IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 13, No. 3, March, 2026, pp. 453-471
Journal homepage: http://iieta.org/journals/mmep

A Structurally Interpretable Hybrid Multi-Head TD3 Framework for Stable and Adaptive Traffic Signal Control

Wahid Chergui*, Saida Lehis, Abdelaali Bekhouche, Brahim Belgroun, Mohamed Boussalem

Department of Computer Science, ICOSI Laboratory, Abbes Laghrour University, Khenchela 40004, Algeria

Corresponding Author Email: Chergui.wahid@univ-khenchela.dz

https://doi.org/10.18280/mmep.130301

Received: 17 January 2026
Revised: 10 March 2026
Accepted: 20 March 2026
Available online: 10 April 2026

Keywords:
adaptive traffic signal control, hybrid action reinforcement learning, TD3, multi-head actor, attention-based policy fusion, interpretability, SUMO simulation, non-stationary traffic demand

ABSTRACT

Adaptive traffic signal control (TSC) under stochastic and non-stationary demand remains challenging because of the limited temporal resolution, learning instability, and poor interpretability of existing deep reinforcement learning (DRL) methods. This study proposes a structurally interpretable Hybrid Multi-Head TD3 (HMH-TD3) framework that jointly optimizes discrete signal phases and continuous green-time durations through a unified hybrid action formulation. The actor is decomposed into multiple behavior-specific heads inspired by self-organizing control principles, whereas an attention-based fusion mechanism enables state-dependent adaptation and provides diagnostic insight into the decision process. Training stability is ensured via TD3 enhancements, including twin critics, target policy smoothing, and explicit anti-oscillation constraints. The framework is evaluated in a high-fidelity Simulation of Urban Mobility (SUMO) environment under uniform, asymmetric, and stochastic traffic demand. Compared with fixed-time, vehicle-actuated, and DRL baselines, the proposed controller achieves 30–60% delay reduction, improved queue regulation, and 10–15% throughput gains in stable regimes while maintaining bounded saturations. Ablation studies confirm the complementary contributions of hybrid action modeling and multi-head attention to robustness and learning consistency.

1. INTRODUCTION

Urban traffic signal control (TSC) remains a core challenge in intelligent transportation systems because of its direct impact on congestion, delay, and network stability. While conventional approaches, such as fixed-time signal plans derived from historical surveys [1] and rule-based vehicle-actuated controllers [2], provide reliable operational baselines, they lack the adaptability required to cope with heterogeneous, stochastic, and rapidly evolving traffic conditions.

Recent advances in deep reinforcement learning (DRL) have demonstrated strong potential for adaptive TSC by learning policies directly from traffic observations [3, 4]. Both value-based and actor–critic methods have achieved measurable improvements in delay reduction and throughput optimization. However, several limitations hinder their practical deployment. First, many DRL-based controllers rely on purely discrete action spaces with predefined green durations, which restrict temporal flexibility. Second, most DRL policies are implemented as monolithic black-box networks, offering limited interpretability and reducing the trust of traffic engineers. Third, learning instability, caused by value overestimation, delayed rewards, and non-stationary traffic dynamics, often leads to oscillatory or unsafe control behavior, which is particularly problematic in safety-critical traffic systems [5-7].

Beyond these well-documented issues, a more fundamental limitation remains unaddressed. In most existing DRL formulations, a single policy must simultaneously resolve conflicting traffic objectives, such as rapid queue clearance, saturation avoidance, and oscillation suppression. Hybrid or parameterized-action approaches partially alleviate temporal rigidity by enabling continuous green-time duration; however, they retain monolithic policies that implicitly conflate these competing behaviors, often resulting in unstable or opaque decision-making. Conversely, multi-head or mixture-of-experts architectures offer policy diversity and robustness but are rarely integrated with hybrid traffic signal actions in a principled manner, particularly in continuous-control actor–critic frameworks.

In parallel, self-organizing traffic control paradigms have been explored as interpretable and robust alternatives based on local behavioral rules that yield coordinated global dynamics [8, 9]. Although transparent and resilient under stochastic demand, these approaches generally do not provide the optimization capability and long-horizon planning afforded by reinforcement learning (RL). While recent studies have explored hybrid rule–learning frameworks, the systematic integration of structured and interpretable behavioral principles into continuous-control DRL architectures remains limited. Unlike hierarchical or option-based RL methods, which introduce additional temporal abstraction and termination learning, this study focuses on behavioral decomposition directly at the action-selection level without increasing temporal complexity.

Rather than proposing a novel RL optimizer, this study

addresses a structural limitation of continuous-control formulations for TSC, namely, the absence of explicit behavioral decomposition at the action-selection level. To this end, we propose a Hybrid Multi-Head TD3 (HMH-TD3) framework for adaptive TSC. The proposed architecture integrates hybrid discrete–continuous action modeling with a multi-head actor structure, in which each head specializes in a distinct behavioral primitive inspired by self-organizing traffic logic. An attention-based fusion mechanism dynamically combines these primitives according to the current traffic context, thereby enabling improved control stability and enhanced structural interpretability.

The main contributions of this study are fourfold: (i) a hybrid action modeling scheme that jointly optimizes discrete signal phase selection and continuous green-time duration; (ii) a multi-head actor decomposition, in which individual heads specialize in complementary traffic behaviors, such as queue clearance and saturation balancing; (iii) a context-aware attention mechanism that coordinates behavioral primitives while providing diagnostic insight into the decision process; and (iv) the integration of stabilized learning mechanisms, including twin critics, target policy smoothing, and anti-oscillation constraints, to ensure convergence under stochastic traffic dynamics.

The proposed HMH-TD3 controller is validated through microscopic traffic simulations under uniform, asymmetric, and time-varying demand scenarios. Comparative experiments against fixed-time control (FTC), vehicle-actuated (VA) logic, and standard DRL baselines demonstrate consistent reductions in delay and congestion while maintaining stable signal operation. The interpretability provided by HMH-TD3 is structural and diagnostic in nature and does not claim a formal causal explanation.

The remainder of this study is organized as follows: Section 2 reviews the related work on adaptive TSC and structured RL architectures. Section 3 introduces the theoretical background. Section 4 formulates the traffic control problem. Section 5 describes the proposed HMH-TD3 framework. Section 6 presents the experimental setup and the results. Finally, Section 7 concludes the study and provides directions for future research.

2. RELATED WORK

Adaptive TSC has long been studied because of its central role in mitigating congestion and improving urban mobility. Early studies primarily followed a self-organizing control paradigm, modeling intersections as decentralized systems governed by local interaction rules. Seminal studies have demonstrated that traffic signals modeled as coupled nonlinear oscillators can achieve emergent synchronization and robust adaptation to fluctuating demand without centralized coordination [10, 11]. Subsequent rule-based extensions, including pressure-based and extension-based strategies, further improved scalability and resilience under stochastic traffic conditions [12, 13]. Despite their transparency and robustness, these approaches generally lack anticipatory optimization and long-horizon planning capabilities.

DRL has emerged as a powerful alternative, enabling adaptive signal control policies to be learned directly from traffic observations [14, 15]. A wide range of DRL-based controllers have been proposed for both isolated intersections and large-scale networks, demonstrating significant improvements in delay reduction and throughput optimization [3, 5, 7, 8]. However, most DRL-based TSC methods rely on monolithic policy representations, are computationally demanding, and offer limited interpretability, which constrains their adoption in safety-critical traffic environments.

To address network-level coordination, recent studies have incorporated attention mechanisms within multi-agent reinforcement learning (MARL) frameworks. Graph-based attention models capture spatial dependencies among neighboring intersections and improve coordination efficiency [16], whereas transformer-based and meta-learning approaches enhance adaptability under non-stationary demand [17-19]. Nevertheless, these methods primarily focus on spatial attention, that is, determining where information is aggregated, while offering limited insight into how individual control actions are internally generated at the agent level.

Interpretability remains a critical challenge in the deployment of DRL-based traffic controllers. More recent intrinsic approaches expose interpretable traffic metrics and parameterized timing decisions within the learning process [9, 20]. Early studies relied on post-hoc explanation techniques, such as surrogate models or decision tree extraction [21, 22], often at the expense of policy fidelity. Nevertheless, interpretability is still frequently treated as an auxiliary objective rather than a structural property of the controller architecture.

Beyond traffic applications, RL research has explored behavioral decomposition through modular policies, mixture-of-experts architectures, and option-based frameworks, demonstrating improved robustness and learning efficiency through specialization [23]. However, such approaches remain underexplored in TSC and are rarely combined with hybrid action formulations or attention mechanisms for structurally interpretable intra-agent decision-making. In contrast to hierarchical or option-based methods that introduce temporal abstraction and option termination learning, the present study performs behavioral decomposition directly at the action-selection level without adding temporal complexity. Moreover, unlike mixture-of-expert models that enforce specialization through auxiliary losses, specialization in the proposed framework emerges implicitly through traffic-induced constraints and attention-based arbitration.

