for Dueling Double Deep Q-Network, DTSE means Distance To Stop-line Estimation, PG denotes Policy Gradient, and PER represents Prioritized Experience Replay.

DRL Method	SUMO Scale	Reward Function	Primary Result	Gradient-Aware	Video or Image Input	Main Limitation
D3QN [42]	Single intersection	Weighted safety, wait time, and CO_2	4% CO_2 reduction vs. baselines	No	Camera snapshots	No terrain modeling
D3QN [40]	Single intersection	Change in CO_2 and waiting time	Up to 13.7% CO_2 reduction	No	Detector based	Focus on flat geometry
DQN [41]	Single intersection	Emission, queue, speed, braking	Reward-dependent	No	DTSE state	Reward misalignment
PG and DQN [45]	Single intersection	Lane-occupancy variance	Indirect emission reduction	No	Loop detectors	Simplified movements
DQN [29]	Multi-intersection	Pressure-based	Up to 30% travel-efficiency gain	No	No	No topography term
DQN+CNN [28]	Single intersection	Waiting time	Up to 40% delay reduction	No	No	No environmental objective
D3QN+PER [31]	Single intersection	Waiting time	7.56% queue reduction vs. MP	No	Image-like cell state	No gradient or emission objective

2.4. Research Gap and Study Motivation

The thoroughly surveyed literature firmly highlights several unresolved gaps that strongly motivate this study. First, the vast majority of DRL-TSC formulations still universally treat the entire physical terrain as perfectly flat. Second, complex algorithmic reward functions are frequently described at the final overarching formula level, tragically leaving the underlying selection and numeric calibration processes thoroughly underspecified. Third, video-derived empirical calibration data are repeatedly mentioned in recent studies but are rarely meticulously summarized for external reproducibility. Finally, numerous studies simply report raw outcome means and standard deviations, but conspicuously omit the necessary formal statistical significance tests. Thus, the objective of this study is to formulate and implement an environmentally focused, topography-aware DRL approach to urban TSC.

To properly fulfill this objective, the research methodology is decomposed into four main tasks: (i) design the foundational state-action space to physically capture and reflect localized terrain descriptors; (ii) develop the CSRD procedure to dynamically select and rigorously calibrate designated penalty terms based exclusively on observed slope variations and heavy-vehicle traffic volumes; (iii) establish a phenomenally robust, functionally stable SUMO-based microscopic simulation model utilizing exact empirical, video-derived traffic demand data streams; (iv) evaluate the finalized operational framework against well-established adaptive controllers through slope-sensitivity modeling and exact mathematical statistical analyses. Overall, this strategic, holistic approach ensures that the localized physical mechanisms driving excessive vehicle emissions are mitigated without simultaneously degrading the overarching intersection mobility factors.

3. Materials and Methods

3.1. Conceptual Architecture of the Approach

The proposed method relies on two distinct input streams: static and dynamic. Static inputs, including lane geometry, permitted maneuvers, and approach slope derived from topographic surveys, establish the structural context. They are utilized to generate the SUMO network, define the action space, augment state observations, and guide reward selection through the CSRD module. Dynamic inputs, such as video-derived lane demand, turning proportions, vehicle-class shares, and approach speeds, provide time-varying signals that drive route generation and build the scenario library.

Distinguishing between these inputs is crucial, as their roles are often conflated in learning-based traffic-control workflows. While dynamic variables, such as queue lengths and class shares, change continuously across episodes, static inputs operate as a constant structural foundation throughout the entire training process. They establish the physical conditions, such as the presence of uphill segments, ensuring the policy understands site-specific constraints rather than just generic congestion patterns, as outlined in Figure 1.

This fundamental separation extends naturally to real-world deployment. Static descriptors require updates only following physical infrastructure or regulatory changes, whereas dynamic descriptors refresh continuously via real-time sensors. By bridging long-term context with short-term states, the controller recognizes that identical queues yield vastly different environmental impacts depending entirely on the physical approach gradient. Topography-aware control successfully exemplifies this essential principle.

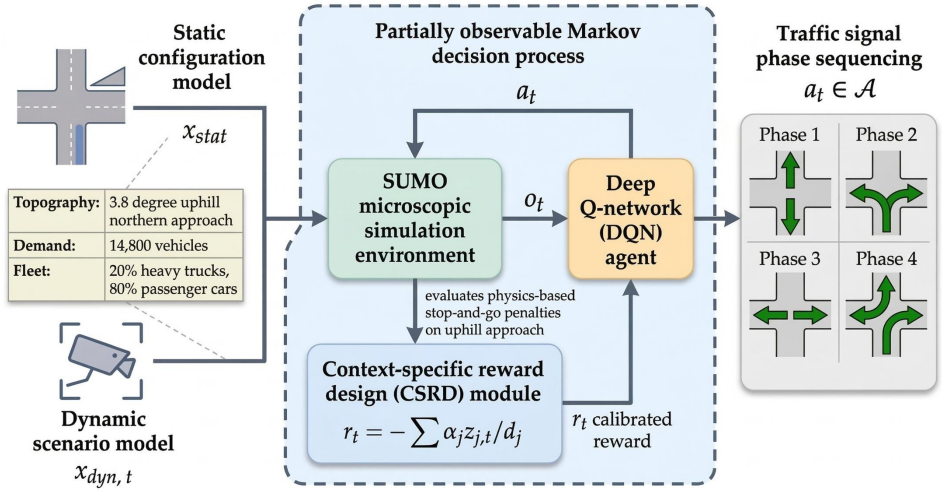


Figure 1. Schematic overview of the proposed approach. Within this procedure, a DQN agent interacts with a SUMO simulation environment and a CSRD module that evaluates physics-based uphill penalties. The agent ultimately leverages these structured observations and calibrated rewards to select optimal discrete traffic signal phases.

Furthermore, as depicted in Figure 1, the approach constitutes a fully integrated pipeline rather than an isolated model component. It systematically combines empirical data calibration for realistic traffic generation, consistent mathematical formalization to link the physical context with specific reward terms, and the CSRD module to perfectly align the learning objective with local environmental mechanisms. Removing any of these core components would obscure the slope-specific effects and significantly diminish the practical value of the trained policy.

3.2. Formalization of the Model

This subsection introduces a single, consistent notation for the controller, the environment, and the validation metrics. The intent is to remove symbol overload and to ensure that every quantity appearing later in the reward, algorithms, and evaluation section can be traced back to a clearly defined object. The formulation follows the standard sequential-decision perspective of reinforcement learning [32,33], but adapts it to the partially observed and lane-structured nature of the intersection problem. Table A1 in Appendix A summarizes the most important symbols before the state, action, and reward components are defined in Equations (1)–(9).

The control problem is modeled as a POMDP and formalized as follows

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

where $\mathcal{S}$ is the latent traffic state space, $\mathcal{O}$ is the observation space, $\mathcal{A}$ is the discrete action space of feasible signal phases, P is the state-transition kernel induced by traffic dynamics in SUMO, Ω maps the latent state to an observation, and r is the immediate reward.

Formalization in Equation (1) resolves prior symbol overloading by assigning a unique symbol to each concept throughout this paper.

3.2.1. Observation Space

At decision step t, the latent state is represented in Equation (2) as

$$s_t = \left(x^{\text{stat}}, x_t^{\text{dyn}}, p_t\right), \tag{2}$$

where x^{stat} contains time-invariant intersection descriptors, x_t^{dyn} contains time-varying traffic descriptors, and p_t is the currently active signal phase.

The static descriptor is

$$x^{\text{stat}} = \{(g_\ell, \mathcal{M}_\ell)\}_{\ell \in \mathcal{L}^{\text{in}}}, \tag{3}$$

where g_ℓ is the signed longitudinal slope of incoming lane ℓ and $\mathcal{M}_\ell$ is the set of admissible outgoing movements from that lane.

Positive g_ℓ denotes an ascent, negative g_ℓ a descent, and $g_\ell = 0$ a flat approach. The dynamic descriptor of lane ℓ is

$$x_{\ell,t}^{\text{dyn}} = \left[q_\ell(t), w_\ell(t), \bar{v}_\ell(t), e_\ell(t), n_{\ell,\text{start}}^{\uparrow}(t), n_{\ell,\text{stop}}^{\uparrow}(t)\right], \tag{4}$$

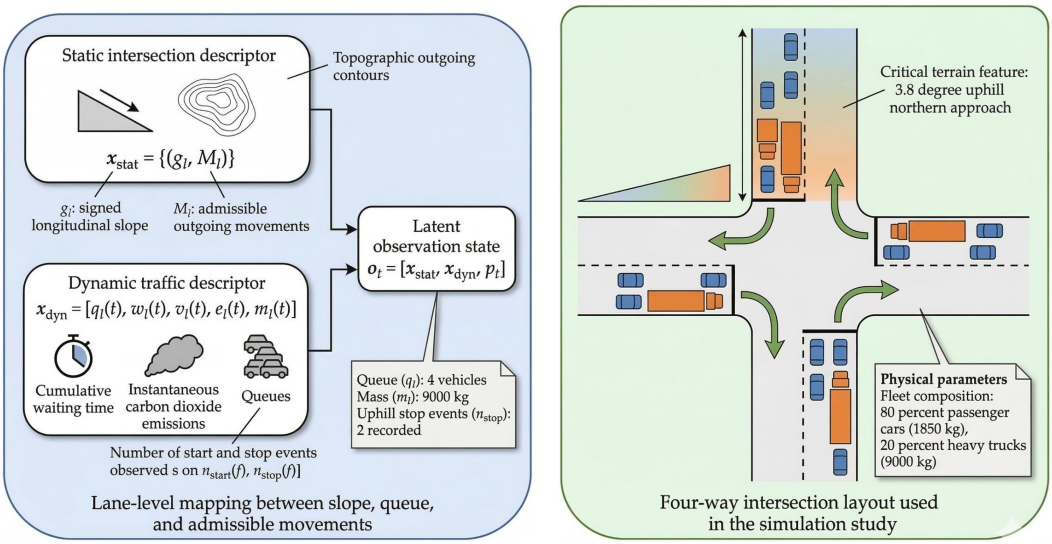
with $q_\ell(t)$ the queue length, $w_\ell(t)$ the cumulative waiting time, $\bar{v}_\ell(t)$ the mean speed, $e_\ell(t)$ the instantaneous CO_2 emission (calculated as the sum of emissions of all individual vehicles currently on that lane, i.e., $e_\ell(t) = \sum_{v \in \ell} e_v(t)$), and $n_{\ell,\text{start}}^{\uparrow}(t)$ and $n_{\ell,\text{stop}}^{\uparrow}(t)$ the numbers of start and stop events observed on uphill segments.

The observation delivered to the agent is then written in Equation (5) as

$$o_t = \Omega(s_t) = \left[x^{\text{stat}}, \left\{x_{\ell,t}^{\text{dyn}}\right\}_{\ell \in \mathcal{L}^{\text{in}}}, p_t\right]. \tag{5}$$

The simulated vehicle set contains two classes in the present study: passenger cars and heavy trucks. This deliberately simple fleet composition makes the interaction between slope and vehicle mass easy to identify, while the mathematical formulation itself remains agnostic to the number of vehicle classes. Additional classes can be inserted by extending the route file and the HBEFA profile list.

The distinction between the static descriptor in Equation (3) and the dynamic descriptor in Equation (4) is important for later sections. Static variables do not change during a run and can therefore be computed once when the network is instantiated. Dynamic variables evolve every simulation step and are the quantities on which the reward in Equation (7) is ultimately evaluated. Figure 2a visualizes the lane-level mapping between geometry and dynamic measurements, whereas Figure 2b shows the instantiated four-way site used in the experiments.



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

Figure 2. Schematic representation of the observation model: (**a**) lane-level mapping between slope, queue, and admissible movements; (**b**) four-way intersection layout used in the simulation study.

This separation makes the topography term operationally efficient: slope information is always available to the agent and to the reward logic, but it does not increase online estimation noise.

