mathematics

MDPI

Article

Explicit Coordinated Signal Timing Optimization Using Soft Actor–Critic and Model Predictive Control Along an Arterial

Kun Zhang [1,2], Helai Huang [1,2], Hanchu Zhou [1,2], Junwu Liu [1,2], Qiming Zheng [3] and Hongfeng Xu [4,*]

[1] School of Traffic and Transportation Engineering, Central South University, Changsha 410075, China
[2] Smart Transport Operation Laboratory, Changsha Communications Investment Holding Group Co., Ltd., Changsha 410014, China
[3] Neusoft Research, Neusoft (China), Shenyang 110179, China
[4] School of Infrastructure Engineering, Dalian University of Technology, Dalian 116024, China
* Correspondence: hfxu@dlut.edu.cn

Abstract

Arterial signal coordination requires coupled decisions on cycle length, phase green times, and offsets in an explicit timing plan. This study presents a modular explicit coordinated signal timing optimization (ECSTO) method for a three-intersection arterial. At each cycle-level decision, a soft actor–critic (SAC) agent selects a common background cycle length and offset corrections for the upcoming cycle. A time–space relationship provides base offsets. Modified model predictive control (MMPC) then allocates phase green times using non-negative, intersection-specific cycle adjustments derived from changes between consecutive offset corrections. It does so without a separately scheduled explicit transition cycle. SAC is trained with a phase-weighted discharged-vehicle reward. System vehicle delay and coordinated-direction space-mean speed are used to evaluate and select policies. Across 1000 common demand scenarios in the simulated arterial, ECSTO had 7.16–11.73% lower aggregate mean system vehicle delay and 5.30–10.19% lower total vehicle travel time than five selected explicit timing-control baselines. These reductions remained statistically significant after Holm adjustment. For coordinated-direction space-mean speed, its mean was significantly below four baselines, illustrating a trade-off between system-wide efficiency and coordinated-direction speed. Ablation results describe differences among four tested timing configurations. The findings are limited to this three-intersection simulation and do not establish field performance or effects in larger networks.

Keywords: arterial signal coordination; model predictive control; offset adjustment; signal timing optimization; soft actor–critic

MSC: 93C40; 93B45; 68T05

Check for updates

Academic Editor: Cheng-Hung Huang

Received: 29 September 2026
Revised: 22 October 2026
Accepted: 30 October 2026
Published: 5 November 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

Arterial signal coordination involves coupled decisions on cycle length, phase green times, and offsets across adjacent intersections. The cycle establishes a shared timing reference, green times allocate service to competing movements, and offsets position the coordinated green intervals in time. These explicit parameters allow the timing plans to be inspected and operational constraints to be stated directly.

Classical coordination methods determine these parameters using timing formulas, precomputed plans, or detector-driven rules. Their performance depends on the assumptions and inputs used to construct the plans. Reinforcement learning (RL) selects control

actions from observed traffic states [1–3]. When an RL policy directly changes phases, its actions may need to be translated into the cycle length, green times, and offsets used by conventional coordinated controllers [4–6]. This motivates assigning learning to selected timing variables and determining the remaining variables under explicit constraints.

The proposed explicit coordinated signal timing optimization (ECSTO) method extends our earlier SAC-based study of cycle length [7]. The extension adds learned offset corrections and incorporates their updates into green-time optimization. A soft actor–critic (SAC) agent selects the common background cycle length and cycle-level offset corrections. A time–space relationship provides base offsets. The previously developed modified model predictive control (MMPC) formulation [8], originally used to optimize green times for a specified cycle length, allocates phase green times and incorporates offset-related cycle adjustments. This organization retains explicit timing parameters and connects the learned cycle length and offset decisions to constrained green-time optimization.

The main contributions of this study are summarized as follows:

- SAC-based cycle-length selection is extended to include offset corrections and connected to the existing MMPC green-time optimization. The resulting modular framework covers the three principal coordinated timing parameters and defines the timing relationships between the learning and optimization modules.
- Changes between consecutive offset corrections are converted into non-negative intersection-specific cycle adjustments. These adjustments preserve relative changes in offset correction and enter MMPC green-time optimization without a separately scheduled explicit transition cycle.
- A three-intersection simulation evaluates the complete framework against three simplified configurations and five selected explicit timing-control baselines. The comparisons assess configuration-level performance and the framework's operational value under the tested conditions.

2. Literature Review

2.1. Traditional Methods for Signal Coordination

Fixed-time, actuated, and adaptive approaches differ in how timing plans are generated and how they respond to observed demand. These differences concern both the calculation of cycle length, green times, and offsets and the rules governing their implementation.

Fixed-time control relies on historical traffic data to design pre-specified timing plans. Owing to its simplicity, low implementation cost, and compatibility with conventional controller settings, it remains widely used in practice, especially on arterials with limited detection infrastructure. It responds to demand variation through scheduled plan changes and periodic retiming rather than direct adjustment to short-term fluctuations. Its suitability depends on how closely design demand matches operating conditions and on the resources available for maintaining the plans [9].

Actuated coordination uses detector information to adjust phase green times in response to short-term demand [10,11]. The background timing plan specifies the cycle, splits, and offset. Within these settings, actuated logic can reallocate unused green time to other phases, but its operational performance still depends on the quality of the background timing plan [12].

Adaptive coordination updates timing decisions from observed traffic conditions [13]. Model-based approaches formulate traffic relationships and control objectives [14–17], whereas rule-based approaches generate timing adjustments from prescribed response relationships [18–20]. Lin et al. combined fuzzy control and differential evolution to optimize a weighted, normalized combination of delay and emissions [21]. Their formulation balances operational and environmental objectives. Its queue model assumes undersatura-

tion, and its stop-rate derivation assumes uniform arrivals. Chala and Kóczy developed decentralized fuzzy controllers that use queues, waiting times, and vehicle proximity to adjust green durations across multiple intersections [22]. Reducing the rule base limits complexity, but the membership-function parameters still require calibration for the application. Such approaches remain dependent on prior design: performance-based decisions rely on the applicability of the traffic-model assumptions, while rule-based decisions rely on rules and parameter settings suited to the operating conditions.

Traditional coordination uses traffic models, engineering rules, or optimization searches to develop timing plans that address control objectives and operational requirements [23]. However, model assumptions and prescribed rules may not fully capture interactions among intersections under complex traffic conditions [24]. In such cases, parameter adjustment alone may be insufficient, and the model structure or control rules may also need revision [25]. RL learns coordination policies from operational feedback, offering another way to address traffic interactions that are difficult to specify adequately in advance.

2.2. RL-Based Methods for Signal Coordination

RL learns control policies through interaction with the environment and reward feedback. Its role in coordination depends on the decisions assigned to the policy. In arterial coordination, existing RL-based studies can be broadly categorized into implicit coordination and explicit coordination [7]. The former develops coordination through learned phase decisions, while the latter retains timing parameters to express the relationships between intersections.

In implicit coordination, a learned policy issues actions such as extending or terminating a green phase, without necessarily producing a common cycle length or an offset. Learning phase decisions allows green times to be adjusted flexibly in response to traffic conditions [26–28]. Coordination can be supported by exchanging traffic information between agents or learning joint decisions across intersections [29–31]. CenLight decomposes the joint state-action space to reduce the search space and uses a recurrent network to learn dependencies between intersection actions. Its reward is the negative sum of approach-queue lengths at each intersection [32]. Ćiprovski et al. used traffic-informed fuzzy priorities to guide phase exploration without changing the learning objective or observations [33]. CenLight organized joint-action search; the fuzzy-priority method guides exploration with traffic knowledge. Without a prescribed common cycle or fixed offsets, these formulations can explore signal sequences beyond conventional cyclic plans, offering potential performance gains. This freedom also broadens the policy's task to learning both green-time service and inter-intersection timing relationships.

In RL-based explicit coordination, timing parameters remain policy outputs or variables in coupled optimization [34,35]. These parameters define a prescribed timing structure within which learning, rules, or models can determine individual timing decisions. Li et al. used a queue-based reward with an additional penalty for severe congestion. Their single agent selects one intersection and adjusts its north–south phase-time allocation, while cycle length, offset, and left-turn durations remain fixed [36]. This restricts the action space to intersection selection and incremental split adjustment. With a given background cycle, HyLight couples RL-based offset control with rule-based phase splits [37]. Its reward combines local and neighboring time-averaged queues to align offset learning with delay reduction. The split mechanism is included in the offset-learning environment to account for interactions between these parameters. Such methods divide timing tasks and concentrate learning on selected parameters. This narrows the policy's task, but the remaining timing responses are constrained by fixed settings or prescribed rules.