While attention mechanisms have been widely used to model spatial dependencies in TSC, establishing a clear relationship between attention weights and observable traffic dynamics remains an open challenge. Existing studies often rely on descriptive visualizations, with limited analysis of how internal signals correspond to observable traffic phenomena such as queue buildup or congestion imbalance.

Recent work emphasizes the importance of linking internal model mechanisms to observable traffic conditions. In this context, attention-based architectures offer the potential to provide behaviorally grounded explanations by revealing how control decisions adapt to variations in traffic states. However, translating these mechanisms into actionable and reliable insights for real-world deployment remains a significant challenge.

More recently, research has increasingly focused on improving the deployment feasibility of DRL-based traffic signal controllers. In particular, lightweight DRL approaches have been proposed to reduce computational overhead through techniques such as model compression, parameter sharing, and efficient network design, enabling real-time inference in resource-constrained environments. In parallel, edge-

computing-based solutions deploy learning and inference closer to traffic infrastructure (e.g., roadside units or intersection controllers), reducing latency and communication costs [24]. Additionally, decentralized RL frameworks have demonstrated strong scalability and reduced computational bottlenecks in large-scale urban networks [25]. Nevertheless, recent surveys emphasize that, despite strong performance, DRL-based TSC still faces practical challenges related to computational complexity, scalability, and real-time constraints [1].

While these developments improve deployment feasibility, they often introduce a trade-off between model simplicity and control optimality, as lightweight designs may limit representational capacity and adaptability in complex traffic scenarios. This highlights the need for architectures that are both computationally efficient and behaviorally expressive, especially in safety-critical environments.

In this context, the proposed multi-head architecture offers a structured form of parameter sharing and modular computation, where multiple behavioral components are learned within a unified framework. This design improves computational efficiency while preserving expressive decision-making capabilities, making it suitable for real-time deployment scenarios.

Despite these advances, several gaps remain. First, self-organizing robustness has not been systematically integrated into continuous-control DRL frameworks. Second, attention mechanisms in TSC predominantly address spatial coordination rather than behavioral interpretability at the agent level. Third, most existing approaches rely on post-hoc explanations rather than architecturally embedded interpretability. To address these gaps, we propose a structurally interpretable HMH-TD3 framework, in which traffic control policies are explicitly decomposed into interpretable behavioral primitives inspired by self-organizing traffic logic and coordinated through attention-based fusion. The proposed framework jointly addresses performance and interpretability, with the attention mechanism providing diagnostic insight into how control strategies adapt to evolving traffic conditions.

3. BACKGROUND AND PRELIMINARIES

This section reviews the theoretical foundations underlying the proposed framework, including RL in Markov Decision Process (MDP), the TD3 algorithm for continuous control, parameterized hybrid action spaces, and multi-head actor architectures.

3.1 Reinforcement learning and Markov Decision Process

RL formulates sequential decision-making problems in dynamic and uncertain environments. Adaptive TSC naturally fits this paradigm, as control decisions must respond to continuously evolving traffic conditions.

An RL problem is commonly modeled as an MDP, defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ indicates the action space, $\mathcal{P}(s'|s,a)$ denotes the stochastic transition kernel, $\mathcal{R}(s,a)$ indicates the immediate reward, and $\gamma \in [0,1]$ denotes the discount factor. At each decision step t, the agent observes s_t, selects an action a_t according to its policy $\pi(a|s)$, receives a reward $\mathcal{R}_t$, and transitions to s_{t+1}.

The objective is to learn an optimal policy maximizing the expected discounted return:

$$\pi^* = \underset{\pi}{argmax}\, \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}_t\right] \quad (1)$$

Tabular RL methods are impractical because of the high dimensionality, stochasticity, and non-stationarity of urban traffic. DRL addresses this limitation by using neural networks to approximate value functions and policies, thereby enabling scalable learning in complex traffic environments [26].

3.2 Twin Delayed Deep Deterministic Policy Gradient

The Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm extends Deep Deterministic Policy Gradient (DDPG) to improve stability and convergence in continuous-control settings [27]. TD3 mitigates overestimation bias and learning instability through three mechanisms.

First, clipped double Q-learning employs two independent critics Q_1 and Q_2, with the target value computed as:

$$Q_{target} = r + \gamma min(Q_1'(s', a'), Q_2'(s', a')) \quad (2)$$

Second, delayed policy updates reduce variance by updating the actor less frequently than the critics. Third, target policy smoothing adds bounded Gaussian noise to the target action:

$$a' = clip(a + \epsilon, a_{min}, a_{max}), \epsilon \sim \mathcal{N}(0, \sigma^2) \quad (3)$$

These mechanisms jointly improve robustness and make TD3 well-suited for nonlinear and stochastic traffic control problems, as demonstrated in recent studies [9, 28].

3.3 Parameterized action space for hybrid traffic signal control

Adaptive TSC requires the selection of both a discrete signal phase and its continuous duration. This hybrid decision structure is modeled using a parameterized action MDP (PAMDP) [29], where the action at time *t* is defined as:

$$a_t = (\varphi_t, \Delta_t) \quad (4)$$

where, $\varphi_t \in \{1, \ldots, M\}$ denotes the selected phase and $\Delta_t \in \mathbb{R}^+$ indicates the associated green-time duration. Parameterized actions enable fine-grained temporal control without inflating the action space and reduce oscillatory behavior compared to purely discrete formulations [29, 30]. The concrete realization of this hybrid structure is described in Section 5.3.

While PAMDP provides a general framework for hybrid decision-making, existing hybrid-action methods such as PADDPG [31], related traffic formulations [9, 32], and hybrid PPO variants typically rely on end-to-end differentiable architectures, where discrete and continuous components are jointly optimized.

In TSC, such formulations may lead to unstable or physically inconsistent decisions, including rapid phase switching or infeasible phase sequences, as no explicit mechanism is designed to enforce operational constraints such as minimum green time or phase persistence.

Compared to these approaches, the proposed method

introduces three key structural differences: (i) a semi-differentiable optimization scheme restricting gradient updates to continuous parameters; (ii) a critic-driven discrete selection mechanism ensuring feasibility and operational validity; (iii) a multi-head policy with attention-based fusion enabling behavioral specialization (e.g., queue dissipation vs. flow balancing) and adaptive arbitration.

This design improves temporal stability and reduces oscillatory behavior under realistic traffic constraints. Unlike monolithic hybrid policies, it captures heterogeneous traffic patterns through specialized behavioral components.

These properties are consistent with the experimental results (Section 6), which show reduced phase-switching frequency and improved traffic stability compared to baseline methods.

Overall, the proposed framework constitutes a domain-specific structural extension of hybrid-action RL, designed to enhance stability, structural interpretability, and operational realism under dynamic and safety-critical traffic conditions.

3.4 Multi-head actor architectures

Multi-head actor architectures enhance robustness and exploration by decomposing the policy into multiple specialized sub-policies [33, 34]. A shared encoder feeds H actor heads, the outputs of which are combined through an attention-based weighted fusion:

$$a_t = \sum_{i=1}^{H} \alpha_i(s_t)\pi_i(s_t), \sum_{i=1}^{H} \alpha_i(s_t) = 1 \tag{5}$$

where, $\pi_i(s_t)$ denotes the output of the i-th actor head and $\alpha_i(s_t)$ denotes its corresponding attention weight. This weighted combination is performed in the latent action space before any discrete decision-making. The attention mechanism fuses continuous latent representations produced by the policy heads rather than averaging discrete phase indices, which would be physically meaningless. This design preserves semantic validity while enabling smooth coordination among heterogeneous behavioral primitives.

By allowing individual heads to specialize in complementary traffic control strategies (e.g., queue clearance, delay balancing, or platoon accommodation), the multi-head formulation improves policy robustness under non-stationary traffic conditions. Moreover, the attention weights $\alpha_i(s_t)$ provide a diagnostic signal indicating the dominant behavioral strategy activated for a given traffic state. This interpretation remains structural and does not imply causal attribution. The integration of this architectural principle within the proposed HMH-TD3 framework is detailed in Section 5.6.

4. PROBLEM DEFINITION

This section formalizes the adaptive TSC problem addressed in this study. We define the intersection model, admissible signal phases, hybrid control actions, and control objective used for RL.

4.1 Intersection model

We consider an isolated four-leg urban intersection regulated by an adaptive traffic signal controller. Traffic dynamics are simulated using the microscopic Simulation of Urban Mobility (SUMO) environment, which captures realistic car-following and lane-changing behaviors under stochastic demand conditions.