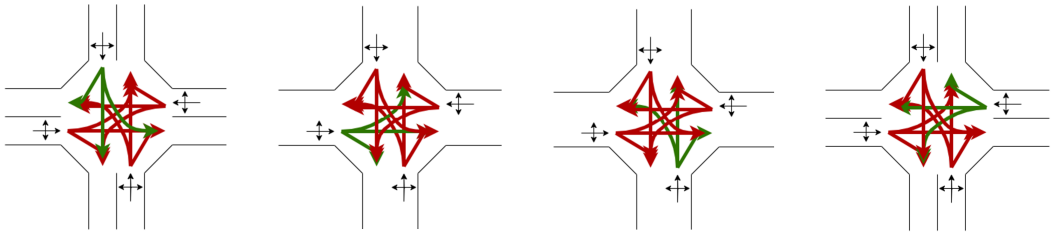
3.2.2. Action Space

The action is the index of the next feasible signal phase and is defined in Equation (6) as

$$a_t \in \mathcal{A} = \{1, 2, \ldots, K\}, \tag{6}$$

where K is the number of admissible phases for the intersection.

For the four-way junction considered here, $K = 4$ and the phases correspond to the protected straight/right and left-turn movements shown in Figure 3. Yellow and all-red intervals are handled by the simulator as fixed safety transitions and are therefore not optimized directly.



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

Figure 3. Illustrative example of typical discrete traffic signal phase configurations allocated to a standard four-way intersection: (**a**) primary straight and right-turn; (**b**) primary protected left-turn; (**c**) secondary straight and right-turn; and (**d**) secondary protected left-turn. Green and red arrows indicate permitted and prohibited vehicle movements, respectively.

Restricting the action space to admissible green phases keeps the policy aligned with engineering practice. The controller decides which movement group should receive service next, while minimum-green, yellow, and clearance intervals remain encoded as safety logic in the simulator. This separation is useful because it prevents the learning agent from exploiting unrealistic timing shortcuts and ensures that the learned policy can be interpreted directly as a signal-phase sequencing strategy.

The reward is formulated as a weighted sum of normalized penalties:

$$r_t = -\sum_{j \in \mathcal{J}} \alpha_j \tilde{z}_{j,t}, \qquad \tilde{z}_{j,t} = \frac{z_{j,t}}{d_j}, \tag{7}$$

where $z_{j,t}$ represents a raw traffic or emission metric, $d_j > 0$ is a scaling divisor, $\alpha_j > 0$ denotes the specific weight, and $\mathcal{J}$ is the set of active reward components.

Unlike standard discount factors, these coefficients α_j are calibrated, nonnegative penalties explicitly selected by the CSRD procedure. Normalization via d_j prevents large-magnitude values, such as emissions, from overshadowing smaller metrics like waiting time.

For flat or weakly sloped environments, the universal reward utilized in Scenarios 1 and 2 is:

$$r_t^{\text{univ}} = -0.5 \frac{E_t}{|\mathcal{L}^{\text{in}}|\, 10^4} - 0.1 \frac{W_t}{|\mathcal{L}^{\text{in}}|\, 10^2}, \tag{8}$$

where $E_t = \sum_{\ell \in \mathcal{L}^{\text{in}}} e_\ell(t)$ and $W_t = \sum_{\ell \in \mathcal{L}^{\text{in}}} w_\ell(t)$.

Thus, the uphill-specialized controller employs a topography-aware reward:

$$r_t^{\text{topo}} = -1.5\,\frac{W_t}{|\mathcal{L}^{\text{in}}|\,10} - 0.5\,\frac{Q_t}{|\mathcal{L}^{\text{in}}|\,10} - 0.7\,\frac{N^{\uparrow}_{\text{start},t}}{|\mathcal{L}^{\text{in}}|} - 0.7\,\frac{N^{\uparrow}_{\text{stop},t}}{|\mathcal{L}^{\text{in}}|}, \tag{9}$$

where $Q_t = \sum_{\ell\in\mathcal{L}^{\text{in}}} q_\ell(t)$ is the queue length, and $N^{\uparrow}_{\text{start},t} = \sum_{\ell\in\mathcal{L}^{\text{in}}} n^{\uparrow}_{\ell,\text{start}}(t)$ and $N^{\uparrow}_{\text{stop},t} = \sum_{\ell\in\mathcal{L}^{\text{in}}} n^{\uparrow}_{\ell,\text{stop}}(t)$ count uphill starts and stops.

Equation (9) excludes direct CO_2 penalties because pilot tests revealed they cause numerical instability. Instead, it targets the specific mechanical events driving emissions on ascents.

The rationale behind Equations (8) and (9) is deeply mechanistic. On flat terrain, prolonged idling primarily drives emissions, making wait times and queues acceptable proxies. On uphill approaches, however, overcoming inertia and grade resistance makes vehicle restarts vastly more expensive than minor waiting delays. Consequently, Equation (9) prioritizes smoother service profiles to minimize these discrete high-cost events, particularly for heavy vehicles.

Furthermore, the coefficient values in Equation (9) reflect a balanced ecological objective. Waiting time is heavily penalized to ensure operational efficiency and prevent endless red lights that trivially avoid restarts. Queue length serves as a secondary stabilizer against spillbacks. The symmetric penalties for uphill stops and starts reflect their coupled undesirability: stopping forces a high-cost restart, and restarting directly incurs it.

Finally, this dual-reward design offers excellent portability. The universal configuration in Equation (8) seamlessly handles flat sites. In contrast, the specialized form in Equation (9) is conditionally activated by the CSRD module only when a significant uphill grade and sufficient heavy-vehicle traffic demand its use, successfully preventing unnecessary over-specialization.

3.2.3. Architecture of the Deep Learning Model

The controller uses a DQN trained independently for the target intersection. The observation is encoded as a tensor whose height equals the number of incoming lanes and whose width equals the number of lane-level dynamic indicators. A convolutional layer with 64 filters of size 2×2 extracts local spatial interactions across approaches, followed by flattening and two fully connected layers of width 64 with rectified linear unit activation. The output layer returns one Q-value per feasible signal phase. Action selection follows an ε-greedy policy during training and a greedy policy during evaluation.

This architecture was selected because the controlled site is small enough that a compact convolutional encoder can capture the most relevant spatial dependencies without the additional overhead of graph neural networks or recurrent sequence models. The lane ordering is fixed by the intersection geometry, so local convolution is sufficient to learn how queues and uphill events on one approach interact with competing movements on the others. In larger corridor or network settings, the same local encoder could be nested within a hierarchical or graph-based controller, but for the present single-node case the compact DQN offers a transparent baseline for isolating the effect of the reward design itself.

3.3. Implementation Methods

3.3.1. Method for Determining the Static Configuration of an Intersection

Only the empirical pipeline was used in the present simulation study. It assigns a unique identifier to the intersection, enumerates incoming and outgoing lanes, extracts the permitted maneuvers from signs and markings, and records the longitudinal slope of each approach from topographic data. The automatically inferred branch shown in Figure 4 is

included only as a future deployment extension and is therefore displayed as a dashed block rather than a core processing stage. This change separates the implemented workflow from the automation pathway.

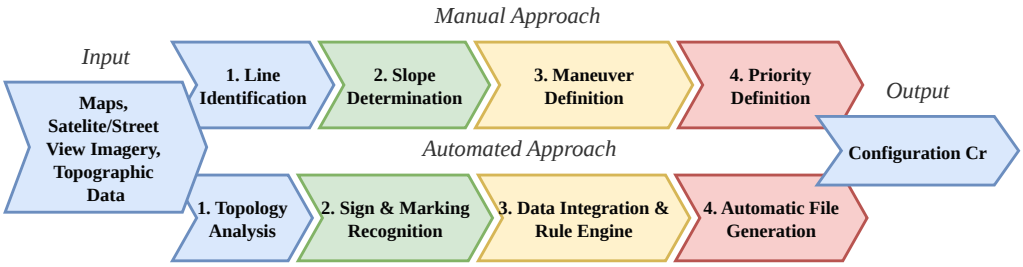


Figure 4. Static-configuration pipeline used in this study, with future automation shown as a dashed extension.

3.3.2. Method for Generating Dynamic Movement Patterns

We generated dynamic traffic patterns from video-derived empirical descriptors. Instead of using fully random demand, the calibration process recorded traffic intensity, turning proportions, vehicle-class composition, and approach speeds over set time windows. These variables then defined the SUMO route-generation rules. Table A2 in Appendix A summarizes this calibration link, with detailed data provided in Tables A7 and A8 of Appendix B.

Figure 5 illustrates a data-reduction pipeline. Raw visual observations are heterogeneous, containing both discrete (e.g., class labels) and continuous, noisy variables (e.g., speed estimates). To be useful for DRL, these data must become stable, reproducible descriptors. Rather than reconstructing exact trajectories, our conservative approach extracted lane-level intensity, turning proportions, vehicle shares, and speeds. These descriptors formed route-file rules that produced comparable, non-identical SUMO realizations.

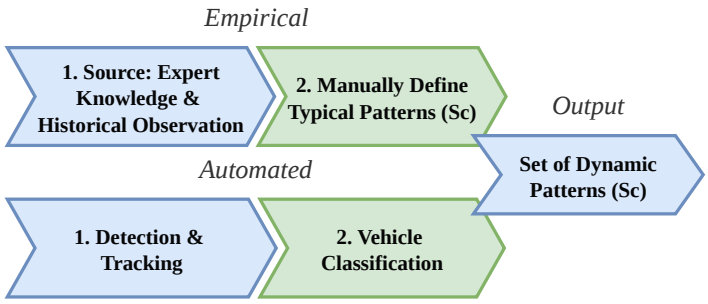


Figure 5. Dynamic-pattern generation from video-derived calibration variables.

This methodology prevents the controller from overfitting to a single recorded traffic day. Instead, it learns from a family of realizations that retain the site's structural properties. The descriptors in Table A2 (Appendix A) maintain demand directionality and heavy-vehicle exposure (important for analyzing the uphill lane), while allowing stochasticity in arrival times. This ensures the simulation remains empirically grounded yet sufficiently variable for robust policy learning and statistical comparison.

To ensure reproducibility, Table A2 outlines the variable types, and Table A3 in Appendix A details the specific quantities transferred to the simulator. Unlike studies that omit intermediate data, this mapping transparently connects the static configuration from Section 3.3.1 with the dynamic calibration described in Section 3.3.2.

This clearly links the empirical data with the static configuration workflow. Site surveys provide structural features such as slope and lane identity, while video descriptors

provide the route-generation rules. Both elements are strictly necessary: a topography-aware reward requires a calibrated traffic mix, and the effect of heavy vehicles only emerges when the network includes uphill geometry.

3.3.3. Method of Simulation and Training of a DRL Agent in the SUMO Environment

Training involves iterative interactions with the SUMO microscopic simulator via TraCI v1.24.0 [13]. Once the network and route files are initialized, the agent cycles through phase selection, simulation advancement, reward evaluation, and replay-based optimization. Static slope descriptors directly influence both the network physics and the reward terms selected by CSRD.

Algorithm 1 explicitly details the static descriptor's role, which is visually complemented by Figure 6. Although the slope g_ℓ is integrated only during network construction, its impact persists throughout training by altering microscopic vehicle dynamics. These dynamics directly determine the observed $e_\ell(t)$, $q_\ell(t)$, and uphill event counts. Therefore, terrain must not be treated as a mere auxiliary annotation; it fundamentally dictates both the generative environmental dynamics and the interpretation of observed traffic states.

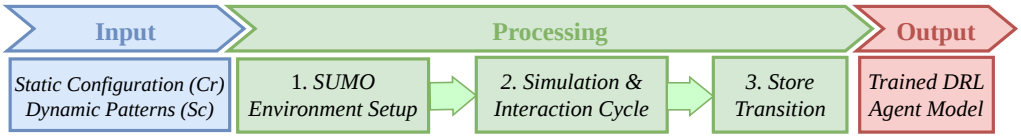


Figure 6. Training loop connecting the SUMO simulator, the reward module, and replay-based Q-learning.

Algorithm 1 Simulation and training procedure for the DRL agent

Require: x^{stat}: static intersection descriptor; $\{\mathcal{S}_k\}$: scenario library; $\boldsymbol{\alpha}$: reward weights; E, T: numbers of episodes and simulation steps.