Implicit and explicit coordination differ in which timing relationships are prescribed and which are learned. Within an explicit timing structure, allowing cycle length, green times, and offsets to vary can remove restrictions imposed by fixing these parameters. Their interactions must then be addressed. Compared with approaches that fix some timing parameters or allocate green times using prescribed rules, this study links coordination-parameter updates with constrained reoptimization of green times. Phase service times can therefore be reallocated under the updated coordination conditions.

3. Modular ECSTO Framework and Methodology

3.1. Overall Framework

Figure 1 shows the cycle-level control sequence. Detector-derived traffic features and the implemented timing plan form the state for the next decision. SAC selects a common background cycle length and offset corrections for the upcoming cycle. The time–space relationship provides base offsets. MMPC uses the background cycle length and offset-related cycle adjustments to determine phase green times. The plan is implemented before the next observation and reward are collected.

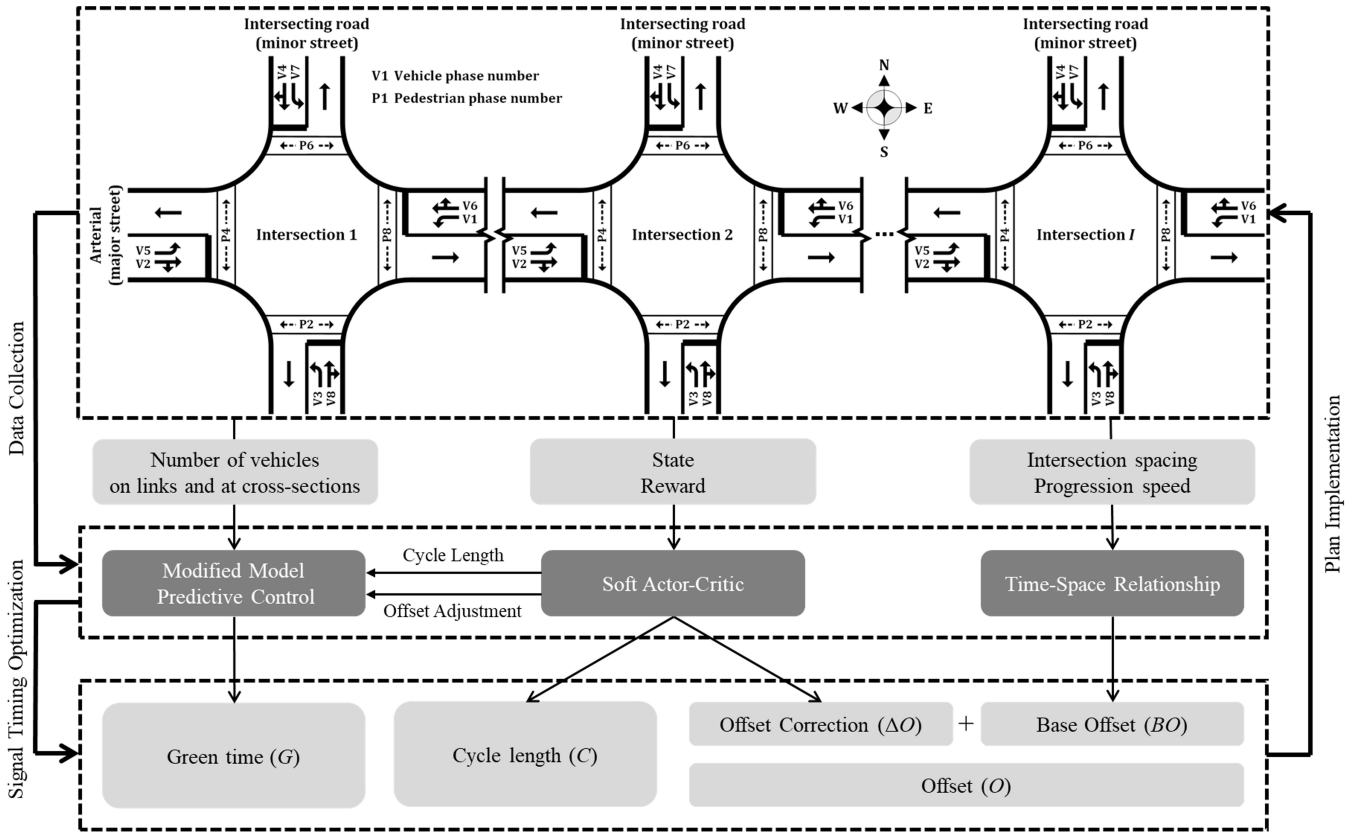


Figure 1. ECSTO modules and data flow, with control decisions updated at cycle level.

SAC selects a common background cycle length. The cycle length implemented at an intersection may also include an offset-related cycle adjustment. Table 1 lists the timing quantities used below.

Table 1. Notation for the cycle-level control sequence.

Symbol	Meaning
$C^{\mathrm{bg}}(n)$	Common background cycle length selected by SAC for cycle n (s).
$C_i^{\mathrm{impl}}(n)$	Cycle length implemented at intersection i (s).
$BO_i, \Delta O_i(n), O_i(n)$	Base offset, offset correction, and resulting offset of the coordinated phase (s).
$u_i(n)$	Adjacent-intersection increment accumulated to obtain $\Delta O_i(n)$ (s).
$OR_i(n), SR(n)$	Offset reference point at intersection i and system time reference point (s).
d_i, m, a_i	Change in offset correction, its minimum across intersections, and the resulting non-negative cycle adjustment (s).

3.2. SAC-Based Optimization of Cycle Length and Offset Correction

Traditional methods for determining cycle length and offset typically rely on explicit modeling of traffic dynamics and progression relationships, which often requires restrictive assumptions. In ECSTO, these two variables are treated as higher-level coordination parameters whose effects unfold over multiple cycles. To improve adaptability while avoiding hand-crafted dynamic models, this study uses SAC to optimize the common background cycle length and cycle-level offset correction through agent–environment interaction. This subsection describes the RL formulation, including the action space, state representation, reward design, and training settings.

3.2.1. Action

During cycle n, SAC selects the common background cycle length and target offset corrections for cycle $n+1$. MMPC then determines phase green times. In $n+1 \mid n$, the first term identifies the cycle when the plan takes effect, and the second identifies the cycle when information is available for selecting the discrete action:

$$A_{n+1|n} = (C^{\mathrm{bg}}(n+1), \Delta O_1(n+1), \Delta O_2(n+1), \ldots, \Delta O_I(n+1)), \tag{1}$$

Here, $C^{\mathrm{bg}}(n+1)$ is the selected background cycle length, and $\Delta O_i(n+1)$ is the decoded target offset correction at intersection i. The background cycle length and decoded offset corrections satisfy

$$C^{\mathrm{bg}}(n+1) \in \left\{ C \in \mathbb{N} \;\middle|\; \begin{array}{c} MinC \le C \le MaxC \\ C \equiv 0 \pmod{CTS} \end{array} \right\} \tag{2}$$

$$\Delta O_1(n+1) = 0, \qquad \Delta O_i(n+1) - \Delta O_{i-1}(n+1) \in \{0,2,4,6,8\}, \quad i = 2, \ldots, I. \tag{3}$$

Here, $MinC$ and $MaxC$ bound the background cycle length selected by SAC, and $CTS = 10$ s is the discretization step. The experimental values of $MinC = 60$ s and $MaxC = 180$ s give 13 background cycle-length actions. The experimental maximum of 180 s applies to this background cycle length. An intersection's implemented cycle length also includes the adjustment described below. Intersection 1 has a fixed zero offset correction and provides the offset reference.

SAC selects independent increments of $u_i(n+1) \in \{0,2,4,6,8\}$ for $i = 2, \ldots, I$, with $u_1(n+1) = 0$. These increments are accumulated along the coordinated direction: $\Delta O_i(n+1) = \sum_{\ell=2}^{i} u_\ell(n+1)$ for $i > 1$. In the three-intersection testbed, $\Delta O_2 = u_2$ and $\Delta O_3 = u_2 + u_3$; the latter can reach 16 s but is coupled to ΔO_2. The 13 background cycle-length choices and two five-valued increments give $|A| = 325$ discrete joint actions.

The tested controller adopts non-negative adjacent increments as a simplifying action representation. A zero increment retains the base adjacent green-start lag, while positive increments provide bounded delays selected jointly with the background cycle length. These delay options can accommodate later platoon arrivals or a smaller downstream queue discharge requirement than assumed in the base calculation.