Let $\mathcal{L}$ denote the set of incoming lanes:

$$\mathcal{L} = \{l_1, l_2, ..l_{|\mathcal{L}|}\} \tag{6}$$

Each lane $l \in \mathcal{L}$ is characterized by time-varying traffic attributes, such as queue length, waiting time, occupancy, and discharge flow. These variables evolve according to stochastic and nonlinear traffic dynamics, which are driven by signal timing, vehicle interactions, and demand fluctuations. Virtual loop detectors provide full observability of these quantities at each control interval.

The traffic system is modeled as an MDP with state transitions governed by:

$$P(s_{t+1}|s_t, a_t) \tag{7}$$

capturing the inherent stochasticity and non-stationarity of urban traffic. An overview of the closed-loop traffic control architecture, including the intersection layout, sensing infrastructure, and hybrid decision process, is illustrated in Figure 1.

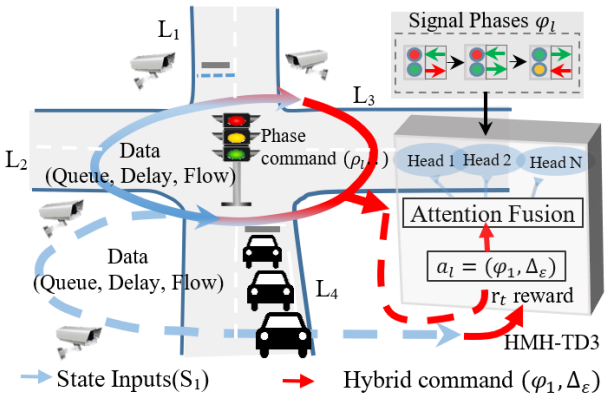


Figure 1. Closed-loop architecture of the proposed Hybrid Multi-Head TD3 (HMH-TD3) traffic signal control (TSC) framework

4.2 Signal phases and hybrid control actions

The traffic signal operates over a predefined set of admissible phases as follows:

$$\Phi = \{\varphi_1, \varphi_2, \ldots, \varphi_M\} \tag{8}$$

Each phase corresponds to a compatible set of traffic movements consistent with standard urban intersection design principles (Figure 2).

Following the PAMDP formulation introduced in Section 3.3, the control action at each decision epoch consists of selecting a discrete phase $\varphi_t \in \Phi$ and an associated continuous green-time duration. In SUMO, continuous duration decisions are operationalized through a green-hold mechanism, whereby the selected phase is maintained for a duration that is proportional to the executed command.

All phase transitions comply with standard safety constraints, including minimum green times, fixed yellow

intervals, and all-red clearance periods, ensuring a physically plausible and safe signal operation [9, 16]. The hybrid phase–duration action execution a_t is integrated within the closed-loop architecture shown in Figure 2.

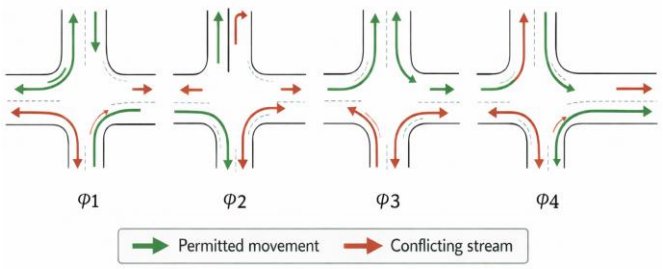


Figure 2. Protected signal phase configurations for the modeled intersection

4.3 Control objective and reward function

The objective of adaptive TSC is to optimize both the selection and duration of signal phases to improve intersection efficiency under time-varying demand. This involves maximizing vehicle discharge and mitigating congestion and excessive delays.

The immediate reward at time step *t* is defined as:

$$\mathcal{R}_t = w_1 N_{veh}(t) - w_2 Q_{len}(t) - w_3 W_{time}(t) \tag{9}$$

where, $N_{veh}(t)$ denotes the number of discharged vehicles, $Q_{len}(t)$ indicates the mean queue length across all incoming lanes, and $W_{time}(t)$ denotes the mean waiting time across all lanes at time *t*. For consistency, these quantities correspond to the aggregated variables $\bar{q}_t$ and $\bar{w}_t$ defined in Section 5.2.

This reward formulation promotes high throughput while penalizing queue accumulation and delays, yielding a smooth and well-conditioned optimization landscape [18, 35]. These properties are critical for stable learning in stochastic traffic environments and motivate the use of the hybrid TD3 framework introduced in Section 3. All quantities are computed from aggregated lane-level measurements as described in Section 5.2.

5. PROPOSED METHODOLOGY

The proposed methodology is based on an enhanced actor–critic RL framework tailored to a hybrid discrete–continuous TSC. The design targets learning stability, operational plausibility, and robustness under non-stationary traffic dynamics, while enabling fine-grained adaptive signal timing. The core architectural components and learning mechanisms are described in Section 5.6.

5.1 Simulation environment and interface

The learning environment is implemented in the microscopic traffic simulator SUMO, which provides a high-fidelity platform for emulating urban traffic dynamics [36]. Although the physical intersection layout is defined in Section 4.1, this section focuses on its integration within the RL loop.

Agent–environment interaction is realized via the traffic control interface (TraCI), which enables synchronous access to traffic state measurements and signal execution at each decision epoch. At each control step *t*, SUMO returns aggregated traffic indicators derived from the incoming lanes (as defined in Eq. (6)), including queue dynamics, throughput, occupancy, and delay-related measures obtained from virtual loop detectors.

The simulator captures nonlinear and time-varying traffic regimes ranging from free-flow to near-saturation, including platoon dispersion, shockwave propagation, and congestion spillback. The controller operates in a closed-loop perception–action cycle: The agent observes the current traffic state, outputs a hybrid control action, and SUMO applies the corresponding signal configuration before advancing the simulation. This synchronous interaction constitutes the operational backbone of the proposed HMH-TD3 controller [37-39].

5.2 Observation space

The observation space is constructed from the lane-level traffic variables defined in Section 4.1, where the set of incoming lanes is denoted by $\mathcal{L} = \{l_1, l_2, .. l_{|\mathcal{L}|}\}$. At each decision epoch t, the environment returns a state vector $s_t \in \mathbb{R}^d$ composed of aggregated macroscopic indicators:

$$s_t = [\bar{w}_t, \sigma_w(t), \bar{q}_t, \sigma_q(t), \bar{s}_t, s_{max}(t), \eta_t] \tag{10}$$

All components of s_t are obtained by aggregating lane-level measurements over the incoming-lane set $\mathcal{L}$, ensuring consistency with the intersection model introduced in Section 4.1. Each lane $l \in \mathcal{L}$ is associated with time-varying traffic attributes measured by SUMO detectors. Let $V_l(t)$ denote the set of vehicles present on lane l at time t, and let $N_l(t) = |V_l(t)|$ denote its cardinality.

The average waiting time on lane l is defined as:

$$w_l(t) = \frac{1}{N_l(t)} \sum_{i \in V_l(t)} w_i(t) \tag{11}$$

where, $w_i(t)$ represents the accumulated time during which vehicle i travels below a threshold speed. If $N_l(t) = 0$, we define $w_l(t) = 0$ to avoid undefined values. Similarly, the queue length on lane l is defined as the number of vehicles with speed below a threshold v_{th} (typically 0.1 m/s) $q_l(t) = |\{i \in V_l(t) : v_i(t) < v_{th}\}|$.

These lane-level quantities are then aggregated across all incoming lanes to form the macroscopic state representation.

The mean waiting time across lanes is defined as:

$$\bar{w}_t = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} w_l(t) \tag{12}$$

and its variability (population standard deviation across lanes) is given by:

$$\sigma_w(t) = \sqrt{\frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} (w_l(t) - \bar{w}_t)^2} \tag{13}$$

The mean queue length and its variability are defined as:

$$\bar{q}_t = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} q_l(t), \sigma_q(t) = \sqrt{\frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} (q_l(t) - \bar{q}_t)^2} \tag{14}$$

Following this, a normalized saturation ratio is defined for each lane as:

$$s_l(t) = \frac{q_l(t)}{q_l^{max}} \tag{15}$$

where, q_l^{max} denotes the maximum admissible queue capacity of lane l, typically derived from lane length and average vehicle spacing. This definition provides a proxy of congestion relative to the physical storage capacity of the lane, which is particularly relevant for capturing spillback effects in urban intersections.