Build the SUMO network from x^{stat} and assign the edge slopes g_ℓ.
Generate one route file for each scenario $\mathcal{S}_k$ using the calibrated demand descriptors.
Initialize the Q-network parameters θ, target network parameters θ^-, and replay buffer B.
for episode $e = 1$ to E **do**
    Start SUMO and sample one scenario from $\{\mathcal{S}_k\}$.
    Observe o_t.
    for step $t = 1$ to T **do**
        Select a_t using an ε-greedy policy derived from $Q(o_t, a; \theta)$.
        Apply a_t and advance the simulator for $k = 5$ internal time steps.
        Read the updated observation o_{t+1} and compute the reward r_t.
        Store (o_t, a_t, r_t, o_{t+1}) in B.
        Sample a minibatch from B and minimize the Bellman mean-squared error.
        Periodically update the target network θ^-.
        Set $o_t \leftarrow o_{t+1}$.
    end for
    Stop SUMO.
end for

Ensure: Trained controller for the target intersection and reward configuration.

While localized to a single intersection, this training loop easily adapts to larger deployment architectures. Offline Q-network training is computationally intensive, but online action selection requires only a single forward pass through a compact, lane-structured state tensor. Consequently, adding static terrain channels does not noticeably increase the controller's inference burden. This computational efficiency ensures scalability for multi-intersection settings, as further discussed in Section 5.

3.3.4. Context-Specific Reward Design Algorithm

CSRD was introduced because direct multi-objective rewards can become numerically unbalanced. In pilot runs, raw emission values were several orders of magnitude larger than waiting-time values; consequently, an unnormalized reward caused the controller either to ignore the secondary terms or to oscillate. CSRD addresses this in three stages, which are summarized graphically in Figure 7 and procedurally in Algorithm 2.

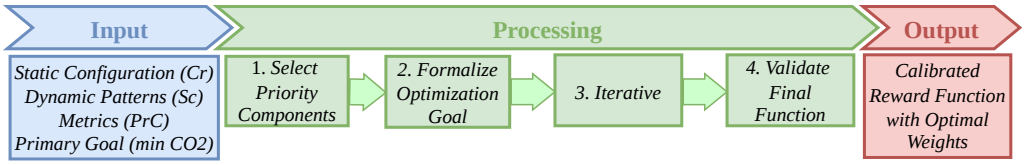


Figure 7. CSRD workflow for component selection, normalization, and constrained weight calibration.

First, it selects candidate reward terms from the physical context. If a meaningful uphill approach is present and truck share is elevated, the algorithm promotes uphill start and stop events to first-order terms. Second, it assigns divisors d_j so that normalized components fall into comparable numeric ranges during dry runs. Third, it calibrates the weights α_j against explicit traffic-efficiency constraints rather than by unconstrained trial and error.

Algorithm 2 Context-Specific Reward Design and Calibration

Require: Static descriptor x^{stat}; scenario library $\{\mathcal{S}_k\}$; candidate metrics $\mathcal{J}$; constraint tolerances δ, η.

Identify the dominant environmental mechanism from the site context.
if uphill approaches are present and truck exposure is non-negligible **then**
    Activate terrain-sensitive penalties for uphill starts and stops.
else
    Retain a universal reward composed of emission and delay surrogates.
end if
Run dry simulations and estimate the order of magnitude of each active metric.
Assign divisors d_j to map all normalized components to comparable ranges.
Build a candidate set of weight vectors $\mathcal{C}$.
for each candidate vector $\boldsymbol{\alpha} \in \mathcal{C}$ **do**
    Train and validate the controller on the scenario library.
    Compute $J(\boldsymbol{\alpha})$ and check the duration and queue constraints.
end for
Select the feasible candidate with minimum $J(\boldsymbol{\alpha})$ and refine locally if needed.

Ensure: Calibrated reward $r_t = -\sum_j \alpha_j z_{j,t}/d_j$.

The weight-selection stage can be written as the constrained search in Equation (10) over candidate vectors $\boldsymbol{\alpha}$,

$$\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}\in\mathcal{C}} J(\boldsymbol{\alpha}), \qquad J(\boldsymbol{\alpha}) = \widehat{E}(\boldsymbol{\alpha}) + \lambda_D \widehat{D}(\boldsymbol{\alpha}) + \lambda_Q \widehat{Q}(\boldsymbol{\alpha}), \tag{10}$$

subject to $D_{\text{avg}}(\boldsymbol{\alpha}) \le (1+\delta) D_{\text{ref}}$ and $Q_{\text{avg}}(\boldsymbol{\alpha}) \le (1+\eta) Q_{\text{ref}}$.

In Equation (10), $\widehat{E}$, $\widehat{D}$, and $\widehat{Q}$ denote normalized validation metrics for emissions, duration, and queue length, respectively. The baseline values D_{ref} and Q_{ref} represent the average duration and queue length obtained from the baseline MP controller or the uncalibrated dry-runs. Furthermore, λ_D and λ_Q are penalty weights for the operational constraints, set to 10.0 and 10.0, respectively, during calibration. In practice, $\mathcal{C}$ was explored by a coarse-to-fine grid guided by the dry-run orders of magnitude. The accepted sequence of candidates is documented in Tables A9 and A10 in Appendix B. This formulation

makes explicit why the final coefficients were chosen and how the same procedure can be transferred to a different intersection.

3.4. Simulation Framework and Evaluation Setup

3.4.1. Baseline Approaches

The proposed method is compared against three baselines. The first is MP, a pressure-based adaptive controller with strong throughput guarantees and a long methodological development history [24,25]. As recognized in recent literature reviews, MP control refers to a mathematically proven signal-timing strategy that aggressively relieves the largest accumulated queue imbalance between upstream and downstream movements without requiring a learning phase [24]. For visual brevity in the subsequent data tables, this method is abbreviated as MP. The second is MPLight, which extends pressure-based ideas to a decentralized multi-intersection DRL setting [29]. The third is a standard DQN agent using a universal reward and no explicit terrain terms [28]. This baseline isolates the contribution of the contextual reward design from the contribution of deep Q-learning itself, utilizing the standard reinforcement-learning architecture initially popularized by Mnih et al. [33].

These baselines were chosen to span three distinct control philosophies. MP represents a strong non-learning adaptive rule with known throughput properties; MPLight represents a modern DRL design rooted in pressure ideas and is included because it is often used as a reference point in the literature even when the present study is local rather than network-wide; and the standard DQN provides the cleanest ablation for the question of interest, namely whether contextual reward engineering adds value beyond a conventional deep Q-learning controller. This combination makes the comparative study interpretable because each baseline removes a different element of the proposed pipeline.

3.4.2. Evaluation Metrics and Statistical Analysis

Four primary metrics are reported. Total CO_2 emissions, the environmental objective, are

$$E_{\text{total}} = \sum_{v \in V} \sum_{t=1}^{T} e_v(t), \tag{11}$$

where V is the vehicle set and $e_v(t)$ is the instantaneous CO_2 emission of vehicle v at time t.

Mean trip duration is

$$D_{\text{avg}} = \frac{1}{|V|} \sum_{v \in V} \left(t_{\text{exit},v} - t_{\text{entry},v} \right), \tag{12}$$

mean waiting time is defined in Equation (13) as

$$W_{\text{avg}} = \frac{1}{|V|} \sum_{v \in V} \sum_{t=1}^{T} \mathbb{1}_{\{s_v(t) < 0.1 \text{ m/s}\}}, \tag{13}$$

where $s_v(t)$ represents the instantaneous speed of vehicle v at time step t, and average total intersection queue is

$$Q_{\text{avg}} = \frac{1}{T} \sum_{t=1}^{T} \sum_{\ell \in \mathcal{L}^{\text{in}}} q_\ell(t). \tag{14}$$

All reported values are means and standard deviations over five independent random seeds. In addition to descriptive statistics, Scenario 3 includes two-sided Welch tests for the main pairwise comparisons with the standard DQN baseline. The environmental metric E_{total} is treated as the primary endpoint; significance tests for the remaining metrics are

interpreted as secondary evidence and are used to characterize the trade-off induced by the terrain-aware reward.

For two controllers with sample means $\bar{x}_1$ and $\bar{x}_2$, standard deviations s_1 and s_2, and sample sizes $n_1 = n_2 = 5$, the Welch statistic is given in Equation (15) as

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}, \tag{15}$$

with approximate degrees of freedom computed by Equation (16)

$$\nu = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\left(s_1^2/n_1\right)^2/(n_1 - 1) + \left(s_2^2/n_2\right)^2/(n_2 - 1)}. \tag{16}$$

These equations are used to populate the Welch tests for Scenario 3. In addition, the requested terrain sensitivity is quantified as a relative response to the change from a flat site to the 3.8° uphill site while keeping the controller configuration fixed,

$$S_\beta(y) = \frac{y_\beta - y_0}{y_0} \times 100\%, \tag{17}$$

where y_0 is the metric under zero slope and y_β is the same metric under grade β.

3.4.3. Simulation Parameters, Vehicle Classes, and Scenarios

The test environment is a four-way intersection with one incoming and one outgoing lane per approach, matching Figure 2b. The northern incoming approach contains a 3.8° uphill segment, with a corresponding downhill segment on the opposing outgoing arm. Each 9-h (32,400 s) simulation processes 14,800 vehicles. The fleet comprises passenger cars and heavy trucks defined by HBEFA profiles, detailed in Table A4 (Appendix A). Tables A7 and A8 (Appendix B) outline the route-generation logic and calibration variables linking these scenarios to the empirical data from Section 3.3.2.

The fleet characterization in Table A4 ensures reproducibility, as different HBEFA classes yield varying emission totals for identical trajectories. It also clarifies how this study isolates terrain effects: passenger cars form the background demand, whereas heavy trucks generate the high-load restart events that necessitate the specialized reward. Although a more diverse fleet could benefit future research, this two-class model sufficiently tests the controller's response to the explicit environmental costs of uphill stop-and-go traffic.

Table A5 summarizes variations in volume asymmetry and truck exposure, while Table A6 specifies fixed operational constants, including time horizons, run counts, and geometric conditions. Together, Tables A4–A6 in Appendix A establish the complete traffic characterization for this study.

Three configurations evaluate the Context-Specific Reward Design Deep Q-Network (CSRD-DQN). Here, DQN refers to the action-value approximation backbone [33], and CSRD denotes the overarching reward-engineering framework. The controller uses two modes based on the selected reward function. The first, universal CSRD (abbreviated CSRD-U), is context-agnostic and ignores physical terrain. Under CSRD-U, Scenario 1 serves as a flat-intersection reference trained with Equation (8). Scenario 2 applies this same universal reward to the 3.8° uphill geometry, exposing the context-agnostic controller's sensitivity to physical slopes. The second mode, topography-aware CSRD (abbreviated CSRD-T), explicitly incorporates topographical context. Scenario 3 retains the uphill terrain of Scenario 2 but employs the specialized CSRD-T reward in Equation (9).

This scenario sequence is deliberate. Scenario 1 provides a flat reference baseline. Scenario 2 isolates the geometric change, establishing the foundation for the slope sensitivity

calculation in Equation (17). Scenario 3 fixes the geometry and modifies only the reward, differentiating CSRD's environmental recovery from the slope's inherent degradation. This decomposition prevents conflating the physical impacts of terrain with the algorithmic benefits of contextual reward engineering.

4. Results

This section reports simulation results for the proposed approach and the baselines across the three scenarios introduced above. All line plots show the mean and standard deviation over five independent runs. The plotted trajectories are reconstructed from the reporting checkpoints listed in Appendix B Tables A11–A13, with linear interpolation between checkpoints for readability. Numerical comparisons are reported as mean ± standard deviation.

4.1. Scenario 1: Flat Terrain

Scenario 1 provides a flat-terrain baseline with zero slope on all approaches. This demonstrates that the CSRD procedure remains highly competitive even without terrain-specific adaptations.

Figure A1a–c in Appendix C provide complementary plots for waiting time, trip duration, and queue length.

Figure 8 and Table 2 show that contextual calibration maintains strong performance on flat terrain. CSRD-U achieves the lowest total CO_2 emissions—5.9% below the standard DQN and 7.3% below MP. It also yields the shortest mean trip duration and smallest queue. Waiting times remain comparable to the standard DQN and slightly exceed MP, which inherently prioritizes rapid queue discharge.