3.2.2. State

The cycle-length feature in the SAC state is the implemented cycle length (C_i^{impl}), which includes any offset-related cycle adjustment. It is distinct from the common background cycle length (C^{bg}) selected by SAC.

As illustrated in Figure 2, the state is constructed using features from all intersections along the arterial. Each intersection includes the cycle length and four legs: north, east, south, and west. Every leg is subdivided into three segments labeled A, B, and C, which represent the through and right-turn lanes, the left-turn lanes, and the upstream segments, respectively. The 11 features for each leg include phase green times and traffic measurements from these segments. The vehicle passing rate is measured between the stop line and the beginning of Segment C, representing the farthest upstream point within radar coverage. Both instantaneous and average densities are calculated based on Segments A, B, and C. The state aggregates observations from the detection segments at all intersections, while spatial coverage remains limited at each approach.

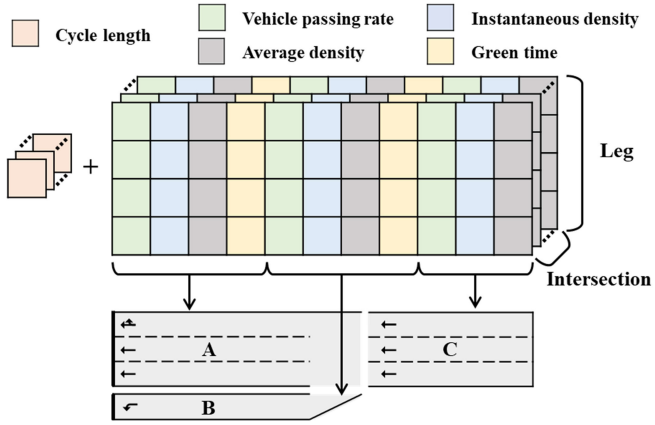


Figure 2. State representation. Lane-use arrows indicate the permitted movements for each lane.

3.2.3. Reward

The SAC training reward counts vehicles discharged past the stop lines during each reward interval [38] and applies phase weights. System vehicle delay and coordinated-direction space-mean speed are used later for policy evaluation and are not additive reward terms. The reward is

$$R_{t+1} = \sum_{i=1}^{I} \sum_{j=1}^{8} \beta_{i,\mathrm{V}j} \cdot NV_{t,t+1}^{i,\mathrm{V}j}, \tag{4}$$

$$\beta_{i,\mathrm{V}j} = \begin{cases} \beta_{\text{Coord}} & \mathrm{V}j \text{ is the coordinated phase} \\ \beta_{\text{NoCoord}} & \mathrm{V}j \text{ is the uncoordinated phase} \end{cases}, \tag{5}$$

where R_{t+1} denotes the reward obtained at time step $t+1$ and $NV_{t,t+1}^{i,\mathrm{V}j}$ is the number of vehicles passing the stop line for phase Vj at intersection i during the interval from time step t to $t+1$. The $\beta_{i,\mathrm{V}j}$ coefficient is the phase weight assigned to that phase.

The $\beta_{i,\mathrm{V}j}$ coefficients are determined by rule-based classification, where β_{Coord} and β_{NoCoord} correspond to coordinated and uncoordinated phases, respectively.

3.2.4. Network Structure and Training

The SAC framework [39,40] includes five neural networks: one actor for policy generation, two critics for Q-value estimation, and two target critics to stabilize training. All networks share a unified structure with two fully connected hidden layers of 512 neurons and ReLU activation functions. The actor network outputs a probabilistic policy distribution, enabling efficient exploration through entropy maximization.

During training, transitions are stored in a replay buffer, and the actor and critic networks are updated iteratively using entropy-regularized loss functions. The target critic networks are updated via soft updates to enhance training stability. The entropy temperature (α) is automatically tuned to balance exploration and exploitation, guided by a target entropy of $\bar{H} = -0.5\log(1/|A|)$ [41], where $|A|$ denotes the number of discrete actions.

Training begins after a warm-up stage during which sufficient transitions are collected. The agent is trained using the Adam optimizer. The discount factor (γ) is set to 0.944. The approximate effective time window of $ETW = \frac{T_R}{1-\gamma}$ describes the discounting timescale, where $T_R = 10$ s is the reward-discounting interval [42]. This interval is distinct from the cycle-length action discretization step (CTS), even though both are 10 s in this implementation. The resulting window is approximately 178.57 s; it is not a finite-horizon guarantee.

A summary of the remaining hyperparameters is provided in Table 2.

Table 2. Hyperparameter settings.

Hyperparameter	Value
Batch size	64 [40]
Learning rate	0.0003 [39]
Replay buffer size	10,000
Replay start threshold	1000
Soft update coefficient (τ)	0.005 [39]

3.3. Base Offset from the Time–Space Relationship

The base offset between coordinated phases is calculated using the time–space relationship, a well-established method that accounts for intersection spacing, travel time, and queue discharge time. In ECSTO, the base offset serves as a progression-oriented reference term rather than a complete solution for offset optimization. It represents an idealized scenario in which the lead vehicle of an upstream coordinated platoon can traverse the downstream intersection without stopping. The base offsets are calculated as follows:

$$BO_{i,\mathrm{V2}} = \begin{cases} 0 & i = 1 \\ BO_{i-1,\mathrm{V2}} + TT_{i-1,\mathrm{V2}} - QD_{i,\mathrm{V2}} & i > 1 \end{cases}, \quad (6)$$

Here, $BO_{i,\mathrm{V2}}$ is the base offset of coordinated phase V2 at intersection i. The term $TT_{i-1,\mathrm{V2}}$ is the travel time from intersection $i-1$ to i, and $QD_{i,\mathrm{V2}}$ is the queue discharge time at intersection i. The resulting offset combines the base offset with SAC's offset correction.

3.4. MMPC-Based Green Time Optimization and Offset Adjustment Without a Separate Transition Cycle

The green-time module uses the MMPC framework from our previous work [8]. That framework introduced a ring-barrier phase structure, link-level flow distribution ratios, and an entering-vehicle participation model for green-time optimization. These components are retained here. The present study specifies how the background cycle length and offset corrections selected by SAC enter the coordinated MMPC timing plan.

In its earlier application, MMPC optimized green times for a specified cycle length. Here, it receives a new background cycle length from SAC and an offset-related cycle adjustment for the first prediction cycle. The MMPC calculation retains the existing phase and green-time constraints.

The offset reference point (OR_i) is the time at which the coordinated phase green starts; the offset is measured relative to the system time reference point (SR) [11,43]. Superscripts old and new denote consecutive implemented reference points, distinguishing their interval from the decision cycle. The new plan uses the SAC action ($A_{n+1|n}$) selected at the current decision; no later action is required. At intersection i, the offsets before and after the plan change are

$$\begin{aligned} O_{i,\mathrm{V2}}^{\mathrm{old}} &= BO_{i,\mathrm{V2}} + \Delta O_{i,\mathrm{V2}}^{\mathrm{old}} = OR_i^{\mathrm{old}} - SR^{\mathrm{old}}, \\ O_{i,\mathrm{V2}}^{\mathrm{new}} &= BO_{i,\mathrm{V2}} + \Delta O_{i,\mathrm{V2}}^{\mathrm{new}} = OR_i^{\mathrm{new}} - SR^{\mathrm{new}}. \end{aligned} \tag{7}$$

The subscript V2 denotes the coordinated phase in the tested phase structure; it is omitted in the action definition and notation table for brevity. Across these consecutive reference points, the base offset remains unchanged while SAC updates the offset correction.

For brevity, write $C^{\mathrm{bg}} = C^{\mathrm{bg}}(n+1)$ for the background cycle length in $A_{n+1|n}$, and let a_i denote the cycle adjustment supplied to MMPC. The interval between implemented reference points is

$$OR_i^{\mathrm{new}} = OR_i^{\mathrm{old}} + C^{\mathrm{bg}} + a_i. \tag{8}$$

The implemented cycle length at intersection i over this interval is therefore $C_i^{\mathrm{impl}} = C^{\mathrm{bg}} + a_i$. The 180 s action bound applies to C^{bg}; it does not automatically constrain this sum.