The saturation-related quantities are then aggregated as follows:

$$\bar{s}_t = \frac{1}{|\mathcal{L}|}\sum_{l\in\mathcal{L}} s_l(t), s_{max}(t) = max_{l\in\mathcal{L}} s_l(t) \tag{16}$$

To improve robustness to sensing uncertainty, a Gaussian noise term $\eta_t \sim \mathcal{N}(0, \sigma_{noise}^2)$ is injected during training. This noise is disabled during evaluation. Each feature $x(t) \in s_t$ is normalized using a clipped z-score transformation as follows:

$$\hat{x}(t) = clip\left(\frac{x(t)-\mu_x}{s_x + 10^{-6}}, -c, c\right) \tag{17}$$

where, μ_x and s_x denote running estimates of the mean and standard deviation of feature *x*, and *c* is a clipping threshold. This normalization mitigates outliers under congested conditions and stabilizes gradient propagation during learning [4, 16, 27].

5.3 Hybrid action space

The controller operates in a parameterized hybrid action space that jointly models discrete phase selection and continuous green-time modulation. The hybrid action defined in Section 3.3, $a_t = (\varphi_t, \Delta_t)$, is parameterized through a latent actor output $\tilde{a}_t = (l_t, \mu_t)$, which is subsequently mapped to executable control variables. At each decision epoch *t*, the actor outputs a latent action:

$$\tilde{a}_t = (l_t, \mu_t) \tag{18}$$

where, $l_t \in \mathbb{R}^M$ denotes the logits associated with the *M* admissible signal phases and $\mu_t \in [-1,1]$ is a continuous scalar controlling the green-time duration [4, 27, 37]. The discrete phase is selected via:

$$\varphi_t = \underset{i\in\{1,\ldots,M\}}{argmax} l_{t,i} \tag{19}$$

The continuous duration command is obtained through affine mapping as follows:

$$\Delta_t = \Delta_{min} + \frac{(\mu_t + 1)}{2}(\Delta_{max} - \Delta_{min}) \tag{20}$$

The hybrid action is executed in SUMO using a green-hold mechanism, whereby the selected phase φ_t is maintained for a duration proportional to Δ_t. This implementation preserves continuous timing flexibility while remaining compatible with the simulator's discrete phase execution model. The optimization of this hybrid action space is detailed in Section 5.5.

5.4 Reward function

This section specifies the normalized reward formulation used for actor–critic optimization, derived from the control objective defined in Section 4.3. At each decision step (*t*), the simulator provides the instantaneous throughput $N_{veh}(t)$, mean queue length $Q_{len}(t)$, and average waiting time $W_{time}(t)$. For consistency, $N_{veh}(t), Q_{len}(t)$ and $W_{time}(t)$ correspond to the aggregated traffic variables defined in Section 5.2, with $Q_{len}(t) = \bar{q}_t$ and $W_{time}(t) = \bar{w}_t$. Each metric is mapped to a bounded range using:

$$\hat{x}(t) = \tanh\left(\frac{x(t)}{k_x}\right) \tag{21}$$

where, the scaling factors k_x are selected based on empirical ranges observed in simulation. The resulting normalized reward is defined as:

$$\begin{aligned}\mathcal{R}_t = w_1 \tanh\left(\frac{N_{veh}(t)}{k_{thr}}\right) - w_2 \tanh\left(\frac{Q_{len}(t)}{k_{queue}}\right) \\ -w_3 \tanh\left(\frac{W_{time}(t)}{k_{wait}}\right)\end{aligned} \tag{22}$$

where, $w_1, w_2, w_3 > 0$ are fixed weighting coefficients inherited from the control objective.

This bounded formulation limits the influence of extreme congestion values, improves numerical conditioning of the critic, and promotes stable convergence under stochastic traffic dynamics [36-40]. The following section focuses on the optimization implications induced by the hybrid discrete–continuous structure of the action space.

5.4.1 Reward scaling and sensitivity analysis

The weighting coefficients (w_1, w_2, w_3) and normalization factors $\left(k_x \in \{k_{thr}, k_{queue}, k_{wait}\}\right)$ were determined based on empirical ranges observed in preliminary simulations. Typical traffic conditions exhibit queue lengths within the range of 0–20 vehicles and waiting times up to approximately 120 s, which guided the selection of scaling parameters to ensure balanced contributions across reward components, consistent with common practices in RL-based TSC [3, 14, 26].

The normalization factors are chosen such that the transformed variables operate within the quasi-linear region of the hyperbolic tangent (tanh) function under nominal conditions. This preserves sensitivity to traffic variations while preventing early saturation under high congestion, thereby improving numerical stability and learning robustness in DRL systems [1, 35].

Due to the bounded and smooth nature of the tanh transformation, the reward exhibits limited sensitivity to moderate perturbations in both scaling factors and weighting coefficients. Variations on the order of ±10–20% primarily affect the trade-off between throughput and delay, without significantly impacting convergence behavior or learning stability. Such robustness is particularly desirable in stochastic traffic environments, where variability in traffic demand can amplify learning instability [39, 41]. These observations were consistent across all demand scenarios.

Furthermore, the bounded normalization mitigates the influence of extreme values and prevents any single metric from dominating the reward signal, which is beneficial under highly dynamic traffic conditions [1, 39].

This robustness is further supported by consistent convergence patterns across multiple random seeds and demand scenarios (Section 6.5). Overall, the proposed formulation does not rely on finely tuned hyperparameters to achieve stable learning. A systematic sensitivity analysis is left for future work.

5.5 Hybrid action space and differentiability considerations

The proposed controller follows the PAMDP formulation, which couples discrete action selection with continuous action parameters. This structure naturally fits TSC, where each decision consists of selecting a signal phase and specifying its green-time duration. Since TD3 is inherently designed for continuous control, specific adaptations are required to accommodate this hybrid action space.

Building upon the formulation introduced in Section 5.3, the actor outputs a latent action $\tilde{a}_t = (l_t, \mu_t)$, which is mapped to the executed hybrid action: $a_t = (\varphi_t, \Delta_t)$. The discrete phase is selected as $\varphi_t = \underset{i\in\{1,\dots,M\}}{argmax} l_{t,i}$, and the continuous duration Δ_t is obtained from μ_t through the affine transformation defined in Eq. (20).

The actor's objective is defined as:

$$J(\theta_\pi) = \mathbb{E}_{s_t\sim\mathcal{D}}[Q(s_t, [onehot(\varphi_t); \Delta_t])] \tag{23}$$

Applying the chain rule, the gradient is:

$$\nabla_{\theta_\pi} J(\theta_\pi) = \mathbb{E}_{s_t\sim\mathcal{D}}\left[\frac{\partial Q}{\partial \Delta_t}\frac{\partial \Delta_t}{\partial \theta_\pi} + \frac{\partial Q}{\partial \varphi_t}\frac{\partial \varphi_t}{\partial l_t}\nabla_{\theta_\pi} l_t\right] \tag{24}$$

Due to the argmax operator, discrete phase selection is non-differentiable [9, 32]. Since argmax is piecewise constant and has zero gradient almost everywhere, we adopt a stop-gradient formulation $\varphi_t = \text{stop\_gradient}(\text{argmax}(l_t))$, which yields $\partial\varphi_t/\partial l_t = 0$. Consequently, the discrete term vanishes, and the gradient propagates exclusively through the continuous component Δ_t.

In the absence of gradient updates, the discrete phase component is optimized indirectly through critic-driven selection dynamics. The critic evaluates state–phase–duration tuples, and phases associated with higher Q-values are selected more frequently, increasing their frequency in the replay buffer. This induces an implicit greedy policy improvement mechanism over the discrete action space [9].

To ensure sufficient exploration during early training, stochasticity is introduced at the logit level by injecting Gaussian noise before the argmax operation. This promotes adequate exploration of the discrete action space before convergence toward deterministic policies.

Following the TD3 framework, target policy smoothing is applied during critic updates [27, 28]. In the proposed formulation, bounded Gaussian noise is added exclusively to the continuous duration parameter, while the discrete phase remains unchanged. This preserves the physical validity of signal phases and avoids introducing unsafe or infeasible transitions.

Overall, the proposed semi-differentiable optimization scheme provides a stable and well-posed framework for learning hybrid discrete–continuous policies within TD3, while ensuring operational validity under realistic traffic control constraints. A detailed derivation and implementation are provided in Section 5.6.4, ensuring clarity and reproducibility of the training procedure.

5.6 Hybrid Multi-Head TD3 architecture

The proposed controller builds upon the Twin Delayed Deep Deterministic Policy Gradient (TD3) framework [27] and is specifically adapted to the hybrid discrete–continuous decision structure of TSC. The architectural design addresses three key challenges: (i) representing diverse control strategies, (ii) stabilizing the value estimation under stochastic traffic dynamics, and (iii) preventing oscillatory timing behavior. To this end, the framework integrates a multi-head actor with attention-based fusion, twin critics with policy smoothing, and domain-specific stabilization mechanisms.