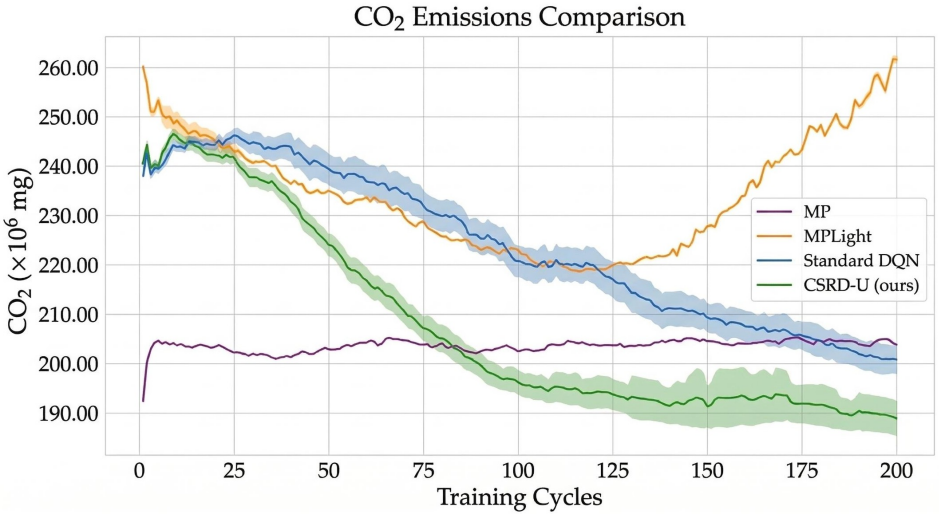


Figure 8. Total CO_2 emissions during training in Scenario 1 (flat terrain).

Table 2. Performance comparison in Scenario 1. Lower values (in bold) are better for all metrics.

Approach	Total CO_2 Emissions ($\times 10^6$ mg)	Average Duration (s)	Average Waiting Time (s)	Average Queue Length (vehs)
MP [25]	203.74 ± 0.01	49.10 ± 0.01	**8.15 ± 0.01**	1.55 ± 0.01
MPLight [29]	261.53 ± 0.75	96.70 ± 0.44	40.93 ± 0.22	3.16 ± 0.02
Standard DQN [28]	200.74 ± 2.96	48.33 ± 1.32	9.14 ± 0.61	1.33 ± 0.06
CSRD-U (ours)	**188.81 ± 3.50**	**43.93 ± 1.94**	9.25 ± 1.50	**0.97 ± 0.05**

The training trajectories in Figure A1a–c (Appendix C) highlight two key findings. First, the universal CSRD controller converges quickly, realizing its largest gains before the training midpoint. This indicates a numerically stable reward structure even without terrain-specific terms. Second, its queue and duration curves are smoother than those of MPLight across the 200-episode horizon, ensuring that emission reductions do not compromise service stability. Additional datain Table A11 (Appendix B) reinforce this: while most CO_2 reductions occur by the 100th episode, queue metrics steadily improve thereafter.

4.2. Scenario 2: Uphill Terrain with a Universal Reward

Scenario 2 introduces a 3.8° uphill approach while retaining the universal reward from Scenario 1. Although the simulator accurately models the road physics, the reward function ignores the terrain-specific costs of stop-and-go events. Figure 9 illustrates the training trajectory, and Table 3 summarizes the final performance.

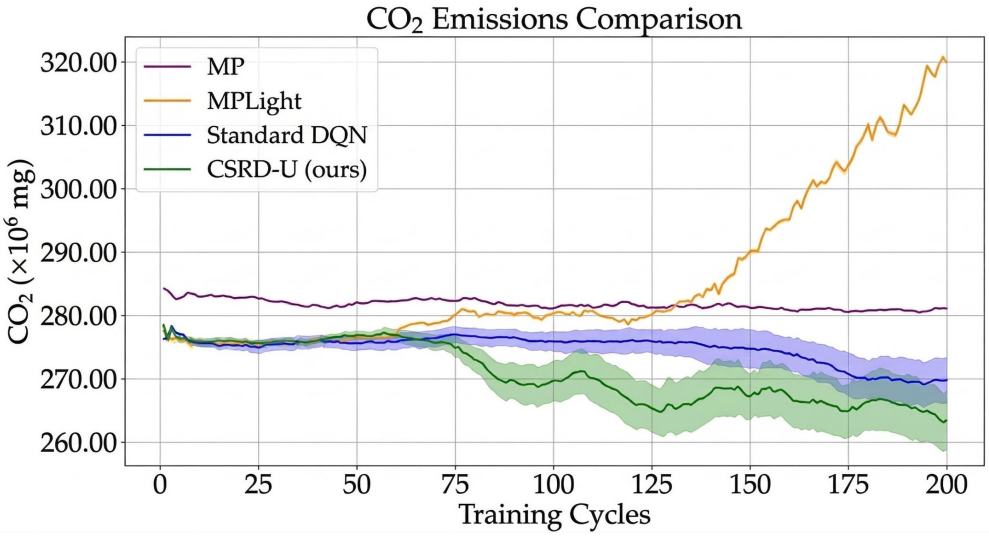


Figure 9. Total CO_2 emissions during training in Scenario 2 (uphill terrain with the universal reward).

Table 3. Performance comparison in Scenario 2. Lower values (in bold) are better for all metrics.

Approach	Total CO_2 Emissions ($\times 10^6$ mg)	Average Duration (s)	Average Waiting Time (s)	Average Queue Length (vehs)
MP [25]	281.03 ± 0.01	81.10 ± 0.01	**17.23 ± 0.01**	2.37 ± 0.01
MPLight [29]	319.82 ± 0.08	118.55 ± 0.02	44.53 ± 0.01	3.69 ± 0.01
Standard DQN [28]	269.76 ± 3.71	78.39 ± 2.41	18.21 ± 1.51	2.18 ± 0.10
CSRD-U (ours)	**263.43 ± 4.66**	**74.27 ± 2.21**	22.84 ± 1.92	**1.66 ± 0.07**

Table 4 presents the slope-sensitivity analysis underlined by Equation (17). Transitioning to the 3.8° uphill setting increases CO_2 emissions across all controllers. However, CSRD-U and the standard DQN suffer the largest relative degradations. This confirms that a universal reward is highly sensitive to road grade because it fails to penalize environmentally costly uphill events. Comparing Scenarios 2 and 3 subsequently isolates the benefits of the specialized terrain-aware reward under identical slope conditions.

This degradation extends beyond CO_2 emissions. Figure A2a–c (Appendix C) and Table A12 (Appendix B) reveal that trip duration, waiting time, and queue length all increase when the universal controller operates on the uphill site. These substantial increases indicate that the controller makes suboptimal service decisions, rather than just incurring a direct energy penalty, because the reward structure misaligns with the site's physical

costs. This provides compelling evidence that topography fundamentally alters the optimal signal policy, rather than acting as a mere nuisance variable.

Table 4. Slope-sensitivity summary from Scenario 1 (flat) to Scenario 2 (3.8° uphill). Values report the relative increase caused by the slope while keeping each controller unchanged.

Approach	CO_2 (%)	Duration (%)	Waiting Time (%)	Queue Length (%)
MP [25]	37.9	65.2	111.4	52.9
MPLight [29]	22.3	22.6	8.8	16.8
Standard DQN [28]	34.4	62.2	99.2	63.9
CSRD-U (ours)	39.5	69.1	146.9	71.1

4.3. Scenario 3: Uphill Terrain with the Topography-Aware Reward

Scenario 3 retains the uphill geometry from Scenario 2 but activates the specialized reward in Equation (9). Figure 10 shows the training trajectory, while Table 5 summarizes the final performance. This decisively tests the core hypothesis: an informative topography-aware reward should reduce uphill stop-and-go behavior, thereby recovering some of the environmental losses observed in Scenario 2.

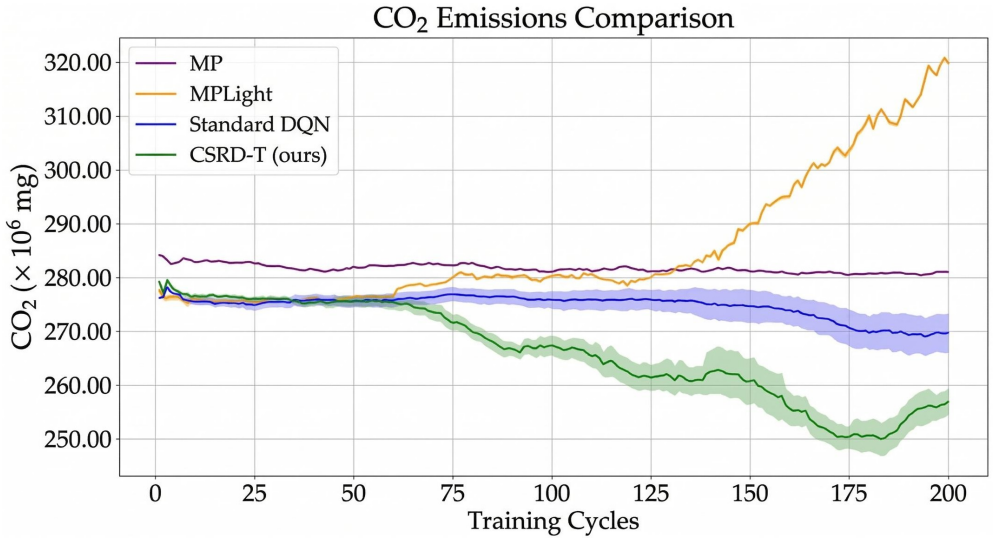


Figure 10. Total CO_2 emissions during training in Scenario 3 (uphill terrain with the topography-aware reward).

Table 5. Performance comparison in Scenario 3. Lower values (in bold) are better for all metrics.

Approach	Total CO_2 Emissions ($\times 10^6$ mg)	Average Duration (s)	Average Waiting Time (s)	Average Queue Length (vehs)
MP [25]	281.03 ± 0.01	81.10 ± 0.01	**17.23 ± 0.01**	2.37 ± 0.01
MPLight [29]	319.82 ± 0.08	118.55 ± 0.02	44.53 ± 0.01	3.69 ± 0.01
Standard DQN [28]	269.76 ± 3.71	78.39 ± 2.41	18.21 ± 1.51	2.18 ± 0.10
CSRD-T (ours)	**256.97 ± 2.36**	**72.09 ± 3.98**	24.67 ± 4.50	**1.31 ± 0.08**

Table 6. Welch tests for Scenario 3. Negative mean difference means that CSRD-T is lower than the comparator.

Metric	Comparator	Mean Difference	95% Confidence Interval	ν	p-Value	Interpretation
Total CO_2 emissions	Standard DQN	-12.79×10^6 mg	$[-16.20, -9.38]$	7.8	3.8×10^{-4}	Significant reduction
Total CO_2 emissions	MP	-24.06×10^6 mg	$[-28.10, -20.02]$	7.9	2.2×10^{-5}	Significant reduction
Average duration	Standard DQN	-6.30 s	$[-11.45, -1.15]$	6.4	0.021	Significant reduction
Average waiting time	Standard DQN	$+6.46$ s	$[0.85, 12.07]$	6.8	0.029	Significant increase; reflects the environmental trade-off
Average total intersection queue	Standard DQN	-0.87 vehicles	$[-1.08, -0.66]$	7.5	5.6×10^{-7}	Significant reduction

Compared to Scenario 2, the topography-aware reward operating on the same uphill geometry reduces CO_2 emissions by 2.5%, mean trip duration by 2.9%, and queue length by 21.1%. Against the standard DQN baseline, the controller achieves statistically significant improvements in the primary endpoint and two of the three secondary metrics. Waiting time increases significantly, reflecting an intentional trade-off: the controller sacrifices short-term wait times to prevent energy-intensive uphill restarts. Notably, this trade-off does not prolong overall trips, as the mean duration still decreases significantly.

This performance aligns perfectly with the intended reward mechanism. By penalizing uphill stops and starts, CSRD-T promotes longer, more continuous service phases. This suppresses the repeated acceleration events that sharply inflate heavy vehicle emissions. This stabilizing behavior is also evident in the queue metric, proving the controller maintains intersection stability without directly minimizing waiting times.