Subtracting the offset equations at the old and new reference points gives

$$\begin{aligned} a_i = &[\Delta O_{i,\mathrm{V2}}^{\mathrm{new}} - \Delta O_{i,\mathrm{V2}}^{\mathrm{old}}] \\ &+ [SR^{\mathrm{new}} - SR^{\mathrm{old}}] - C^{\mathrm{bg}}. \end{aligned} \tag{9}$$

At the first intersection, $\Delta O_1^{\mathrm{old}} = \Delta O_1^{\mathrm{new}} = 0$. Define the raw change in offset correction (d_i), its minimum (m), and the non-negative cycle adjustment (a_i) as follows:

$$d_i = \Delta O_{i,\mathrm{V2}}^{\mathrm{new}} - \Delta O_{i,\mathrm{V2}}^{\mathrm{old}}, \tag{10}$$

Since $d_1 = 0$, the minimum m is nonpositive. All raw changes receive the same shift:

$$m = \min_{1 \le \ell \le I} d_\ell, \qquad a_i = d_i - m \ge 0. \tag{11}$$

With the first intersection as the system reference, its implemented cycle length is $C^{\mathrm{bg}} - m$. Hence, $SR^{\mathrm{new}} - SR^{\mathrm{old}} = C^{\mathrm{bg}} - m$, and substitution into the preceding relation gives $a_i = d_i - m$. The common shift preserves relative changes in offset correction and keeps every a_i non-negative, avoiding local cycle shortening solely for offset adjustment. MMPC incorporates these adjustments into its existing green-time optimization and timing constraints, without a separately scheduled explicit transition cycle. Non-negativity alone does not establish feasibility under every possible timing limit.

The phase structure, minimum-green settings, and clearance durations used in this study give a minimum cycle of 60 s. Since the background-cycle action satisfies $C^{\mathrm{bg}} \ge 60$ s and $a_i \ge 0$, the implemented cycle satisfies

$$C_i^{\mathrm{impl}} = C^{\mathrm{bg}} + a_i \ge C^{\mathrm{bg}} \ge 60 \text{ s}. \tag{12}$$

Within this cycle, MMPC imposes $G_{ij}(n) \geq G_{ij}^{\min}$ and retains the ring–barrier constraints on the prescribed phase sequence, equal durations of parallel rings between barriers, and synchronized barrier crossings.

4. Simulation Setup and Experimental Evaluation

The evaluation uses a three-intersection arterial in microscopic simulation. Experiment 1 selects the phase-weight setting and a stored policy. Experiment 2 compares four timing configurations, and Experiment 3 compares the complete configuration with five selected timing-control baselines. The evaluation is limited to the specified geometry and demand range; it does not represent a field deployment or a larger network.

The simulation environment was built in PTV Vissim 6.00-19. All control algorithms were implemented in Python 3.7.16 through the Vissim COM interface, and the deep learning models were developed using PyTorch 1.13.1. All experiments were executed on computers equipped with an Intel 3.6 GHz CPU, 32 GB RAM, and an Nvidia GeForce RTX 2080Ti GPU.

The Wiedemann 74 car-following model uses an average standstill distance of 2.0 m, an additive safety-distance term of 3.3, and a multiplicative safety-distance term of 4.3. These car-following parameters were calibrated against a target saturation flow rate of 1900 veh/h/lane. Desired free-flow speeds are uniformly distributed over 48–58 km/h. The lane-change waiting time before vehicle removal is 45 s. Other car-following and lane-changing parameters retain Vissim defaults.

4.1. Arterial Testbed and Sensing Assumptions

The arterial has three adjacent signalized intersections, numbered from west to east, with the shared geometry shown in Figure 3. Stop-line spacings are 400 m between intersections 1 and 2 and 300 m between intersections 2 and 3. No additional access points are modeled along the arterial. The posted speed limit is 50 km/h.

To reflect practical roadside data collection conditions, the simulation assumes radar observation extending 130 m upstream of each stop line. Vehicle counts and density features from this observation window form part of the control state. The study evaluates coordinated timing-control performance under this prescribed sensing setup. This sensing assumption does not establish the reliability of field detectors.

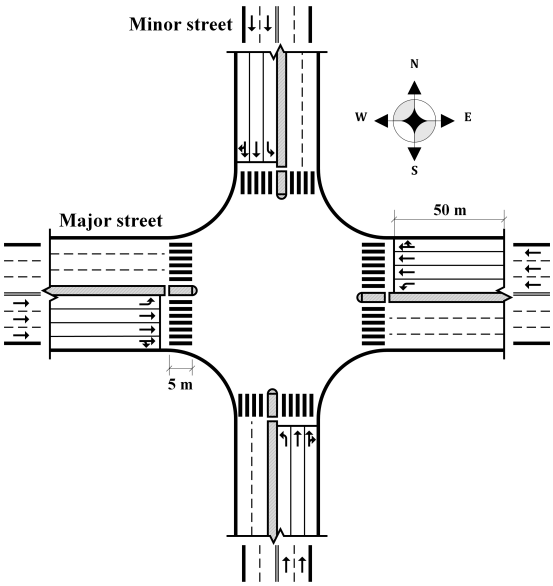


Figure 3. Shared intersection layout in the simulated three-intersection arterial. Lane-use arrows indicate the permitted movements for each lane.

4.2. Demand Scenario Generation

Traffic scenarios were generated by sampling per-lane vehicle inputs from 300 to 600 veh/h, giving total arterial inputs from 5400 to 10,800 veh/h. For result analysis, the generated scenarios were classified as low (5400–7200 veh/h), medium (7200–9000 veh/h), and high (9000–10,800 veh/h).

Rejection sampling retained scenarios that met the total-demand and directional-dominance rules. These rules determined which scenarios entered the evaluations, but they do not establish an exactly uniform distribution of the retained total demands.

(1) Total-demand rule: The lane-weighted average of the sampled per-lane input factors differs from the system demand reference by no more than 0.05 on the normalized sampling scale.
(2) Directional-dominance rule: Before conversion to integer vehicle inputs, the sampled per-lane input factor for the west leg of intersection 1 is the largest, and that for the east leg of intersection 3 is the second largest.

Samples that failed either rule were discarded and replaced. The accepted factors were then converted to integer directional vehicle inputs within the stated demand range. Rounding can produce equal inputs on the two arterial entry legs.

Pedestrian volumes and left- and right-turn ratios were generated using simple random sampling within their respective prescribed bounds. Pedestrian demand ranged from 100 to 400 ped/h, with pedestrians interacting with right-turning vehicles. Pedestrian service durations and pedestrian–vehicle intergreen intervals were fixed, without pedestrian actuation. Demand variation therefore affected simulated interactions without directly changing the phase structure, minimum-green bounds, clearance intervals, or MMPC timing constraints. Left-turn ratios ranged from 10 to 20% on the major street and 15 to 25% on the minor street, while right-turn ratios ranged from 5 to 10% on the major street and 10 to 15% on the minor street.

Random seeds controlled demand sampling and the randomization settings of the microscopic simulation. Experiments 2 and 3 used all integer seeds from 5000 through 5999, each corresponding to one demand scenario. Vehicle demand, turning ratios, and pedestrian volumes were identical across methods for a given seed. Each method was run once in Vissim for each demand scenario.

For each scenario, system vehicle delay is the system-level mean vehicle-delay measurement. Coordinated-direction space-mean speed is the total recorded travel distance divided by the total recorded travel time over two intervals bounded by the west-approach stop lines at intersections 1–2 and 2–3, expressed in km/h. Total vehicle travel time is the sum of vehicle travel times recorded during the measurement period. The tables report arithmetic means and standard deviations across scenarios; the standard deviation uses N as the divisor, where N is the number of scenarios. These standard deviations reflect variation in demand inputs as well as microscopic-simulation randomness.

Experiments 2 and 3 use scenario-matched, two-sided paired t tests to compare ECSTO with each alternative. Pointwise 95% confidence intervals are reported for the mean paired difference ($\bar{d}$). Effect sizes use $d_z = \bar{d}/s_d$, where s_d is the sample SD of paired differences. Holm adjustment covers all 21 comparisons ($\alpha = 0.05$). For the non-normal paired differences, t inference uses a large-sample approximation assuming independent scenarios. Paired bias-corrected and accelerated bootstrap intervals agreed on the direction of the mean differences and whether the intervals included zero.

4.3. Baseline Methods

4.3.1. Ablation Baselines

ECSTO-G retains MMPC green-time optimization but uses a conventional cycle length and base offsets. ECSTO-C retains SAC cycle length selection but uses conventional green-time allocation and base offsets. ECSTO-CG retains SAC cycle length selection and MMPC green-time optimization but uses base offsets alone. The complete ECSTO configuration also includes SAC offset correction. The suffixes C, G, and O denote SAC cycle selection, MMPC green-time optimization, and SAC offset correction, respectively. Some comparisons replace more than one mechanism, so the observed differences do not isolate the causal effects of individual modules.