5.6.1 Multi-head actor with attention fusion

The actor network adopts a multi-head structure inspired by ensemble and policy diversification approaches [41, 42]. A shared encoder maps the observation state s_t into a latent feature vector h_t, which is passed to H independent policy heads:

$$\pi_i(s_t) = f_i(h_t), i = 1, \dots, H \tag{25}$$

Each head generates a candidate latent action and is encouraged to specialize in distinct behavioral primitives (e.g., queue clearance, delay balancing, congestion dissipation). An attention mechanism computes adaptive fusion weights:

$$\alpha_i(s_t) = \frac{\exp(g_i(h_t))}{\sum_{j=1}^{H} exp(g_i(h_t))} \tag{26}$$

where, g_i denotes a lightweight scoring network. The final latent action is obtained as follows:

$$\tilde{a}_t = \sum_{i=1}^{H} \alpha_i(s_t)\pi_i(s_t) \tag{27}$$

This formulation maintains policy diversity during training while enabling context-dependent selection of behavioral primitives, thereby improving robustness and providing diagnostic insight into policy behavior under heterogeneous traffic conditions [41].

The attention weights provide a structural and diagnostic indication of how behavioral primitives are combined under different traffic conditions. However, they do not constitute a causal explanation or formal feature attribution in the sense of explainable AI.

5.6.2 Twin critics and policy smoothing

Value estimation follows the TD3 framework, which is adapted to the parameterized hybrid action space. Two independent critics, Q_1 and Q_2, are maintained to mitigate the overestimation bias. The target value is computed using the clipped double-Q formulation:

$$Q_{target} = r_t + \gamma min(Q_1'(s_{t+1}, a'_{t+1}), Q_2'(s_{t+1}, a'_{t+1})) \tag{28}$$

Target policy smoothing is applied during critic updates by

adding bounded Gaussian noise only to the continuous duration component, as detailed in Section 5.6.4. This regularization prevents the exploitation of sharp Q-function peaks and suppresses abrupt green-time variations while preserving the integrity of the discrete phase decisions. Actor updates are delayed, as prescribed in TD3, to ensure that the policy gradients rely on stabilized value estimates.

5.6.3 Warm-up and anti-oscillation mechanisms

Learning in hybrid discrete–continuous action spaces is particularly sensitive to cold-start instability and abrupt policy fluctuations. To mitigate these effects, the proposed controller incorporates a structured warm-up phase, followed by explicit anti-oscillation mechanisms.

During the initial training episodes, a structured exploration strategy is employed, combining random discrete phase selection with bounded noise injection on the continuous duration parameter. This strategy promotes broad state–action coverage and stabilizes early critic learning by preventing premature convergence to suboptimal timing patterns [4].

During execution, two complementary constraints are enforced to ensure a physically plausible and stable signal operation. First, a minimum-hold constraint guarantees that once a signal phase is activated, it remains active for a minimum duration consistent with traffic engineering standards, thereby preventing rapid and unsafe phase toggling. Second, the raw green-time duration Δ_t is filtered using an exponential moving average to produce the executed duration:

$$\Delta_t^{smooth} = \beta\Delta_t + (1-\beta)\Delta_{t-1}^{smooth}, 0 < \beta < 1 \tag{29}$$

This temporal smoothing compensates for noisy DRL action outputs, reduces oscillatory signal behavior, and yields smoother control trajectories in hybrid discrete–continuous settings.

5.6.4 Replay buffer and training updates

Training is performed off-policy using a replay buffer $\mathcal{D}$ that stores transition tuples $(s_t, a_t, r_t, s_{t+1}, d_t)$. Uniform mini-batch sampling reduces temporal correlations and improves sample efficiency.

The twin critics are updated by minimizing the Bellman error using clipped double-Q learning:

$$y_t = r_t + \gamma(1-d_t)\min_{i=1,2} Q_{\theta_i'}(s_{t+1}, a'_{t+1}) \tag{30}$$

where the target action is defined as:

$$\begin{aligned} a'_{t+1} &= [onehot(\varphi_{t+1}); \Delta'_{t+1}], \\ \Delta'_{t+1} &= clip(\Delta_{t+1} + \epsilon, \Delta_{min}, \Delta_{max}), \epsilon \sim \mathcal{N}(0, \sigma^2) \end{aligned} \tag{31}$$

Target policy smoothing is applied exclusively to the continuous duration Δ_{t+1}, while the discrete phase φ_{t+1} remains unchanged to preserve the categorical structure.

The actor is updated in a delayed manner by maximizing the expected value estimated by the primary critic:

$$\nabla_{\theta_\pi} J(\theta_\pi) = \mathbb{E}_{s_t \sim \mathcal{D}}[\nabla_{\theta_\pi} Q_{\theta_1}(s_t, [onehot(\varphi_t); \Delta_t])] \tag{32}$$

Due to the non-differentiability of the argmax operator used for phase selection, a stop-gradient formulation is adopted, i.e., $\partial\varphi_t/\partial l_t = 0$. Applying the chain rule, the policy gradient decomposes into:

$$\nabla_{\theta_\pi} J(\theta_\pi) = \mathbb{E}_{s_t \sim \mathcal{D}}\left[\frac{\partial Q}{\partial \Delta_t}\nabla_{\theta_\pi}\Delta_t + \frac{\partial Q}{\partial \varphi_t}\frac{\partial \varphi_t}{\partial l_t}\nabla_{\theta_\pi} l_t\right] \tag{33}$$

Since $\partial\varphi_t/\partial l_t = 0$, the policy gradient reduces to the continuous component:

$$\nabla_{\theta_\pi} J(\theta_\pi) = \mathbb{E}_{s_t \sim \mathcal{D}}\left[\frac{\partial Q}{\partial \Delta_t}\nabla_{\theta_\pi}\Delta_t\right] \tag{34}$$

The discrete phase component is therefore optimized indirectly through critic-driven selection dynamics, where phases associated with higher Q-values are selected more frequently, inducing an implicit greedy policy improvement over the discrete action space.

To ensure sufficient exploration during early training, stochasticity is introduced at the logit level $l_t^{noisy} = l_t + \epsilon_l$, $\epsilon_l \sim \mathcal{N}(0, \sigma_l^2)$.

This promotes adequate exploration of the discrete action space before convergence toward deterministic policies.

To further clarify the training procedure, the actor update can be summarized as follows:

Forward pass:
Compute logits and duration $(l_t, \Delta_t) = \pi(s_t)$
Discrete phase selection (non-differentiable):
Add exploration noise: $l_t^{noisy} = l_t + \epsilon_l, \epsilon_l \sim \mathcal{N}(0, \sigma_l^2)$
Select Phase: $\varphi_t = \text{argmax}(l_t^{noisy})$
Detach discrete branch $\varphi_t = \text{stop\_gradient}(\varphi_t)$
Critic evaluation
Compute Q-value: $\text{Q}_{\text{val}} = Q_1(s_t, onehot(\varphi_t), \Delta_t)$
Actor update
$\text{L}_{\text{actor}} = -\text{Q}_{\text{val}}.\text{mean}()$
$\theta_\pi \leftarrow \theta_\pi - \alpha_\pi \nabla_{\theta_\pi} \text{L}_{\text{actor}}$

Gradients propagate only through Δ_t, while the discrete phase φ_t receives no gradient updates.

Finally, target networks are softly updated:

$$\theta' \leftarrow \tau\theta + (1-\tau)\theta', 0 < \tau \ll 1 \tag{35}$$

This decoupled update scheme provides a stable and well-posed optimization framework, where continuous parameters are learned via gradient descent and discrete decisions evolve through value-based selection.

5.7 Training pipeline and algorithm

The training pipeline defines the interaction between the SUMO simulation environment and the proposed HMH-TD3 controller, enabling stable off-policy learning under stochastic and nonlinear traffic dynamics. Training follows the TD3 actor–critic scheme and integrates warm-up exploration, replay-based updates, multi-head policy optimization, and periodic evaluation.

Training is conducted episodically. At the beginning of each episode, the environment is reset according to the specified demand scenario. At each decision step, the agent observes the current traffic state and selects a hybrid control action composed of a discrete signal phase and a continuous green-time duration. During an initial warm-up period, exploratory actions are applied to populate the replay buffer and stabilize early critic learning. Once sufficient experience is collected, actions are generated by the learned multi-head actor and executed using the green-hold mechanism.

Learning updates follow the TD3 schedule. Twin critics are

updated at every step using mini-batches sampled from the replay buffer, whereas actor updates are performed at a lower frequency to improve stability. Target networks are synchronized via soft updates. Policy performance and generalization are periodically assessed in deterministic (noise-free) evaluation mode using traffic-level metrics, and internal policy behavior can be analyzed through attention-weight evolution.