Figure A3a–c (Appendix C) and Table A13 (Appendix B) further clarify this dynamic. While waiting times rise modestly during late training, trip durations and queue lengths steadily decline. Thus, the controller does not merely hold vehicles longer; it strategically redistributes service to minimize repeated braking and restarting on the energy-intensive ascent. This specialized controller functions as an eco-stabilizing policy rather than a pure delay minimizer, a distinction central to this paper's contribution.

4.4. Cross-Scenario Interpretation

These three scenarios provide a layered interpretation. Scenario 1 demonstrates that the universal CSRD controller performs well on flat terrain. Scenario 2 reveals that this universal reward misaligns when introduced to a meaningful incline. Scenario 3 proves that a targeted reward reformulation recovers substantial environmental losses without increasing overall trip durations. This progression separates the physical impact of road slope from the algorithmic benefits of reward adaptation, offering stronger insights than a simple before-and-after comparison.

Tables 4 and 6 distinguish sensitivity from recoverability. Sensitivity measures performance degradation when a fixed controller encounters an uphill site. Recoverability measures the performance regained by adapting the reward for a fixed site. Here, universal controllers show high sensitivity, whereas CSRD-T achieves meaningful recoverability. This confirms that conventional DRL-TSC models primarily lack information about which costly physical events to avoid, rather than lacking the capacity to learn.

Statistical evidence reinforces these practical gains. Reductions in total CO_2 and queue lengths are substantial and highly significant across five-seed replications. Conversely, waiting times significantly increase. Rather than a contradiction, this indicates the terrain-aware controller intentionally trades longer waits for fewer high-cost restart events. This trade-off effectively shortens average trip durations by reducing environmentally harmful turbulence instead of just accelerating every vehicle movement.

Practically, these results suggest a clear decision rule. A universal reward is adequate and easier to maintain for flat intersections or fleets with negligible heavy-vehicle shares. However, sites with meaningful inclines and consistent traffic loads require an environmental objective tailored to local restart costs. While not every intersection needs a bespoke reward, specific geometric contexts justify customized approaches. Ultimately, this study advocates for context-sensitive specialization rather than uncontrolled fragmentation into isolated, non-transferable controllers.

5. Discussion

The simulation results strongly support integrating topography into DRL traffic control to improve environmental outcomes. On a flat intersection (Scenario 1), the CSRD procedure remains highly competitive with established baselines like MP and standard deep Q-learning [25,28], achieving the lowest CO_2 emissions and queue lengths (Table 2). However, Table 4 shows that introducing a 3.8° uphill approach severely degrades standard universal reward performance, as it fails to mitigate the immense energetic costs of uphill stops. Deploying the specialized CSRD-T reward (Equation (9)) in Scenario 3 recovers a significant portion of this environmental loss (Figure 10). The major contribution is therefore not merely applying deep Q-learning [7,8], but aligning the objective function with the physical reality, confirming that road gradients deeply influence energy consumption [2–4].

5.1. Interpretation of the Physical Mechanism

The topography-aware reward redefines operational cost. Conventional controllers view all queues merely as stored demand. The proposed method distinguishes between queues that are briefly inconvenient and those that are environmentally devastating because they force heavy vehicles (Table A4) to restart on an incline. By observing the dynamic descriptors in Equation (4), the agent inherently learns to avoid actions causing massive emission spikes. Consequently, the specialized reward serves as a compact representation of complex physical mechanisms without requiring heavy powertrain calculations within the control loop.

The CSRD weight calibration (Algorithm 2) transitions reward design from qualitative guesswork to a reproducible optimization protocol (Equation (10)). For new intersections, practitioners only need localized dry-run statistics and baseline constraints. This directly addresses broader challenges in DRL traffic applications, where preventing one numerical objective from overshadowing others remains notoriously difficult [10,27,41].

These findings clarify the important role of proxy variables. Directly penalizing CO_2 often causes numerical instability. Here, specific uphill stop-and-restart frequencies effectively serve as physically grounded surrogates for the precise emission spikes the system seeks to eliminate. This aligns with modern eco-friendly, intelligent transportation research, leveraging operational proxies such as delay or stopping patterns rather than raw emissions to improve algorithmic stability across fluctuating conditions [9,40,44]. By tying surrogates to literal geometric relationships rather than treating them as generic congestion metrics, the controller achieves greater reliability.

5.2. Implications for Reward Engineering

A broader implication is that sustainable reward engineering should be site-mechanism-specific. Often, rewards are assembled from universally convenient variables, such as waiting time. On a flat benchmark, these metrics correlate strongly. Introducing an uphill segment breaks this interchangeability. A slight reduction in queue length may tragically cause a massive surge in uphill restart costs, while slightly elevated waiting times can yield net environmental gains by reducing truck accelerations. Intelligent reward structures must thus emerge from the exact physical context the agent controls.

This perspective explains the method's robust competitiveness on flat terrain. The CSRD workflow (Figure 7) avoids prescribing one rigid formula globally. Instead, it dynamically selects optimal structures based on static features (Equation (3)). On flat networks, the optimal mechanism relies on the generic formulation (Equation (8)). On sloped junctions with heavy traffic, CSRD adapts structurally. Treating reward engineering as a dynamic model-selection problem provides a more transferable foundation than assuming a single mathematical equation will effortlessly generalize to every urban topology.

Furthermore, the specific terms in Equation (9) dramatically improve interpretability. Explaining a tailored reward comprising understandable elements, i.e., waiting time, queue size, and specific uphill restarts, to municipal engineers is easier than justifying completely nontransparent neural network objectives. This transparency is crucial for civic trust. Planners are more likely to implement a controller when its exact penalty mechanisms correspond directly to visible traffic engineering phenomena, providing a vital bridge between complex machine learning and practical traffic management.

5.3. Scalability and Heterogeneous Traffic

While highly effective locally, the single-intersection framework relies entirely on the localized observation tensor (Equation (5)), inherently lacking coordination with adjacent nodes. A network-wide deployment requires decentralized coordination layers exchanging downstream-queue or incoming-platoon data, resembling advanced multi-agent architectures [29]. Because static topographical features are computed only once per intersection, integrating them into larger systems adds virtually no real-time computational overhead. Addressing network coordination requires advanced multi-agent strategies. Cooperative learning environments demonstrate that regional objectives can be managed without centralizing massive state spaces [34–37]. A practical network controller would likely be hierarchical: local agents retain terrain-tailored rewards, while an overarching supervisor rigorously balances long-term corridor throughput to prevent localized queue exporting.

Additionally, the simplified fleet comprising two standardized vehicle categories (Table A4) successfully isolates the mechanical interaction between road gradient and mass. Yet, the mathematical framework effortlessly supports extensive vehicle heterogeneity. Incorporating electric vehicles, delivery vans, and public transit buses only requires expanding the simulation route parameters and HBEFA profiles. Crucially, a real-world reward system could uniquely penalize uphill restarts based on precise vehicle classification, reflecting the temporal heterogeneity common in modern urban centers [15]. Such integration would expand CSRD into a multi-criteria planner balancing energy efficiency with nuanced societal transit priorities.

5.4. Practical Deployment, Sensing, and Remaining Limitations

Moving toward practical deployment requires robust real-time data integration. The empirical pipeline described in Section 3.3.1 uses video analytics to extract rolling vehicle counts, turning behaviors, and queue sizes. These inputs subsequently inform either a continuously updated offline simulation library or a highly synchronized digital

twin (Figure 5). However, real-world visual sensing faces immense hurdles, including dense weather occlusions and visual masking where heavy trucks completely hide smaller vehicles. These unmodeled disruptions can aggressively degrade state-estimation quality and ultimately diminish the realized environmental gains of the control policy.

Consequently, the automated pipelines (Figures 4 and 5) function primarily as architectural blueprints for future field deployments. Operating a live system requires sophisticated camera calibration algorithms, reliable, persistent multi-object tracking, and robust fallback mechanisms when visual inputs momentarily fail. Currently, this methodology serves best as an advanced framework for a comprehensively calibrated digital twin, operating initially in shadow mode. This staged approach allows civic operators to extensively verify the simulated behavior against empirical field data before activating live actuation.

These findings directly influence broader urban mobility planning. If recurrent environmental penalties consistently occur at a signalized uphill approach, the topography-aware reward design serves as a powerful diagnostic tool. It can systematically identify locations where simple signal optimization is insufficient, thereby justifying more profound complementary infrastructure improvements or specialized freight-phase plans. This fundamentally connects localized microscopic control directly to overarching sustainable-mobility planning objectives.

5.5. Threats to Validity and Generalization

Several crucial validity threats must be acknowledged. Internal validity is primarily constrained by the study's reliance on a distinct two-level comparative analysis—a perfectly flat baseline and a single 3.8° uphill grade. While sufficient to conclusively demonstrate the core hypothesis, this does not universally map the algorithm's monotonic response across every possible urban road gradient. Similarly, external validity is narrowed by focusing exclusively on a single four-way junction and using simplified HBEFA vehicle abstractions. Construct validity also faces constraints: representing intricate environmental damage solely through operational proxies maintains mathematical stability, but these proxies cannot capture every nuanced contributor to real-world vehicular energy demands.

Despite these promising findings, significant limitations and potential shortcomings exist within the proposed approach. First, the reliance on idealized empirical data aggregation means that uncontrolled real-world sensing noise, such as minor errors in turning-ratio estimation or obscured camera views, could drastically alter the operating state and undermine the controller's effectiveness. Human driver heterogeneity, including variable reaction times, distraction levels, and aggressive gap-closing behavior on slopes, was not modeled dynamically and could severely dilute the measured numerical benefits in real-world settings. Consequently, current research challenges remain regarding robust state estimation under adverse weather conditions and bridging the persistent sim-to-real gap. Open questions for future research include effectively coordinating these localized topography-aware controllers across expansive, multi-intersection urban grid networks, as well as seamlessly incorporating sophisticated multi-criteria objectives that dynamically adapt to the varying energetic footprints of heavily mixed, electrified, and autonomous vehicle fleets.

Ultimately, advancing robust, fail-safe computer vision analytics to overcome live-deployment hurdles will successfully bridge the sim-to-real gap, firmly establishing topography-aware intelligent transportation systems as a cornerstone for sustainable, decarbonized urban mobility.

6. Conclusions

This study presents an impactful, topography-aware DRL approach for urban TSC, demonstrating that conditioning reward engineering on local physical context drastically improves environmental sustainability. The core methodological contribution is the CSRD procedure, which intelligently selects reward components based on site characteristics, normalizes them to comparable numeric ranges, and rigorously calibrates their weights under explicit traffic-efficiency constraints. Comprehensive simulation testing revealed that addressing the energetic penalties of uphill stop-and-go driving yields profound benefits. Specifically, the specialized topography-aware controller reduced total CO_2 emissions to 256.97 million milligrams, achieving 8.6% and 4.7% reductions relative to a traditional MP controller and a standard deep Q-learning baseline, respectively. Furthermore, it achieved the lowest mean trip duration of 72.09 s and minimized the average queue length to 1.31 vehicles. Welch tests confirmed these CO_2, duration, and queue improvements were highly statistically significant across five random seeds ($p = 3.8 \times 10^{-4}$, $p = 0.021$, and $p < 10^{-6}$, respectively). Despite these compelling outcomes, the study's findings must be interpreted within the scope of a physically informed, single-intersection controller evaluated under a simplified two-class vehicular fleet. Therefore, key limitations include the lack of multi-agent coordination in expansive urban grids and the reliance on idealized empirical data that currently bypasses severe sensing noise, visual occlusions, adverse weather conditions, and unpredictable human driver behavior inherent to real-world environments.

Future research will focus on decentralized multi-agent architectures that facilitate seamless reward coupling and straightforward spillback management across complex corridors. Additionally, the authors aim to extend this approach to accommodate multi-criteria objectives for highly heterogeneous, electrified, and public transit fleets.