4.3.2. Comparative Baselines

Fixed-time control (FC) and actuated control (AC) denote the two variants of the Webster and Vistro baselines. The five comparative baselines are Webster-FC, Webster-AC, Vistro-FC, Vistro-AC, and traffic-responsive control (TC). Within the explicit arterial-coordination framework, Webster represents classical timing design, Vistro represents commercial signal-timing optimization, and TC represents advanced adaptive control. All use the reported arterial geometry, demand-generation procedure and assumed 130 m roadside detection coverage.

Webster-FC and Webster-AC use the Webster cycle-length calculation, green-time allocation based on the assignable cycle length [19], and base offsets from the time–space relationship. Webster-AC also uses actuated logic with fixed force-off points.

Vistro-FC and Vistro-AC use Vistro timing plans for cycle length, green times, and offsets. Demand determines which plan is selected from the existing set. Vistro-AC also uses the actuated logic applied in Webster-AC.

TC [18] generates timing plans from cycle groups. Its implementation estimates green-time demand by exponentially smoothing detector data and signal status. The cycle length is based on estimated major- and minor-street demands, and base offsets are obtained from the time–space relationship.

4.4. Results and Analysis

4.4.1. Experiment 1: Reward Coefficient Tuning and Policy Selection

The first stage compares six phase-weight settings for the discharged-vehicle reward. The coefficient for uncoordinated phases, β_{NoCoord}, is fixed at 1.0; β_{Coord} ranges from 1.0 to 3.5, as listed in Table 3. A training episode denotes one complete simulation run used for learning. For comparisons among phase-weight settings, the saved policy is evaluated after every 10 training episodes using the same fixed set of 40 random seeds, with one run per demand scenario and greedy action selection. Network parameters remain fixed during evaluation. The evaluation uses outcomes from 900 to 4500 s.

Policy evaluation seeks lower system vehicle delay (D) and higher coordinated-direction space-mean speed (V). Let S_k denote the nondominated set for coefficient combination k, and let R denote the reference nondominated set constructed from the pooled candidate sets. This empirical reference front is not a known global optimum. Generational distance (GD) is the average Euclidean distance from each point in S_k to its nearest point in R. Inverted generational distance (IGD) averages the distance in the reverse direction, from each point in R to its nearest point in S_k. Both distances are calculated in the original (D, V) coordinates without normalization.

Hypervolume (HV) is the union area of the dominated rectangles in the delay–speed plane. It uses the reference point $(1.05 \max D,\ 0.95 \min V)$ from the pooled sets (S_k), with delay minimized and speed maximized. Combination 5 ($\beta_{\mathrm{Coord}} = 3.0$) has the lowest GD

and IGD and the highest HV among the six settings in Table 3. For policy selection within this setting, saved candidate policies are evaluated after every five training episodes using the same 40 random seeds and evaluation procedure.

Table 3. Two-objective policy-selection metrics for the six phase-weight settings, computed without normalization. Lower GD and IGD and higher HV are preferred.

Id	Coefficient Combination		Performance		
	β_{Coord}	β_{NoCoord}	GD	IGD	HV
1	1.0	1.0	0.50	2.99	38.00
2	1.5	1.0	1.24	0.82	40.79
3	2.0	1.0	0.89	1.26	41.61
4	2.5	1.0	0.50	2.80	37.53
5	3.0	1.0	0.11	0.32	46.88
6	3.5	1.0	1.75	2.24	43.40

Figure 4 shows the checkpoint evaluation histories for this setting through episode 700, corresponding to 26,921 completed action steps. After substantial early fluctuations, delay decreases and speed increases. Both metrics fluctuate less later in training. These observations describe empirical stabilization of traffic performance; they do not establish strict convergence.

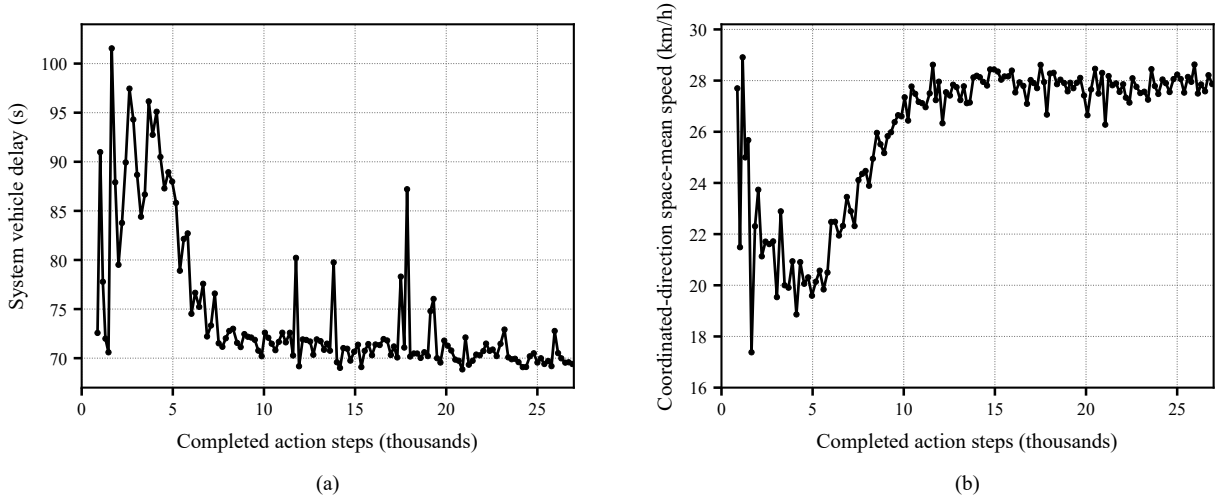


Figure 4. Training performance for combination 5: (**a**) system vehicle delay; (**b**) coordinated-direction space-mean speed.

With this coefficient setting, the evaluated policies are ranked on a common relative scale using utility quantiles. For each metric, the empirical quantile is $(r-1)/(M-1)$, where r is its ascending rank among the $M = 135$ evaluated policies. Delay utility is one minus the empirical delay quantile, so lower delay gives higher utility. Speed utility is the empirical speed quantile, so higher speed gives higher utility. Each utility quantile indicates a candidate's rank among evaluated policies, not a percentage improvement over a baseline.

Figure 5 shows the evaluated policies in the delay–speed plane. The inset compares the delay and speed utility quantiles and their sums for policies on the obtained Pareto front. The policy at training step 20,884 has the highest sum of utility quantiles: 100.00% for delay plus 91.79% for speed equals 191.79%. The sum is a selection score, not a 191.79% performance gain. In this evaluation, the policy had a system vehicle delay of 68.86 s and a coordinated-direction space-mean speed of 28.30 km/h. Its stored network parameters were then used in the separate ablation and baseline comparisons.

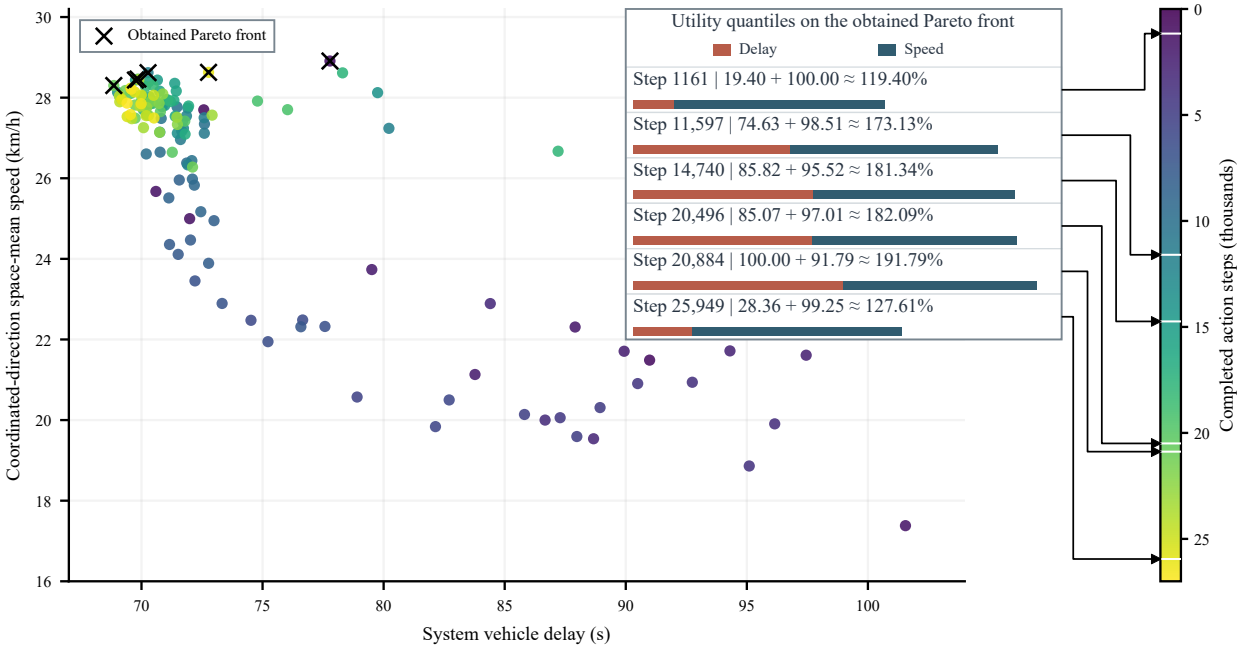


Figure 5. Evaluated policies and utility quantiles of the obtained Pareto front.