During training, exploration noise and target policy smoothing are applied exclusively to the continuous duration component, whereas discrete phase selection is evaluated through the critic without gradient propagation. The complete training procedure is summarized in Algorithm 1.

Algorithm 1. HMH-TD3 for Traffic Signal Control

Input: SUMO environment $\mathcal{E}$, episode limit N_{ep}, maximum steps per episode T, actor network π_θ with $\mathcal{H}$ heads, critic networks $Q_{\theta_1}, Q_{\theta_2}$, replay buffer $\mathcal{D}$.

Initialize: Target networks $\theta' \leftarrow \theta$, Critic targets $\theta_1' \leftarrow \theta_1$, $\theta_2' \leftarrow \theta_2$, Replay buffer $\mathcal{D} = \emptyset$, smoothed duration $\Delta_0^{smooth} = 0$.

For episode k = 1 …N_{ep} **do**

Reset SUMO, observe initial state s_0

For t = 0…T-1 do

a. **Action Selection**

If the warm-up phase **then**

Sample φ_t uniformly from Φ

Sample Δ_t uniformly from $[\Delta_{min}, \Delta_{max}]$.

Else (use learned policy):

Compute latent action: $\tilde{a}_{t=} \sum_{i=1}^{H} \alpha_i(s_t)\pi_i(s_t)$

Decompose $\tilde{a}_t = (l_t, \mu_t)$

Select discrete phase: $\varphi_t = \underset{i\in\{1,\dots,M\}}{argmax}\, l_{t,i}$

Map duration: $\Delta_t = \Delta_{min} + \frac{(\mu_t+1)}{2}(\Delta_{max} - \Delta_{min})$

Endif

Apply temporal smoothing:

$\Delta_t^{smooth} = \beta\Delta_t + (1-\beta)\Delta_{t-1}^{smooth}$

b. Environment Step

Execute action $a_t = [onehot(\varphi_t); \Delta_t^{smooth}]$, observe (s_{t+1}, r_t, d_t).

c. **Store transition:**

Add $(s_t, a_t, r_t, s_{t+1}, d_t)$ to buffer $\mathcal{D}$.

d. **Critic Updates**

Sample mini-batch $\mathcal{B} \subset \mathcal{D}$

For each $(s_j, a_j, r_j, s_{j+1}, d_j)$ in B **do**

Compute target action from target actor:

$(l_j', \mu_j') = \pi_{\theta'}(s_{j+1})$

$\varphi_j' = \underset{i\in\{1,\dots,M\}}{argmax}\, l_{j,i}'$

$\Delta_j = \Delta_{min} + \frac{(\mu_j'+1)}{2}(\Delta_{max} - \Delta_{min})$

Apply target policy smoothing (duration **only)**:

$\Delta_j' = clip(\Delta_j + \epsilon, \Delta_{min}, \Delta_{max}),\ \epsilon \sim \mathcal{N}(0, \sigma^2)$

Form target action: $a'_j = [onehot(\varphi_j'); \Delta_j']$

Compute target:

$y_j = r_j + \gamma(1 - d_j)\min_{i=1,2} Q_{\theta_i'}(s_{j+1}, a'_j)$

End for

Update the critic by minimizing, for $i \in \{1,2\}$:

$\mathcal{L}_i = \frac{1}{|B|}\sum_{j\in B}\left(Q_{\theta_i}(s_j, a_j) - y_j\right)^2$

e. **Delayed Actor Update**

If t mod 2 = 0 **then**

$\theta \leftarrow \theta + \alpha_\pi \nabla_\theta \frac{1}{|B|}\sum_{j\in B} Q_{\theta_1}(s_j, [onehot(\varphi_j); \Delta_j])$

Endif

f. **Soft Target Update**

$\theta' \leftarrow \tau\theta + (1-\tau)\theta'$

$\theta_i' \leftarrow \tau\theta_i + (1-\tau)\theta_i',\ i \in \{1,2\}$

If $d_t = 1$ **then** break **Endif**

End For

Periodically evaluate the policy in noise-free mode.

End For

6. EXPERIMENTAL SETUP AND EVALUATION

This section describes the experimental protocol used to evaluate the proposed HMH-TD3 controller in a controlled microscopic traffic simulation environment. The evaluation focuses on delay reduction, congestion mitigation, throughput, and control stability under stochastic and time-varying demand regimes.

6.1 Simulation environment and protocol

Experiments were conducted using the SUMO 1.18.0 microscopic traffic simulator, which was interfaced via TraCI with a simulation step of 1 s. The controlled network corresponds to the isolated four-leg intersection defined in Section 4, with identical geometry, detector placement, and signal constraints across all experiments. Vehicle dynamics follow the Krauss car-following model, and signal operations respect standard safety constraints, including minimum green times, a 3 s yellow phase, and a 2 s all-red clearance.

To ensure full reproducibility, fixed random seeds were applied to all the stochastic components (SUMO, Python, NumPy, and PyTorch). Each experiment was repeated over five independent runs (seeds: 42, 123, 2024, 7, and 888), and the results are reported as mean values with standard deviation bands. To further ensure the robustness of the reported results, statistical significance tests are conducted across runs, as detailed in Section 6.5.

6.2 Traffic demand scenarios

Traffic demand was generated using non-stationary Poisson arrival processes over 3600 s episodes. Three representative regimes were considered:

Uniform demand: balanced arrivals across all approaches with rates ($\lambda \in [\lambda_{min}, \lambda_{max}]\ veh/h/lane$).

Asymmetric demand: one dominant approach with ($\lambda_{major} \gg \lambda_{minor}$), stressing prioritization and spillback prevention.

Stochastic demand: time-varying arrivals following sinusoidal or step-based profiles $\lambda(t) = \lambda_0 + Asin(\omega t)$ capturing rapid transitions and short-term disturbances.

6.3 Baseline methods for comparison

The proposed controller is benchmarked against four baselines spanning classical, rule-based, and RL methods:

FTC: A pre-timed plan with fixed cycle lengths and splits computed offline, serving as a non-adaptive, lower bound.

VA: A rule-based controller adjusting green times based on detector gaps, providing limited heuristic responsiveness.

Deep Q-Network (DQN): A discrete RL controller selecting

phases from a fixed set without continuous duration control, representing standard discrete RL approaches.

Standard TD3 (Single-Head): A continuous actor–critic baseline using the same hybrid action space but a single MLP actor. This baseline isolates the structural impact of the multi-head attention mechanism under identical training conditions. However, we note that the multi-head architecture introduces additional parameters, which may partially influence performance. This aspect is further discussed in Section 6.6.

6.4 Performance metrics

Performance is evaluated using standard traffic-engineering and RL metrics computed from lane-level SUMO measurements [4, 5, 16]. All metrics are directly derived from the macroscopic state variables defined in Section 5.2 (Eqs. (10)-(17)).

Operational efficiency: Three indicators are used to quantify traffic efficiency.

Average queue length (veh): The average queue length measures congestion intensity as the temporal mean of the aggregated queue length $\bar{q}_t$, introduced in Eq. (14):

$$\bar{q} = \frac{1}{T}\sum_{t=0}^{T-1} \bar{q}_t \tag{36}$$

where, T denotes the total number of decision steps in the evaluation episode.

Average waiting time (s/veh): The average waiting time captures the mean delay experienced by vehicles and is defined as the temporal average of the aggregated waiting time $\bar{w}_t$ (Eq. (12)):

$$\bar{w} = \frac{1}{T}\sum_{t=0}^{T-1} \bar{w}_t \tag{37}$$

Intersection throughput (veh/h) is defined as the rate of vehicles successfully discharged from the intersection:

$$TH = \frac{\mathcal{N}_{exit}}{T_{sim}} \times 3600 \tag{38}$$

where, $\mathcal{N}_{exit}$ denotes the total number of vehicles exiting the network during the simulation time T_{sim} (in seconds).

Congestion and control stability: To monitor congestion buildup and safety-critical operating conditions, the Saturation Ratio (%) is analyzed based on the definition in Eqs. (15)-(16). The mean saturation and the maximum saturation are computed as temporal averages of $\bar{s}_t$ and $s_{max}(t)$ respectively.

$$\bar{s} = \frac{1}{T}\sum_{t=0}^{T-1} \bar{s}_t\,, \qquad s_{max} = \frac{1}{T}\sum_{t=0}^{T-1} s_{max}(t) \tag{39}$$

These metrics capture both global congestion levels and localized spillback effects. In addition, learning stability indicators, including episodic return, actor and critic loss trajectories, and update variance, are analyzed to assess the convergence behavior and robustness to noise, following standard continuous-control RL practices [27].