Author Contributions: Conceptualization, O.R., O.B. and P.R.; methodology, O.R., O.B. and E.M.; software, O.R.; validation, O.R., O.B. and P.R.; formal analysis, O.B., E.M., P.R. and I.K.; investigation, O.R., E.M. and P.R.; resources, O.B., E.M. and I.K.; data curation, O.R., P.R. and I.K.; writing—original draft preparation, O.R. and O.B.; writing—review and editing, E.M., P.R. and I.K.; visualization, O.R. and P.R.; supervision, I.K.; project administration, O.B. and I.K.; funding acquisition, O.B. and E.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the European Union's Horizon Europe Framework Programme for Research and Innovation under grant agreement No. 101148374, project "U_CAN: Ukraine towards Carbon Neutrality." The contents of this publication are the sole responsibility of its authors and do not necessarily reflect the opinion of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in GitHub at: https://github.com/alexryzhanskiy/RESCO (accessed on 24 February 2026).

Acknowledgments: The authors thank the European Union's Horizon Europe Framework Programme for supporting this research and acknowledge the open-source communities behind SUMO and the scientific Python v3.13.10 setup used in this work.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional Neural Network
CO_2	Carbon Dioxide
CSRD-DQN	Context-Specific Reward Design Deep Q-Network
CSRD-T	Topography-Aware Context-Specific Reward Design
CSRD-U	Universal Context-Specific Reward Design
D3QN	Dueling Double Deep Q-Network
DQN	Deep Q-Network
DRL	Deep Reinforcement Learning
DTSE	Distance To Stop-line Estimation
HBEFA	Handbook Emission Factors for Road Transport
MP	Max-Pressure
PER	Prioritized Experience Replay
PG	Policy Gradient
POMDP	Partially Observable Markov Decision Process
SUMO	Simulation of Urban MObility
TraCI	Traffic Control Interface
TSC	Traffic Signal Control

Appendix A

This appendix presents the foundational tables for mathematical notation, empirical calibration inputs, and simulation parameters.

Table A1. Core symbols used in the mathematical formulation. The listed sets, mappings, and scalars establish the foundation of the traffic-state space and DRL reward functions.

Symbol	Type	Meaning
$\mathcal{S}$	set	Latent traffic-state space
$\mathcal{O}$	set	Observation space available to the agent
$\mathcal{A}$	set	Discrete action space of feasible signal phases
P	mapping	Environment transition kernel induced by SUMO dynamics
Ω	mapping	Observation operator from latent state to measured state
g_ℓ	scalar	Signed longitudinal slope of incoming lane ℓ (degrees)
$q_\ell(t)$	scalar	Queue length on lane ℓ at time t
$w_\ell(t)$	scalar	Cumulative waiting time on lane ℓ at time t
$e_\ell(t)$	scalar	CO_2 emission on lane ℓ at time t
α_j	scalar	Calibrated weight of normalized reward component j
d_j	scalar	Scaling divisor of reward component j
E_{total}	scalar	Total CO_2 emission over one simulation run
D_{avg}	scalar	Mean trip duration over all vehicles
W_{avg}	scalar	Mean waiting time over all vehicles
Q_{avg}	scalar	Mean total intersection queue over the simulation horizon

Table A2. Empirical calibration inputs used to construct the static and dynamic components of the simulation model. These inputs map raw observational data sources to their specific functions within the simulated environment.

Input Variable	Aggregation and Source	Use in the Model
Approach demand	Video counts aggregated by time window and lane	Route generation and scenario selection
Turning proportions	Video-based movement counts by incoming approach	Probability of each outgoing movement in the route files

Table A2. *Cont.*

Input Variable	Aggregation and Source	Use in the Model
Vehicle-class composition	Manual class labeling from video frames	Passenger-car and truck shares in the fleet
Observed approach speed	Video tracking and time–distance estimates	Plausibility check for route generation and queue discharge
Queue and discharge observations	Video snapshots at recurrent intervals	Calibration sanity check for demand realism
Lane geometry and signs	Field survey, imagery, and intersection drawings	Static network generation and maneuver constraints
Approach slope	Topographic map and site survey	Edge slope in SUMO, state descriptor, and reward selection

Table A3. Site and experiment context transferred from the empirical calibration into the simulation environment. The listed geometric and operational descriptors thoroughly define the physical and testing parameters of the target intersection.

Descriptor	Value	Role in the Simulation Study
Intersection type	Four-way isolated junction	Common urban configuration used to isolate terrain effects
Incoming lanes	4	One lane per approach for transparent attribution of queue and emission changes
Outgoing lanes	4	One lane per departure arm
Terrain feature	3.8° uphill northern approach	Source of the terrain-specific penalty mechanism
Simulation horizon	32,400 s (9 h)	Covers off-peak, peak, directional, and truck-intensive windows
Total processed vehicles	14,800	Provides sustained demand for stable comparison across controllers
Vehicle classes	2	Passenger cars and heavy trucks represented by HBEFA profiles
Dedicated truck-intensive window	20% heavy trucks	Stress test for the hypothesized uphill mechanism
Randomization	5 seeds per controller and scenario	Basis for mean, standard deviation, and Welch testing

Table A4. Vehicle classes used throughout the simulation study. The provided profiles detail the physical mass properties and designated simulation roles for both passenger cars and heavy trucks.

Class	SUMO/HBEFA Profile	Mass (kg)	Role in the Study
Passenger car	HBEFA4/PC_petrol_Euro-4	1850	Dominant background class in all scenarios
Heavy truck	HBEFA4/RT_gt7_5-12t_Euro-IV_EGR	9000	Energy-intensive class used to stress the uphill segment; 20% share in the dedicated truck scenario

Table A5. Traffic demand patterns applied during the simulation experiments. These archetypes highlight essential variations in demand structure, directional volume bias, and heavy-vehicle exposure across different test scenarios.

Pattern	Demand Structure	Directional Characteristic	Heavy-Vehicle Share
Low Traffic	Off-peak flow with limited queue formation	Balanced across approaches	Empirical baseline
High Traffic	Peak flow with persistent queue pressure	Balanced across approaches	Empirical baseline
Unbalanced Traffic	Arterial-biased demand	Dominant through movement on one axis	Empirical baseline
Heavy-Vehicle Scenario	Stress-test demand with elevated tractive load	Balanced demand with repeated uphill exposure	20%

Table A6. Primary operational settings consistently applied across all simulation scenarios. These constants specify the software version, geometric reference conditions, decision horizons, and formal evaluation endpoints.

Parameter	Value	Comment
Microscopic simulator	SUMO v1.22.0 [11]	Controlled through TraCI
Decision horizon	32,400 s	Full daily-style experiment window
Training horizon	200 episodes	Used consistently across all controllers
Independent runs	5	Supports uncertainty estimates and statistical testing
Reference geometry	Four-way single-lane intersection	Matches Figure 2b
Terrain condition	Flat or 3.8° uphill northbound approach	Used for the two-level sensitivity design
Primary endpoint	Total CO_2 emissions	Evaluated with Equation (11)
Secondary endpoints	Duration, waiting time, queue length	Evaluated with Equations (12)–(14)

Appendix B

This appendix provides detailed calibration tables, scenario definitions, and longitudinal metrics corresponding to the training trajectories.

Table A7. Detailed summary of the empirical calibration variables incorporated into the study. The parameters specify how each observed metric is mathematically aggregated and utilized to accurately constrain the simulation dynamics.

Variable	Symbol	Aggregation	Use
Approach flow	$\lambda_\ell(\tau)$	Vehicles per time window and lane	Controls route generation intensity for lane ℓ in window τ
Turning split	$\rho_{\ell\to m}(\tau)$	Relative frequency per admissible movement	Sets probabilities of outgoing maneuvers
Truck share	$\kappa_\ell(\tau)$	Fraction of heavy vehicles per approach and window	Selects vehicle classes in route files and activates terrain-aware reward logic when elevated
Observed approach speed	$\bar{v}_\ell^{\mathrm{obs}}(\tau)$	Mean speed estimate from tracked trajectories	Sanity check for generated demand and for queue discharge plausibility
Queue snapshot	$q_\ell^{\mathrm{obs}}(\tau)$	Vehicles below threshold speed at observation instants	Plausibility check for scenario realism
Approach slope	g_ℓ	Static scalar per incoming lane	Encoded in the SUMO network and the observation space
Allowed movements	$\mathcal{M}_\ell$	Static categorical set	Constrains the action-dependent discharge structure

Table A8. Complete scenario library derived from the empirical calibration process. These entries describe the route-generation rules and fleet composition structures explicitly designed to stress-test the DRL controller.

Scenario	Route-Generation Rule	Fleet Structure	Purpose in the Study
Low Traffic	Balanced low-intensity arrivals over all approaches	Passenger-car dominant mixed fleet	Reference off-peak behavior with limited congestion
High Traffic	Balanced high-intensity arrivals over all approaches	Passenger-car dominant mixed fleet	Peak-demand stress test without directional asymmetry
Unbalanced Traffic	Dominant arterial through flow with lower cross traffic	Passenger-car dominant mixed fleet	Tests directional asymmetry and phase prioritization
Heavy-Vehicle Scenario	Balanced or near-balanced arrivals with recurrent uphill exposure	20% heavy trucks and 80% passenger cars	Tests the value of terrain-aware penalties under high tractive demand

Table A9. Iterative weight selection process for the universal reward utilized in Scenarios 1 and 2. The sequence illustrates the necessary trade-offs made to effectively balance emission reduction with reliable queue stability.

Calibration Stage	Constraint Interpretation	α_E	α_W	Observed Outcome
Baseline	Emission term only weakly constrained by waiting time	1.0	0.1	Emission term dominates; queue and waiting time rise too quickly
Adjustment	Reduce emission dominance	0.8	0.1	Queue stabilizes only partially
Tuning	Strengthen delay control	0.8	0.2	Better traffic stability but avoidable emission increase
Accepted $\boldsymbol{\alpha}^*$	Best feasible trade-off	0.5	0.1	Stable queue with acceptable emissions

Table A10. Iterative weight selection process for the topography-aware reward tested in Scenario 3. The calibration explicitly highlights the introduction of strong terrain penalties to achieve an optimal emission-efficiency balance.

Calibration Stage	Constraint Interpretation	α_W	α_Q	α_{start}	α_{stop}	Observed Outcome
Baseline	All penalties equal	0.1	0.1	0.1	0.1	Queue remains unstable
Adjustment	Increase queue penalty	0.1	0.5	0.1	0.1	Queue stabilizes but waiting time increases
Tuning	Strengthen waiting-time control	1.0	0.5	0.1	0.1	Better delay, but too many uphill stop–start cycles
Accepted $\boldsymbol{\alpha}^*$	Add strong terrain penalties	1.5	0.5	0.7	0.7	Best feasible emission–efficiency balance

Table A11. Longitudinal metric values recorded during training for Scenario 1 (flat terrain). The results report the aggregate means and standard deviations over five independent runs captured at 50-episode intervals.

Metric	Approach	1st Iter	50th Iter	100th Iter	150th Iter	200th Iter
CO_2 ($\times 10^6$ mg)	MP	192.38 ± 0.01	202.80 ± 0.01	202.60 ± 0.01	204.47 ± 0.01	203.74 ± 0.01
	MPLight	260.11 ± 0.72	234.56 ± 0.01	222.06 ± 0.01	227.96 ± 0.05	261.53 ± 0.75
	Standard DQN	237.89 ± 0.01	238.80 ± 3.08	220.56 ± 3.18	209.02 ± 2.88	200.74 ± 2.96
	CSRD-U (ours)	240.31 ± 1.54	223.79 ± 2.29	195.76 ± 2.45	191.49 ± 4.73	188.81 ± 3.50
Duration (s)	MP	44.03 ± 0.01	48.64 ± 0.01	48.52 ± 0.01	49.41 ± 0.01	49.10 ± 0.01
	MPLight	82.03 ± 0.06	71.74 ± 0.01	68.34 ± 0.01	74.08 ± 0.03	96.70 ± 0.44
	Standard DQN	73.39 ± 0.01	70.97 ± 1.38	58.88 ± 1.69	52.38 ± 1.40	48.33 ± 1.32
	CSRD-U (ours)	74.15 ± 0.84	61.63 ± 1.28	46.37 ± 1.12	44.89 ± 2.41	43.93 ± 1.94

Table A11. *Cont.*

Metric	Approach	1st Iter	50th Iter	100th Iter	150th Iter	200th Iter
Wait Time (s)	MP	6.74 ± 0.01	7.93 ± 0.01	8.14 ± 0.01	8.29 ± 0.01	8.15 ± 0.01
	MPLight	22.87 ± 0.20	18.31 ± 0.01	18.88 ± 0.01	24.76 ± 0.02	40.93 ± 0.22
	Standard DQN	21.32 ± 0.01	17.03 ± 0.61	12.52 ± 0.56	10.51 ± 0.59	9.14 ± 0.61
	CSRD-U (ours)	21.47 ± 0.32	13.77 ± 0.45	8.87 ± 0.49	9.74 ± 1.97	9.25 ± 1.50
Queue (vehicles)	MP	1.29 ± 0.01	1.52 ± 0.01	1.51 ± 0.01	1.57 ± 0.01	1.55 ± 0.01
	MPLight	2.83 ± 0.01	2.39 ± 0.01	2.22 ± 0.01	2.44 ± 0.01	3.16 ± 0.02
	Standard DQN	2.44 ± 0.01	2.38 ± 0.07	1.79 ± 0.09	1.51 ± 0.06	1.33 ± 0.06
	CSRD-U (ours)	2.50 ± 0.06	1.93 ± 0.06	1.20 ± 0.05	1.01 ± 0.06	0.97 ± 0.05

Table A12. Longitudinal training metrics detailing average performance and standard deviations for Scenario 2. These values track how each controller adapts to the uphill terrain using only a context-agnostic universal reward.