4.4.2. Experiment 2: Ablation Evaluation

This experiment compares the complete ECSTO configuration with three simplified configurations under the same 1000 demand scenarios. Table 4 lists the source of each timing parameter. ECSTO-CO is not reported because the conventional green-time allocation in ECSTO-C does not implement the cycle-by-cycle offset adjustment considered here. These comparisons do not isolate each module's independent causal effect.

Table 4. Sources of cycle length, phase green times, and offsets in the four tested configurations.

Configuration	Cycle Length	Green Time	Offset
ECSTO	SAC algorithm	MMPC method	Time–space relationship SAC algorithm
ECSTO-G	Webster method Critical intersection method	MMPC method	Time–space relationship
ECSTO-C	SAC algorithm	Green time allocation based on assignable cycle length [19]	Time–space relationship
ECSTO-CG	SAC algorithm	MMPC method	Time–space relationship

Table 5 reports system vehicle delay across the test scenarios. ECSTO has the lowest mean among the four configurations, at 70.51 s, compared with 82.07 s for ECSTO-G, 71.52 s for ECSTO-C, and 71.07 s for ECSTO-CG. The reported relative mean reductions are 14.09%, 1.42%, and 0.78%, respectively. The reductions relative to ECSTO-G and ECSTO-C remain significant after Holm adjustment ($p < 0.001$ and $p = 0.031$), with $d_z = 0.916$ and 0.077, respectively. The 0.56 s reduction relative to ECSTO-CG is not statistically significant (95% CI: $[-0.07, 1.18]$ s; Holm-adjusted $p = 0.082$).

Table 5. Descriptive statistics and paired comparisons for system vehicle delay across the four tested configurations (s).

Measure	Timing Configuration			
	ECSTO-G	ECSTO-C	ECSTO-CG	ECSTO
Median (s)	54.84	53.01	46.37	45.66
Mean (s)	82.07	71.52	71.07	70.51
SD (s)	53.73	42.30	48.73	49.11
$\bar{d}$ (s)	11.56	1.02	0.56	/
95% t CI	[10.78, 12.35]	[0.19, 1.84]	[−0.07, 1.18]	/
Holm-adjusted p	<0.001	0.031	0.082	/
Effect size d_z	0.916	0.077	0.055	/

$\bar{d}$: baseline minus ECSTO; positive values favor ECSTO.

Table 6 reports coordinated-direction space-mean speed. ECSTO has the highest mean among the four configurations, at 28.36 km/h, compared with 27.34, 25.14, and 25.15 km/h for ECSTO-G, ECSTO-C, and ECSTO-CG. The mean differences relative to ECSTO are 1.02, 3.22, and 3.21 km/h. All three speed increases have positive 95% confidence intervals and remain significant after Holm adjustment (all $p < 0.001$), with d_z ranging from 0.284 to 0.795. ECSTO-C and ECSTO-CG have nearly identical means, so their comparison does not support attributing an independent speed gain to the green-time module.

Table 6. Descriptive statistics and paired comparisons for coordinated-direction space-mean speed across the four tested configurations (km/h).

Measure	Timing Configuration			
	ECSTO-G	ECSTO-C	ECSTO-CG	ECSTO
Median (km/h)	29.20	27.67	28.98	30.50
Mean (km/h)	27.34	25.14	25.15	28.36
SD (km/h)	7.13	8.65	7.95	7.60
$\bar{d}$ (km/h)	1.02	3.22	3.21	/
95% t CI	[0.80, 1.25]	[2.88, 3.56]	[2.96, 3.46]	/
Holm-adjusted p	<0.001	<0.001	<0.001	/
Effect size d_z	0.284	0.580	0.795	/

$\bar{d}$: ECSTO minus baseline; positive values favor ECSTO.

4.4.3. Experiment 3: Comparative Evaluation

This experiment compares ECSTO with five selected explicit timing-control baselines in the same three-intersection simulation, using the same 1000 demand scenarios as the ablation experiment. The results describe these scenarios and their aggregate statistics.

Figure 6 groups system vehicle delay by demand. The differences between method means are smaller in the plotted low-demand bins and more visible in several medium-demand bins. At high demand, the differences vary by bin. The FC/AC pairs show that the ordering of fixed-time and actuated variants also varies across bins. These are descriptive observations; the error bars show dispersion across scenarios, not statistical significance.

Table 7 summarizes system vehicle delay across scenarios. ECSTO has the lowest mean, at 70.51 s, compared with 75.95–79.88 s for the five selected baselines. The reported relative mean reductions range from 7.16% to 11.73%. ECSTO also has the lowest median and across-scenario standard deviation in the table. The paired mean reductions range from 5.44 to 9.37 s, with positive 95% confidence intervals, Holm-adjusted $p < 0.001$ for all five comparisons, and d_z ranging from 0.484 to 0.724.

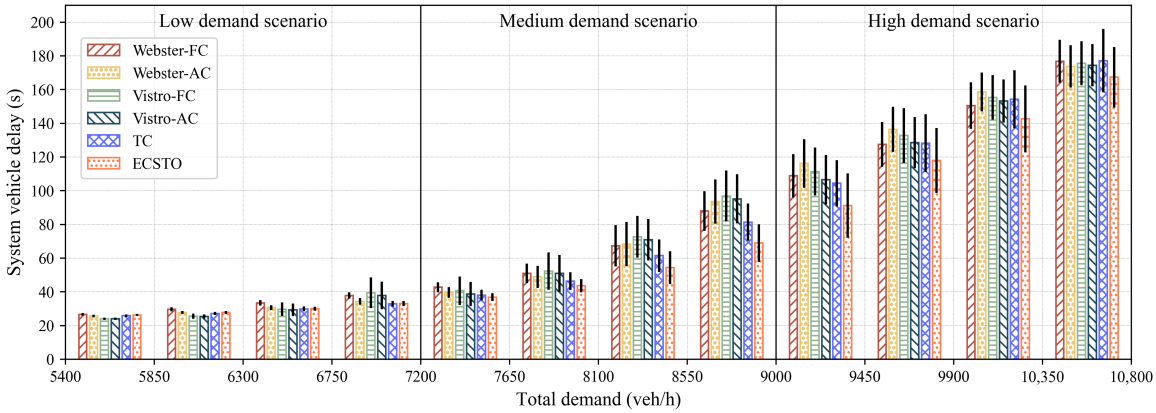


Figure 6. Mean system vehicle delay for the five baselines and ECSTO by total-demand bin (veh/h). Error bars show the across-scenario standard deviation within each bin (s).

Table 7. Descriptive statistics and paired comparisons for system vehicle delay across the evaluated demand scenarios (s).

Measure	Control Method					
	Webster		Vistro			
	FC	AC	FC	AC	TC	ECSTO
Median (s)	53.83	52.79	58.72	57.12	50.06	45.66
Mean (s)	78.65	79.55	79.88	78.14	75.95	70.51
SD (s)	50.90	53.63	53.13	52.41	53.15	49.11
$\bar{d}$ (s)	8.15	9.04	9.37	7.63	5.44	/
95% t CI	[7.45, 8.84]	[8.22, 9.86]	[8.48, 10.26]	[6.81, 8.46]	[4.74, 6.13]	/
Holm-adjusted p	<0.001	<0.001	<0.001	<0.001	<0.001	/
Effect size (d_z)	0.724	0.680	0.654	0.572	0.484	/

$\bar{d}$: baseline minus ECSTO; positive values favor ECSTO.