Deployment feasibility: Practical deployability is assessed using two metrics. Phase switching frequency (switches/h) assesses signal smoothness and operational plausibility. The frequency of phase transitions is computed as:

$$F_{switch} = \frac{\mathcal{N}_{switch}}{T_{sim}} \times 3600 \tag{40}$$

where, $\mathcal{N}_{switch}$ is the total number of phase changes during the simulation. This metric ensures compliance with anti-oscillation constraints described in Section 5.6.3. Computational efficiency (ms/decision) measures the average inference time per control step to confirm real-time feasibility.

6.5 Results and discussion

This section evaluates the learning dynamics, traffic-level performance, and internal policy behavior of the proposed HMH-TD3 controller across three demand scenarios. The analysis integrates training dynamics (Figure 3), multi-head diagnostics (Figure 4), traffic-level performance metrics (Figure 5), and attention–traffic alignment (Figure 6) to provide a comprehensive understanding of both performance and internal policy structure.

Learning dynamics and convergence: Training exhibits a consistent two-phase pattern (Figure 3). The initial exploration phase (Episodes 1–60) is characterized by high return variance, reflecting active exploration of the hybrid phase–duration action space. Stabilization occurs after approximately 80–120 episodes, followed by steady convergence between Episodes 300 and 500, with overall return improvements of 35–45%. Actor and critic losses remain bounded and smooth across all scenarios, confirming that TD3 stabilization mechanisms effectively handle the hybrid semi-differentiable action formulation. These results demonstrate that discrete phase selection and continuous duration control can be reliably learned within an off-policy actor–critic framework.

Traffic-level performance: The learned policy translates into significant operational improvements (Figure 5). Mean waiting time decreases from values exceeding 40 s to stable levels between 7 and 12 s across all demand regimes. Simultaneously, throughput increases by approximately 10–15%, reaching 0.30–0.32 veh/s, while queue length and saturation remain within stable, non-critical ranges.

These results indicate that performance gains arise from adaptive and anticipatory signal timing rather than short-term reward exploitation. The controller effectively balances responsiveness and phase persistence, avoiding oscillatory behavior while maintaining efficient traffic flow. As shown in Figure 4, different attention heads exhibit distinct activation patterns associated with specific traffic regimes (e.g., congestion clearance vs. flow stabilization), contributing to the balanced performance observed in Figure 5.

Computational efficiency: As defined in Section 6.4, the average inference time per decision step is 8.7 ± 1.2 ms for the proposed HMH-TD3 controller, measured over five independent runs under identical experimental conditions on a desktop-class CPU (Intel Core i7 @ 3.0 GHz or equivalent), without GPU acceleration.

For comparison, the single-head TD3 and DQN baselines require 6.1 ± 0.9 ms and 4.3 ± 0.7 ms per decision, respectively, while rule-based controllers (VA and FTC) incur negligible computational overhead (< 1 ms per decision).

Despite the additional complexity introduced by the multi-head architecture and attention-based fusion, the proposed

controller remains computationally efficient. All methods operate well below the 1 s control interval used in the simulation environment, confirming real-time feasibility. Low variance across runs and the absence of inference spikes further indicate stable and predictable deployment behavior.

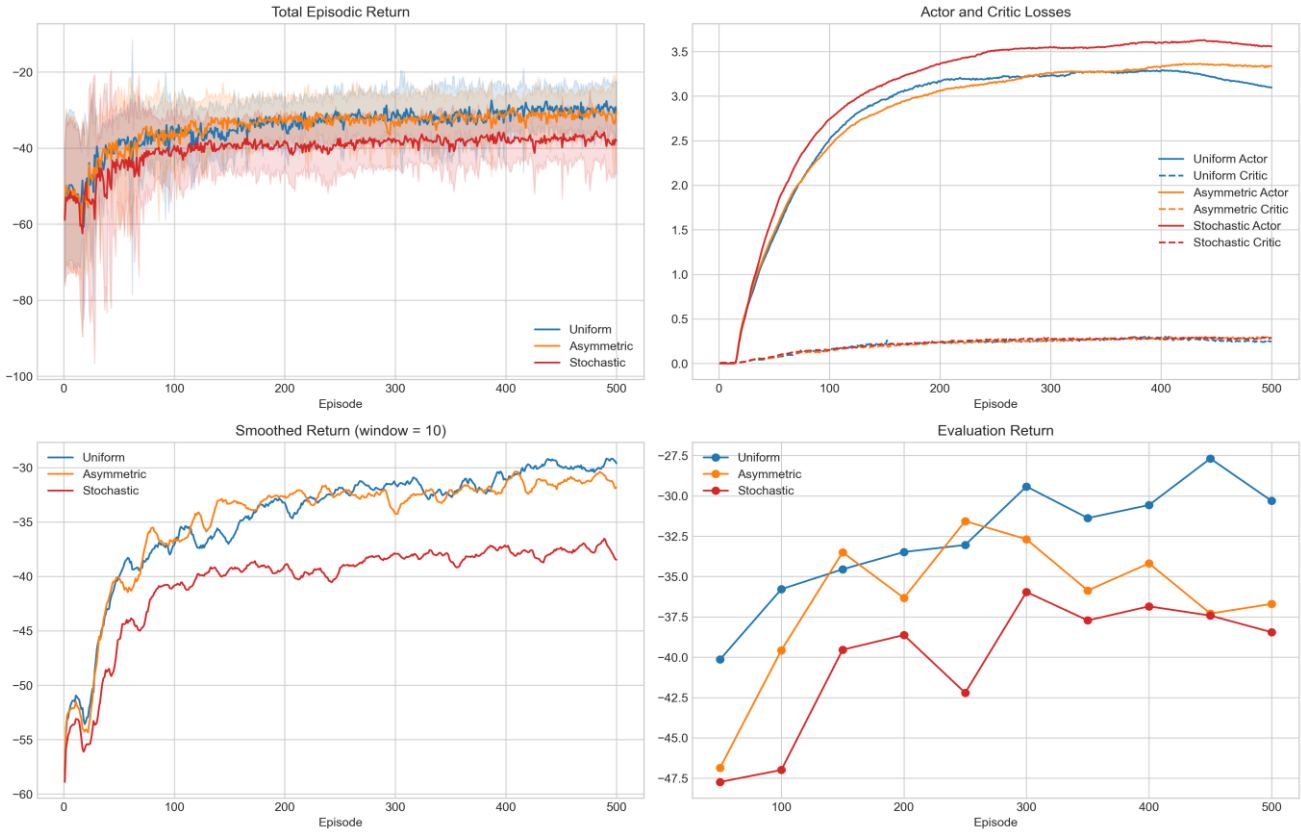
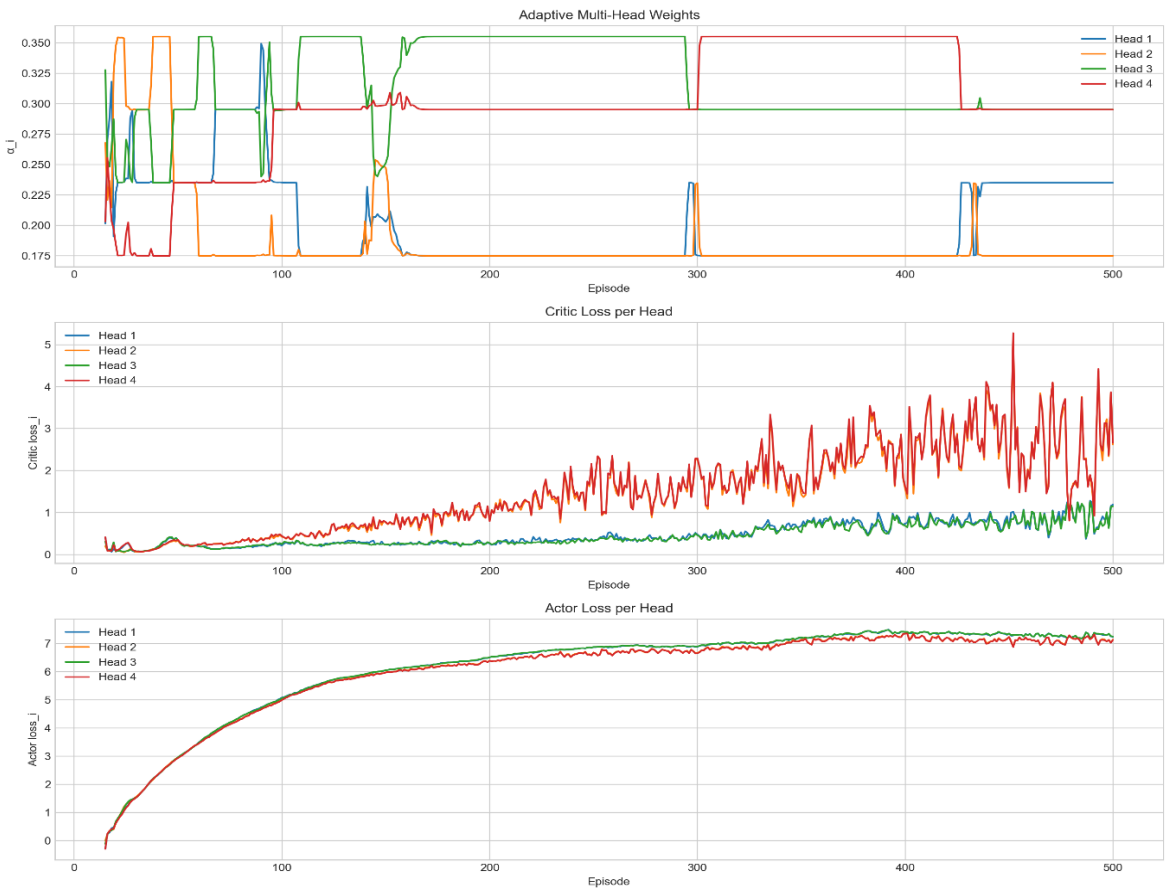
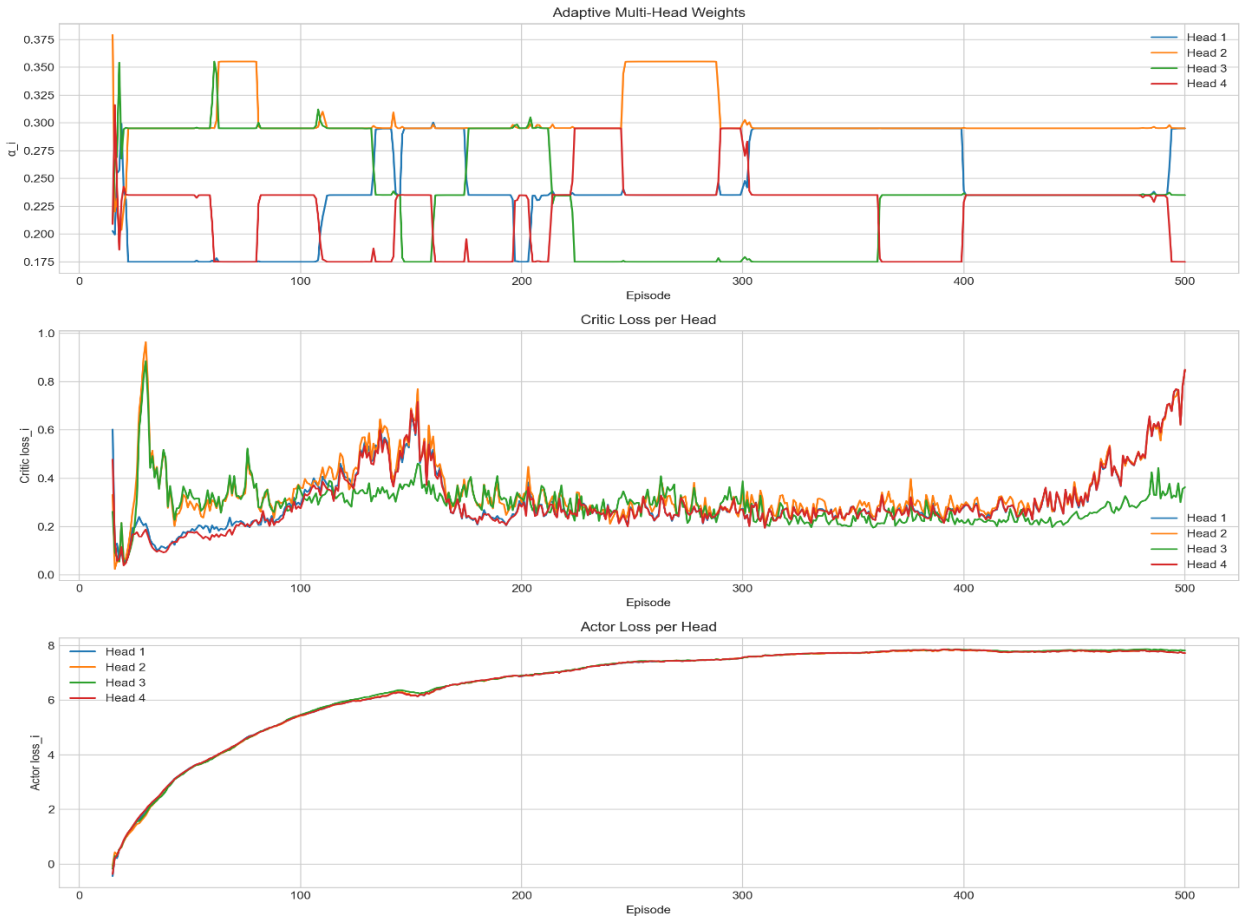