Metric	Approach	1st Iter	50th Iter	100th Iter	150th Iter	200th Iter
CO_2 ($\times 10^6$ mg)	MP	284.22 ± 0.01	281.83 ± 0.01	281.27 ± 0.01	281.10 ± 0.01	281.03 ± 0.01
	MPLight	277.74 ± 0.61	276.60 ± 0.01	280.45 ± 0.24	290.13 ± 0.38	319.82 ± 0.08
	Standard DQN	276.24 ± 0.01	275.59 ± 1.05	275.84 ± 1.65	274.68 ± 2.99	269.76 ± 3.71
	CSRD-U (ours)	278.45 ± 0.25	276.83 ± 0.79	269.66 ± 3.31	267.79 ± 3.60	263.43 ± 4.66
Duration (s)	MP	82.08 ± 0.01	81.95 ± 0.01	81.67 ± 0.01	81.62 ± 0.01	81.10 ± 0.01
	MPLight	90.74 ± 0.04	86.27 ± 0.01	87.52 ± 0.10	95.96 ± 0.18	118.55 ± 0.02
	Standard DQN	87.76 ± 0.01	84.53 ± 0.71	82.06 ± 1.10	81.15 ± 1.76	78.39 ± 2.41
	CSRD-U (ours)	87.81 ± 0.04	82.36 ± 0.34	78.74 ± 2.10	76.10 ± 2.47	74.27 ± 2.21
Wait Time (s)	MP	17.94 ± 0.01	17.50 ± 0.01	17.06 ± 0.01	17.09 ± 0.01	17.23 ± 0.01
	MPLight	22.32 ± 0.01	19.87 ± 0.01	21.67 ± 0.02	28.23 ± 0.09	44.53 ± 0.01
	Standard DQN	22.92 ± 0.01	18.99 ± 0.44	18.25 ± 0.56	18.55 ± 1.11	18.21 ± 1.51
	CSRD-U (ours)	22.56 ± 0.14	19.11 ± 0.26	21.07 ± 1.79	22.44 ± 2.05	22.84 ± 1.92
Queue (vehicles)	MP	2.47 ± 0.01	2.40 ± 0.01	2.42 ± 0.01	2.38 ± 0.01	2.37 ± 0.01
	MPLight	2.89 ± 0.02	2.67 ± 0.01	2.70 ± 0.01	2.99 ± 0.01	3.69 ± 0.01
	Standard DQN	2.78 ± 0.01	2.61 ± 0.05	2.44 ± 0.06	2.34 ± 0.09	2.18 ± 0.10
	CSRD-U (ours)	2.76 ± 0.01	2.43 ± 0.02	2.14 ± 0.10	1.81 ± 0.08	1.66 ± 0.07

Table A13. Longitudinal training metrics measuring progressive performance improvements for Scenario 3. The data present means and standard deviations over five independent runs specifically when the specialized topography-aware reward is activated.

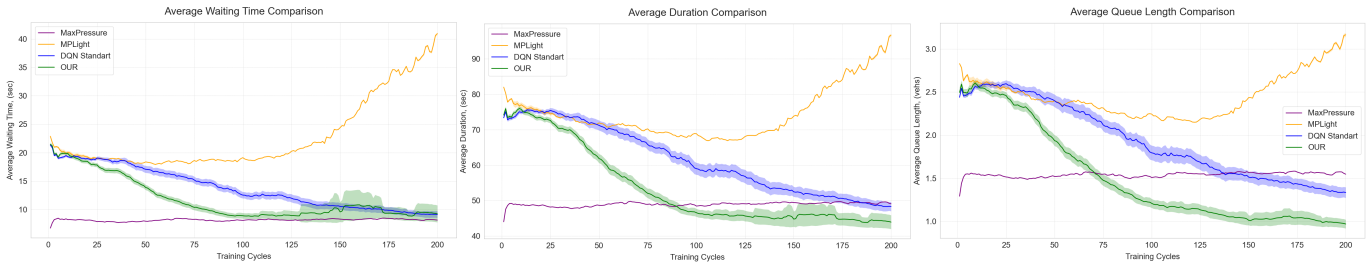
Metric	Approach	1st Iter	50th Iter	100th Iter	150th Iter	200th Iter
CO_2 ($\times 10^6$ mg)	MP	284.22 ± 0.01	281.83 ± 0.01	281.27 ± 0.01	281.10 ± 0.01	281.03 ± 0.01
	MPLight	277.74 ± 0.61	276.60 ± 0.01	280.45 ± 0.24	290.13 ± 0.38	319.82 ± 0.08
	Standard DQN	276.24 ± 0.01	275.59 ± 1.05	275.84 ± 1.65	274.68 ± 2.99	269.76 ± 3.71
	CSRD-T (ours)	279.20 ± 0.10	275.79 ± 1.01	267.18 ± 1.94	260.94 ± 4.21	256.97 ± 2.36
Duration (s)	MP	82.08 ± 0.01	81.95 ± 0.01	81.67 ± 0.01	81.62 ± 0.01	81.10 ± 0.01
	MPLight	90.74 ± 0.04	86.27 ± 0.01	87.52 ± 0.10	95.96 ± 0.18	118.55 ± 0.02
	Standard DQN	87.76 ± 0.01	84.53 ± 0.71	82.06 ± 1.10	81.15 ± 1.76	78.39 ± 2.41
	CSRD-T (ours)	87.96 ± 0.16	82.50 ± 0.78	76.26 ± 1.35	71.79 ± 3.35	72.09 ± 3.98
Wait Time (s)	MP	17.94 ± 0.01	17.50 ± 0.01	17.06 ± 0.01	17.09 ± 0.01	17.23 ± 0.01
	MPLight	22.32 ± 0.01	19.87 ± 0.01	21.67 ± 0.02	28.23 ± 0.09	44.53 ± 0.01
	Standard DQN	22.92 ± 0.01	18.99 ± 0.44	18.25 ± 0.56	18.55 ± 1.11	18.21 ± 1.51
	CSRD-T (ours)	22.90 ± 0.01	19.00 ± 0.34	18.69 ± 0.90	21.25 ± 2.71	24.67 ± 4.50

Table A13. *Cont.*

Metric	Approach	1st Iter	50th Iter	100th Iter	150th Iter	200th Iter
Queue (vehicles)	MP	2.47 ± 0.01	2.40 ± 0.01	2.42 ± 0.01	2.38 ± 0.01	2.37 ± 0.01
	MPLight	2.89 ± 0.02	2.67 ± 0.01	2.70 ± 0.01	2.99 ± 0.01	3.69 ± 0.01
	Standard DQN	2.78 ± 0.01	2.61 ± 0.05	2.44 ± 0.06	2.34 ± 0.09	2.18 ± 0.10
	CSRD-T (ours)	2.78 ± 0.02	2.43 ± 0.05	2.00 ± 0.05	1.55 ± 0.06	1.31 ± 0.08

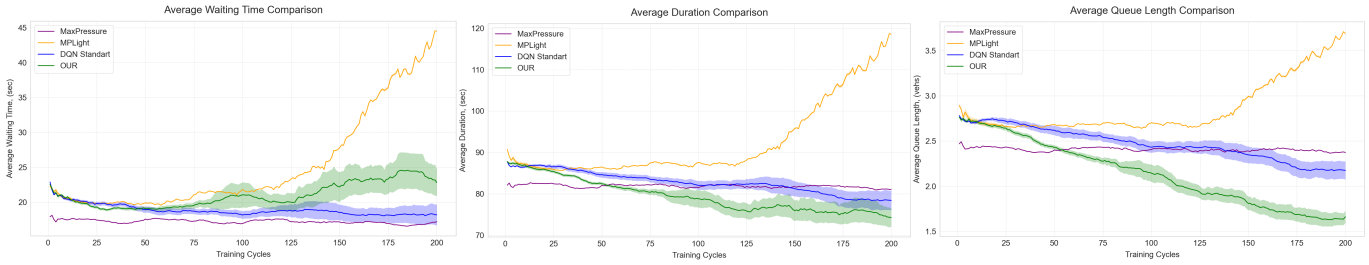
Appendix C

This appendix contains the additional figures visualizing the training trajectories for waiting time, trip duration, and queue length across the three evaluated scenarios.



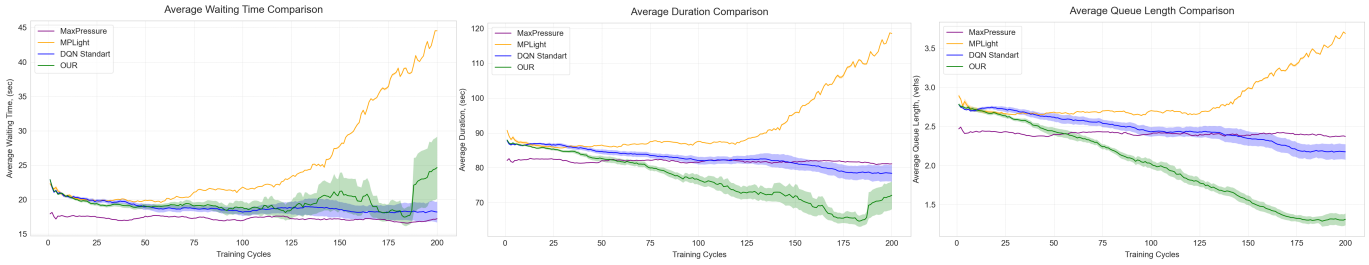
(**a**) (**b**) (**c**)

Figure A1. The training trajectories of the traffic controllers evaluated under the flat terrain configuration of Scenario 1 across three secondary metrics: (**a**) average waiting time, (**b**) average trip duration, and (**c**) average queue length.



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

Figure A2. The performance degradation observed when the context-agnostic universal controllers operate within the 3.8-degree uphill geometry of Scenario 2 across three secondary metrics: (**a**) average waiting time, (**b**) average trip duration, and (**c**) average queue length.



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

Figure A3. The recovery of operational stability in Scenario 3 following the activation of the specialized topography-aware reward across three secondary metrics: (**a**) average waiting time, (**b**) average trip duration, and (**c**) average queue length.