Table 8 reports total vehicle travel time, summed over vehicles during the measurement period. ECSTO has the lowest aggregate mean among the six methods, at 1.93×10^6 s. The five baseline means range from 2.04×10^6 to 2.15×10^6 s. The relative mean reductions, calculated from the unrounded scenario means, range from 5.30% to 10.19%. ECSTO also has the lowest median and across-scenario standard deviation in this table. All five paired mean reductions have positive 95% confidence intervals and remain significant after Holm adjustment (all $p < 0.001$), with d_z ranging from 0.546 to 0.790. The standard deviations describe dispersion across the evaluated scenarios; they do not establish temporal stability or field robustness.

Table 8. Descriptive statistics and paired comparisons for total vehicle travel time across the evaluated demand scenarios (10^6 s).

Measure	Control Method					
	Webster		Vistro			
	FC	AC	FC	AC	TC	ECSTO
Median ($\times 10^6$s)	1.49	1.49	1.58	1.55	1.44	1.38
Mean ($\times 10^6$s)	2.06	2.15	2.14	2.10	2.04	1.93
SD ($\times 10^6$s)	1.24	1.36	1.31	1.29	1.28	1.19
$\bar{d}$ (10^6 s)	0.1339	0.2188	0.2081	0.1737	0.1079	/
95% t CI	[0.1222, 0.1455]	[0.2017, 0.2360]	[0.1906, 0.2256]	[0.1577, 0.1897]	[0.0957, 0.1202]	/
Holm-adjusted p	<0.001	<0.001	<0.001	<0.001	<0.001	/
Effect size d_z	0.713	0.790	0.737	0.673	0.546	/

$\bar{d}$: baseline minus ECSTO; positive values favor ECSTO.

Table 9 gives a different ordering for coordinated-direction space-mean speed. ECSTO has a mean of 28.36 km/h, above Webster-FC (26.36 km/h) but below Webster-AC, Vistro-FC, Vistro-AC, and TC (31.25–33.90 km/h). All five speed differences have 95% confidence intervals excluding zero and Holm-adjusted $p < 0.001$; d_z is 0.344 relative to Webster-FC and ranges from −1.039 to −0.463 relative to the other four baselines. Its across-scenario standard deviation of 7.60 km/h is the lowest among the six methods. Coordinated movements and other conflicting movements share the available signal service, so a trade-off may arise between system-wide efficiency and coordinated-direction progression. The phase-weighted discharged-vehicle reward gives relative preference to coordinated movements while also considering service to other movements. Final policy selection considers both system delay and coordinated-direction speed. The system-level gains and lower coordinated-direction speed relative to four baselines illustrate this performance trade-off.

Table 9. Descriptive statistics and paired comparisons for coordinated-direction space-mean speed across the evaluated demand scenarios (km/h).

Measure	Control Method					
	Webster		Vistro			
	FC	AC	FC	AC	TC	ECSTO
Median (km/h)	29.51	35.90	33.11	31.46	35.16	30.50
Mean (km/h)	26.36	33.90	31.60	31.25	32.18	28.36
SD (km/h)	9.38	8.52	10.53	10.60	10.09	7.60
$\bar{d}$ (km/h)	2.00	−5.54	−3.24	−2.88	−3.82	/
95% t CI	[1.64, 2.36]	[−5.87, −5.21]	[−3.65, −2.83]	[−3.27, −2.50]	[−4.19, −3.44]	/
Holm-adjusted p	<0.001	<0.001	<0.001	<0.001	<0.001	/
Effect size d_z	0.344	−1.039	−0.489	−0.463	−0.637	/

$\bar{d}$: ECSTO minus baseline; positive values favor ECSTO.

5. Conclusions

This study extended SAC-based cycle-length selection to include offset corrections and connected these decisions to the existing MMPC green-time optimization within a modular ECSTO framework. A time–space relationship provides base offsets. The method translates changes between consecutive offset corrections into non-negative intersection-specific cycle adjustments supplied to MMPC. This mechanism links learned cycle-length and offset decisions to phase green-time optimization within the same cycle-level control process, without a separately scheduled explicit transition cycle.

In the evaluated three-intersection simulation, ECSTO had lower aggregate mean system vehicle delay (70.51 s) and total vehicle travel time (1.93×10^6 s) than the five selected timing-control baselines. These reductions remained statistically significant after Holm adjustment. For coordinated-direction space-mean speed, its mean of 28.36 km/h was significantly higher than Webster-FC and lower than the other four baselines. Its across-scenario speed standard deviation of 7.60 km/h was the lowest among the six methods. Among the four ablation configurations, ECSTO had the lowest mean delay and highest mean speed. However, its delay difference relative to ECSTO-CG was not statistically significant. These configuration-level differences do not isolate the independent causal contribution of each timing module.

ECSTO is a candidate optimization layer for existing arterial coordination. It is formulated for finite roadside radar observations and NEMA-style dual-ring-and-barrier phase organization. Cycle-level updates allow more time for computation and plan transmission than second-by-second decisions, subject to implementation deadlines. A plausible application is a short arterial with varying demand and a clear coordination direction, where both system delay and progression are control objectives. Deployment requires adequate

traffic features, synchronization and timing-update interfaces. Costs include detection, communications, computation, integration and maintenance. Where existing facilities suffice, additional investment may focus on software, computing and commissioning; otherwise, equipment upgrades may be needed.

The empirical findings are limited to the evaluated policies, tested demand scenarios, baseline implementations, and three-intersection arterial. They do not establish field deployability or performance in heterogeneous or larger networks. Restricting adjacent offset increments to non-negative values may exclude useful progression patterns under different directional traffic conditions. Performance has not been compared with an otherwise matched domain allowing signed increments. Training repeatability and sensitivity to training seeds have not been quantified through independent retraining under a fixed configuration. We plan to evaluate ECSTO across additional arterial geometries, numbers of intersections and demand conditions. Comparisons with other advanced controllers will use clearly specified conditions. These evaluations will distinguish transfer of a fixed policy from performance after retraining or adaptation. Alternative base-offset calculations also warrant investigation.

Author Contributions: Conceptualization, K.Z.; methodology, H.X. and H.H.; software, Q.Z.; validation, H.Z. and J.L.; formal analysis, Q.Z.; investigation, H.Z. and J.L.; resources, K.Z.; data curation, K.Z.; writing (original draft preparation), K.Z. and H.X.; writing (review and editing), H.H., H.Z., J.L. and Q.Z.; supervision, H.H. and H.X.; funding acquisition, K.Z. and H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China (grant number 2023YFB2504700); the Changsha Major Science and Technology Projects, China (grant number kh2401002); the Ministry of Education Foundation for Humanities and Social Sciences (grant number 24YJCZH460); and the Natural Science Foundation of Hunan Province, China (grant number 2024JJ7624).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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

Conflicts of Interest: Q.Z. is employed by Neusoft (China). Kun Zhang, Helai Huang, Hanchu Zhou, and Junwu Liu were affiliated with the company Changsha Communications Investment Holding Group Co., Ltd. The authors declare no other conflicts of interest.

References

1. Zhao, H.; Dong, C.; Cao, J.; Chen, Q. A survey on deep reinforcement learning approaches for traffic signal control. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108100. [CrossRef]
2. Zheng, Q.; Xu, H.; Chen, J.; Zhang, K. Multi-dimensional double deep dynamic Q-network with aligned Q-fusion for dual-ring barrier traffic signal control. *Appl. Sci.* **2025**, *15*, 1118. [CrossRef]
3. Du, Y.; ShangGuan, W.; Chai, L. Traffic signal control in mixed traffic environment based on advance decision and reinforcement learning. *Transp. Saf. Environ.* **2022**, *4*, tdac027. [CrossRef]
4. Zhu, Y.; Yin, X.; Chen, C. Extracting decision tree from trained deep reinforcement learning in traffic signal control. *IEEE Trans. Comput. Soc. Syst.* **2023**, *10*, 1997–2007. [CrossRef]
5. Hu, Y.; Du, L.; Easa, S.M. Explainable reinforcement learning for improved traffic signal control. *Comput.-Aided Civ. Infrastruct. Eng.* **2025**, *40*, 3911–3933. [CrossRef]
6. Cai, S.; Fang, J.; Xu, M. XLight: An interpretable multi-agent reinforcement learning approach for traffic signal control. *Expert Syst. Appl.* **2025**, *273*, 126938. [CrossRef]
7. Zhang, K.; Xu, H.; Pan, B.; Zheng, Q. Explicit coordinated signal control using soft actor–critic for cycle length determination. *IET Intell. Transp. Syst.* **2024**, *18*, 1396–1407. [CrossRef]