References

1. United Nations General Assembly; United Nations Economic and Social Council. Progress Towards the Sustainable Development Goals: Report of the Secretary-General. Report of the Secretary-General A/79/79-E/2024/54, United Nations. 2024. Available online: https://unstats.un.org/sdgs/files/report/2024/SG-SDG-Progress-Report-2024-advanced-unedited-version.pdf (accessed on 24 February 2026).
2. Coiret, A.; Deljanin, E.; Vandanjon, P.O. Vehicle energy savings by optimizing road speed-sectioning. *Eur. Transp. Res. Rev.* **2020**, *12*, 41. [CrossRef]
3. Ryzhanskyi, O.; Manziuk, E.; Barmak, O.; Krak, I.; Bacanin, N. An approach to optimizing CO_2 emissions in traffic control via reinforcement learning. In *Proceedings of the 5th International Workshop on Intelligent Information Technologies & Systems of Information Security (IntelITSIS 2024), Khmelnytskyi, Ukraine, 28 March 2024*; Hovorushchenko, T., Savenko, O., Popov, P.T., Lysenko, S., Eds.; CEUR: Aachen, Germany, 2024; Volume 3675, pp. 137–155. Available online: https://ceur-ws.org/Vol-3675/paper10.pdf (accessed on 24 February 2026).
4. Pavlyshyn, V.; Manziuk, E.; Barmak, O.; Krak, I.; Damaševičius, R. Modeling environment intelligent transport system for eco-friendly urban mobility. In *Proceedings of the 5th International Workshop on Intelligent Information Technologies & Systems of Information Security (IntelITSIS 2024), Khmelnytskyi, Ukraine, 28 March 2024*; Hovorushchenko, T., Savenko, O., Popov, P.T., Lysenko, S., Eds.; CEUR: Aachen, Germany, 2024; Volume 3675, pp. 118–136. Available online: https://ceur-ws.org/Vol-3675/paper9.pdf (accessed on 24 February 2026).
5. Ryzhanskyi, O.; Pavlyshyn, V.; Radiuk, P.; Manziuk, E.; Barmak, O.; Krak, I. AI-driven traffic signal control system to reduce CO_2 emissions. In *Proceedings of the the Workshops "AI for Environmental and Social Sustainability" and "AI and Interdisciplinary Innovations for Sustainable Development" (YAISD-WS 2025), Ternopil-Skomorochy, Ukraine, 8–9 May 2025*; Pitsun, O., Dyvak, M., Eds.; CEUR: Aachen, Germany, 2025; Volume 3974, pp. 18–27. Available online: https://ceur-ws.org/Vol-3974/paper02.pdf (accessed on 24 February 2026).
6. Pavlyshyn, V.; Ryzhanskyi, O.; Manziuk, E.; Radiuk, P.; Barmak, O.; Krak, I. Establishing patterns of the urban transport flows on clustering analysis. In *Proceedings of the the Workshops "AI for Environmental and Social Sustainability" and "AI and Interdisciplinary Innovations for Sustainable Development" (YAISD-WS 2025), Ternopil-Skomorochy, Ukraine, 8–9 May 2025*; Pitsun, O., Dyvak, M., Eds.; CEUR: Aachen, Germany, 2025; Volume 3974, pp. 1–9. Available online: https://ceur-ws.org/Vol-3974/paper01.pdf (accessed on 24 February 2026).
7. Louw, J.; Labuschagne, L.; Woodley, T. A comparison of reinforcement learning agents applied to traffic signal optimisation. In *Proceedings of the SUMO User Conference 2022*; SUMO Conference Proceedings; TIB Open Publishing: Hannover, Germany , 2022; Volume 3, pp. 15–43. [CrossRef]
8. Wei, H.; Zheng, G.; Gayah, V.V.; Li, Z. Recent advances in reinforcement learning for traffic signal control: A survey of models and evaluation. *ACM SIGKDD Explor. Newsl.* **2020**, *22*, 12–18. [CrossRef]
9. Haddad, T.A.; Hedjazi, D.; Aouag, S. A new deep reinforcement learning-based adaptive traffic light control approach for isolated intersection. In Proceedings of the 2022 5th International Symposium on Informatics and Its Applications (ISIA), M'sila, Algeria, 29–30 November 2022; pp. 1–6. [CrossRef]
10. Ouyang, C.; Zhan, Z.; Lv, F. A comparative study of traffic signal control based on reinforcement learning algorithms. *World Electr. Veh. J.* **2024**, *15*, 246. [CrossRef]
11. Eclipse SUMO Project. Eclipse SUMO. 2026. Available online: https://eclipse.dev/sumo/ (accessed on 24 February 2026).
12. HBEFA. Handbook of Emission Factors for Road Transport. 2026. Available online: https://www.hbefa.net/ (accessed on 24 February 2026).
13. Eclipse SUMO Project. TraCI: Traffic Control Interface. 2026. Available online: https://sumo.dlr.de/docs/TraCI/index.html (accessed on 24 February 2026).
14. Krajzewicz, D.; Behrisch, M.; Wagner, P.; Luz, R.; Krumnow, M. Second generation of pollutant emission models for SUMO. In *Modeling Mobility with Open Data*; Behrisch, M., Weber, M., Eds.; Lecture Notes in Mobility; Springer: Cham, Switzerland, 2015; pp. 203–221. [CrossRef]
15. Pavlyshyn, V.; Manziuk, E.; Barmak, O.; Radiuk, P.; Krak, I. An adaptive machine learning approach to sustainable traffic planning: High-fidelity pattern recognition in smart transportation systems. *Future Transp.* **2025**, *5*, 152. [CrossRef]
16. Eom, M.; Kim, B.I. The traffic signal control problem for intersections: A review. *Eur. Transp. Res. Rev.* **2020**, *12*, 50. [CrossRef]
17. Chala, T.D.; Kóczy, L.T. Intelligent fuzzy traffic signal control system for complex intersections using fuzzy rule base reduction. *Symmetry* **2024**, *16*, 1177. [CrossRef]
18. Wang, A.; Zhang, K.; Li, M.; Shao, J.; Li, S. Game theory-based signal control considering both pedestrians and vehicles in connected environment. *Sensors* **2023**, *23*, 9438. [CrossRef]
19. Faqir, N.; Loqman, C.; Boumhidi, J. Combined extreme learning machine and max pressure algorithms for traffic signal control. *Intell. Syst. Appl.* **2023**, *19*, 200255. [CrossRef]

20. Ren, W.; Zhang, J.; Li, L.; Zhou, Q. An intersection platoon speed control model considering traffic efficiency and energy consumption in CVIS. *Math. Probl. Eng.* **2021**, *2021*, 2891247. [CrossRef]
21. Abohashima, H.; Eltawil, A.B.; Gheith, M.S. A mathematical programming model and a firefly-based heuristic for real-time traffic signal scheduling with physical constraints. *IEEE Access* **2021**, *9*, 128314–128327. [CrossRef]
22. Othmani, T.; Boubaker, S.; Rehimi, F.; Alimi, S.E. Improving energy consumption efficiency and environmental sustainability through smart traffic control adaptation. In Proceedings of the 2023 IEEE International Conference on Artificial Intelligence & Green Energy (ICAIGE), Sousse, Tunisia, 12–14 October 2023; pp. 1–6. [CrossRef]
23. Sachan, A.; Kumar, N. Intelligent traffic control system for emergency vehicles. In *IoT and Analytics for Sensor Networks*; Nayak, P., Pal, S., Peng, S.L., Eds.; Lecture Notes in Networks and Systems; Springer: Singapore, 2022; Volume 244, pp. 151–160. [CrossRef]
24. Levin, M.W. Max-pressure traffic signal timing: A summary of methodological and experimental results. *J. Transp. Eng. Part A Syst.* **2023**, *149*, 03123001. [CrossRef]
25. Kouvelas, A.; Lioris, J.; Fayazi, S.A.; Varaiya, P. Maximum pressure controller for stabilizing queues in signalized arterial networks. *Transp. Res. Rec. J. Transp. Res. Board* **2014**, *2421*, 133–141. [CrossRef]
26. Haydari, A.; Yilmaz, Y. Deep reinforcement learning for intelligent transportation systems: A survey. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 11–32. [CrossRef]
27. Michailidis, P.; Michailidis, I.; Lazaridis, C.R.; Kosmatopoulos, E. Traffic signal control via reinforcement learning: A review on applications and innovations. *Infrastructures* **2025**, *10*, 114. [CrossRef]
28. Ault, J.; Hanna, J.P.; Sharon, G. Learning an interpretable traffic signal control policy. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, 9–13 May 2020*; Fallah-Seghrouchni, A.E., Sukthankar, G., An, B., Yorke-Smith, N., Eds.; International Foundation for Autonomous Agents and Multiagent Systems: Pittsburgh, PA, USA, 2020; pp. 88–96. Available online: https://www.ifaamas.org/Proceedings/aamas2020/pdfs/p88.pdf (accessed on 24 February 2026).
29. Chen, C.; Wei, H.; Xu, N.; Zheng, G.; Yang, M.; Xiong, Y.; Xu, K.; Li, Z. Toward a thousand lights: Decentralized deep reinforcement learning for large-scale traffic signal control. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 3414–3421. [CrossRef]
30. Ivander, W.; Kozhevnikov, S.; Sontheimer, M.; Chou, S. Enhancing urban traffic management in Taipei: A reinforcement learning approach. In Proceedings of the 2024 Smart City Symposium Prague (SCSP), Prague, Czech Republic, 23–24 May 2024; pp. 1–6. [CrossRef]
31. Sun, Z.; Jia, X.; Cai, Y.; Ji, A.; Lin, X.; Liu, L.; Wang, W.; Tu, Y. Joint control of traffic signal phase sequence and timing: A deep reinforcement learning method. *Digit. Transp. Saf.* **2025**, *4*, 118–126. [CrossRef]
32. Bellman, R. A Markovian decision process. *J. Math. Mech.* **1957**, *6*, 679–684. [CrossRef]
33. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [CrossRef]
34. Liu, D.; Li, L. A traffic light control method based on multi-agent deep reinforcement learning algorithm. *Sci. Rep.* **2023**, *13*, 9396. [CrossRef]
35. Kolat, M.; Kővári, B.; Bécsi, T.; Aradi, S. Multi-agent reinforcement learning for traffic signal control: A cooperative approach. *Sustainability* **2023**, *15*, 3479. [CrossRef]
36. Zhang, Z.; Zhang, W.; Liu, Y.; Xiong, G. Mean field multi-agent reinforcement learning method for area traffic signal control. *Electronics* **2023**, *12*, 4686. [CrossRef]
37. Fang, W.; Zhao, X.; Zhang, C. Fairness-aware multi-agent reinforcement learning and visual perception for adaptive traffic signal control. *Optoelectron. Lett.* **2024**, *20*, 764–768. [CrossRef]
38. Nie, Q.; Ou, J.; Zhang, H.; Lu, J.; Li, S.; Shi, H. A robust integrated multi-strategy bus control system via deep reinforcement learning. *Eng. Appl. Artif. Intell.* **2024**, *133*, 107986. [CrossRef]
39. Ran, Q.; Liang, C.; Liu, P. A safe lane-changing strategy for autonomous vehicles based on deep Q-networks and prioritized experience replay. *Digit. Transp. Saf.* **2025**, *4*, 170–174. [CrossRef]
40. Wang, Z.; Xu, L.; Ma, J. Carbon dioxide emission reduction-oriented optimal control of traffic signals in mixed traffic flow based on deep reinforcement learning. *Sustainability* **2023**, *15*, 16564. [CrossRef]
41. Schumacher, M.E.H.; Adriano, C.M.; Giese, H. Challenges in reward design for reinforcement learning-based traffic signal control: An investigation using a CO_2 emission objective. In *Proceedings of the SUMO User Conference 2023*; SUMO Conference Proceedings; TIB Open Publishing: Hannover, Germany, 2023; Volume 4, pp. 131–151. [CrossRef]
42. Zhang, G.; Chang, F.; Jin, J.; Yang, F.; Huang, H. Multi-objective deep reinforcement learning approach for adaptive traffic signal control system with concurrent optimization of safety, efficiency, and decarbonization at intersections. *Accid. Anal. Prev.* **2024**, *199*, 107451. [CrossRef]
43. Ounoughi, C.; Touibi, G.; Yahia, S.B. EcoLight: Eco-friendly traffic signal control driven by urban noise prediction. In *Database and Expert Systems Applications*; Strauss, C., Cuzzocrea, A., Kotsis, G., Tjoa, A.M., Khalil, I., Eds.; Lecture Notes in Computer Science; Springer: Vienna, Austria, 2022; Volume 13426, pp. 205–219. [CrossRef]

44. Jung, J.; Kim, I.; Yoon, J. EcoMRL: Deep reinforcement learning-based traffic signal control for urban air quality. *Int. J. Sustain. Transp.* **2025**, *19*, 720–729. [CrossRef]
45. Kővári, B.; Szőke, L.; Bécsi, T.; Aradi, S.; Gáspár, P. Traffic signal control via reinforcement learning for reducing global vehicle emission. *Sustainability* **2021**, *13*, 11254. [CrossRef]

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