8. Zhang, K.; Xu, H.; Pan, B.; Zheng, Q.; Chen, H. Modified model predictive control for coordinated signals along an arterial under relaxing assumptions. *J. Adv. Transp.* **2024**, *2024*, 9967121. [CrossRef]
9. Gordon, R.L. *Traffic Signal Retiming Practices in the United States*; Transportation Research Board: Washington, DC, USA, 2010. [CrossRef]
10. Wang, S. A flexible model-free tram signal priority method with a large coordination scope in China. *Transp. Saf. Environ.* **2024**, *6*, tdad027. [CrossRef]
11. Urbanik, T.; Tanaka, A.; Lozner, B.; Lindstrom, E.; Lee, K.; Quayle, S.; Beaird, S.; Tsoi, S.; Ryus, P.; Gettman, D.; et al. *Signal Timing Manual*, 2nd ed.; Transportation Research Board: Washington, DC, USA, 2015. [CrossRef]
12. Wang, Z.; Wang, X.; Jerome, Z.; Liu, H.X. Improving coordinated actuated control performance by measuring "early return to green" using probe vehicle data. *Transp. Res. Part C Emerg. Technol.* **2026**, *191*, 105853. [CrossRef]
13. Stevanovic, A. *Adaptive Traffic Control Systems: Domestic and Foreign State of Practice*; National Academies Press: Washington, DC, USA, 2010. [CrossRef]
14. Long, G.; Wang, A.; Jiang, T. Traffic signal self-organizing control with road capacity constraints. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 18502–18511. [CrossRef]
15. Wang, H.; Zhu, M.; Hong, W.; Wang, C.; Tao, G.; Wang, Y. Optimizing signal timing control for large urban traffic networks using an adaptive linear quadratic regulator control strategy. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 333–343. [CrossRef]
16. Yang, H.; Guo, B.; Yan, C.; Chen, Z.; Wang, P. A new mobility field and gradient-based traffic signal control approach applicable to large-scale road networks. *Transp. Saf. Environ.* **2025**, *7*, tdae024. [CrossRef]
17. Zhang, L.; Zhang, L. Distributed Traffic Signal Optimization at V2X Intersections. *Mathematics* **2024**, *12*, 773. [CrossRef]
18. Xu, H.; Zhang, K.; Zhang, D.; Zheng, Q. Traffic-responsive control technique for fully actuated coordinated signals. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 5460–5469. [CrossRef]
19. Ali, M.E.M.; Durdu, A.; Celtek, S.A.; Yilmaz, A. An adaptive method for traffic signal control based on fuzzy logic with Webster and modified Webster formula using SUMO traffic simulator. *IEEE Access* **2021**, *9*, 102985–102997. [CrossRef]
20. Elkari, B.; Ourabah, L.; Sekkat, H.; Chaibi, Y. Minimizing intersection delays: A novel fuzzy logic-based architecture for traffic signal control. *Int. J. Intell. Transp. Syst. Res.* **2024**, *22*, 542–560. [CrossRef]
21. Lin, H.; Han, Y.; Cai, W.; Jin, B. Traffic signal optimization based on fuzzy control and differential evolution algorithm. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 8555–8566. [CrossRef]
22. Chala, T.D.; Kóczy, L.T. Agent-Based Intelligent Fuzzy Traffic Signal Control System for Multiple Road Intersection Systems. *Mathematics* **2025**, *13*, 124. [CrossRef]
23. Gong, Y.; Riis, C.; Zhong, S.; Wang, T.; Rodrigues, F.; Azevedo, C.L.; Jiang, Y. Active learning-enhanced constrained Bayesian optimization for multi-modal and multi-objective traffic signal control. *Transp. Res. Part C Emerg. Technol.* **2026**, *192*, 105872. [CrossRef]
24. Han, Y.; Han, Z.; Ding, F.; Li, F.; Wang, H.; Wang, X. Field-tested signal controller to mitigate spillover using trajectory data. *Comput.-Aided Civ. Infrastruct. Eng.* **2024**, *39*, 1354–1369. [CrossRef]
25. Zoabi, R.; Haddad, J. Enhanced queue-based Max-Pressure traffic signal control. *Transp. Res. Part C Emerg. Technol.* **2026**, *186*, 105538. [CrossRef]
26. Kodama, N.; Harada, T.; Miyazaki, K. Traffic signal control system using deep reinforcement learning with emphasis on reinforcing successful experiences. *IEEE Access* **2022**, *10*, 128943–128950. [CrossRef]
27. Saiki, T.; Arai, S. Flexible traffic signal control via multi-objective reinforcement learning. *IEEE Access* **2023**, *11*, 75875–75883. [CrossRef]
28. Zhang, G.; Chang, F.; Jin, J.; Yang, F.; Huang, H. Multi-objective deep reinforcement learning approach for adaptive traffic signal control system with concurrent optimization of safety, efficiency, and decarbonization at intersections. *Accid. Anal. Prev.* **2024**, *199*, 107451. [CrossRef]
29. Chu, T.; Wang, J.; Codecà, L.; Li, Z. Multi-agent deep reinforcement learning for large-scale traffic signal control. *IEEE Trans. Intell. Transp. Syst.* **2020**, *21*, 1086–1095. [CrossRef]
30. Li, Z.; Yu, H.; Zhang, G.; Dong, S.; Xu, C.Z. Network-wide traffic signal control optimization using a multi-agent deep reinforcement learning. *Transp. Res. Part C Emerg. Technol.* **2021**, *125*, 103059. [CrossRef]
31. Hu, T.; Hu, Z.; Lu, Z.; Wen, X. Dynamic traffic signal control using mean field multi-agent reinforcement learning in large-scale road networks. *IET Intell. Transp. Syst.* **2023**, *17*, 1715–1728. [CrossRef]
32. Wu, J.; Ran, Y.; Lou, Y.; Shu, L. CenLight: Centralized traffic grid signal optimization via action and state decomposition. *IET Intell. Transp. Syst.* **2023**, *17*, 1247–1261. [CrossRef]
33. Ćiprovski, D.; Ilić, N.; Božilović, B.; Vučetić, M. Fuzzy-Guided Exploration for Multi-Agent Reinforcement Learning in Traffic Signal Control. *Mathematics* **2026**, *14*, 1942. [CrossRef]
34. Huang, H.; Hu, Z.; Lu, Z.; Wen, X. Network-scale traffic signal control via multiagent reinforcement learning with deep spatiotemporal attentive network. *IEEE Trans. Cybern.* **2023**, *53*, 262–274. [CrossRef]

35. Liu, D.; Li, L. A traffic light control method based on multi-agent deep reinforcement learning algorithm. *Sci. Rep.* **2023**, *13*, 9396. [CrossRef]
36. Li, Q.; Zhuang, X.; Liu, L.; Chen, B.; Zhang, Y.; Zhao, Y. Single-Agent Reinforcement Learning Model for Adaptive Traffic Signal Control in Urban Corridors. *J. Adv. Transp.* **2026**, *2026*, 5134018. [CrossRef]
37. Kim, H.; Yeo, H. Hybrid approach for coordinated traffic signal control capturing offset-split interdependency using average phase statistics. *Transp. Res. Part C Emerg. Technol.* **2026**, *188*, 105718. [CrossRef]
38. Zheng, Q.; Xu, H.; Chen, J.; Zhang, D.; Zhang, K.; Tang, G. Double deep Q-network with dynamic bootstrapping for real-time isolated signal control: A traffic engineering perspective. *Appl. Sci.* **2022**, *12*, 8641. [CrossRef]
39. Haarnoja, T.; Zhou, A.; Hartikainen, K.; Tucker, G.; Ha, S.; Tan, J.; Kumar, V.; Zhu, H.; Gupta, A.; Abbeel, P.; et al. Soft actor-critic algorithms and applications. *arXiv* **2019**, arXiv:1812.05905. . [CrossRef]
40. Christodoulou, P. Soft actor-critic for discrete action settings. *arXiv* **2019**, arXiv:1910.07207. [CrossRef]
41. Xu, Y.; Hu, D.; Liang, L.; McAleer, S.; Abbeel, P.; Fox, R. Target entropy annealing for discrete soft actor-critic. *arXiv* **2021**, arXiv:2112.02852. [CrossRef]
42. OpenAI; Berner, C.; Brockman, G.; Chan, B.; Cheung, V.; Dębiak, P.; Dennison, C.; Farhi, D.; Fischer, Q.; Hashme, S.; et al. Dota 2 with large-scale deep reinforcement learning. *arXiv* **2019**, arXiv:1912.06680. [CrossRef]
43. Zheng, M.; Xu, H.; Zhang, K.; Yao, R. Shortest-way: An improved empirical transition method for signal coordination. *J. Adv. Transp.* **2017**, *2017*, 7670521. [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.