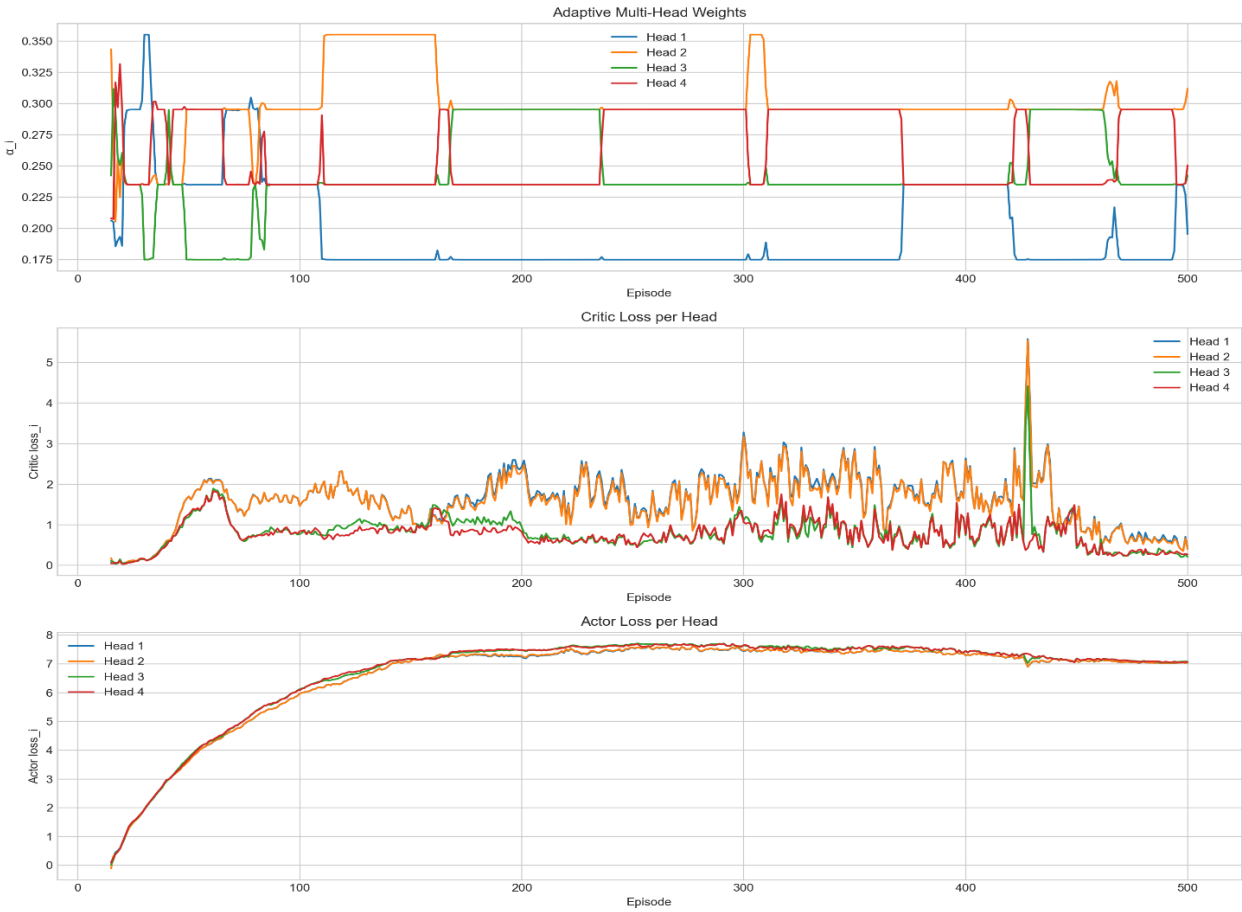
Figure 3. Training summary



(a) Asymmetric



(b) Stochastic



(c) Uniform

Figure 4. Multi-head diagnostics

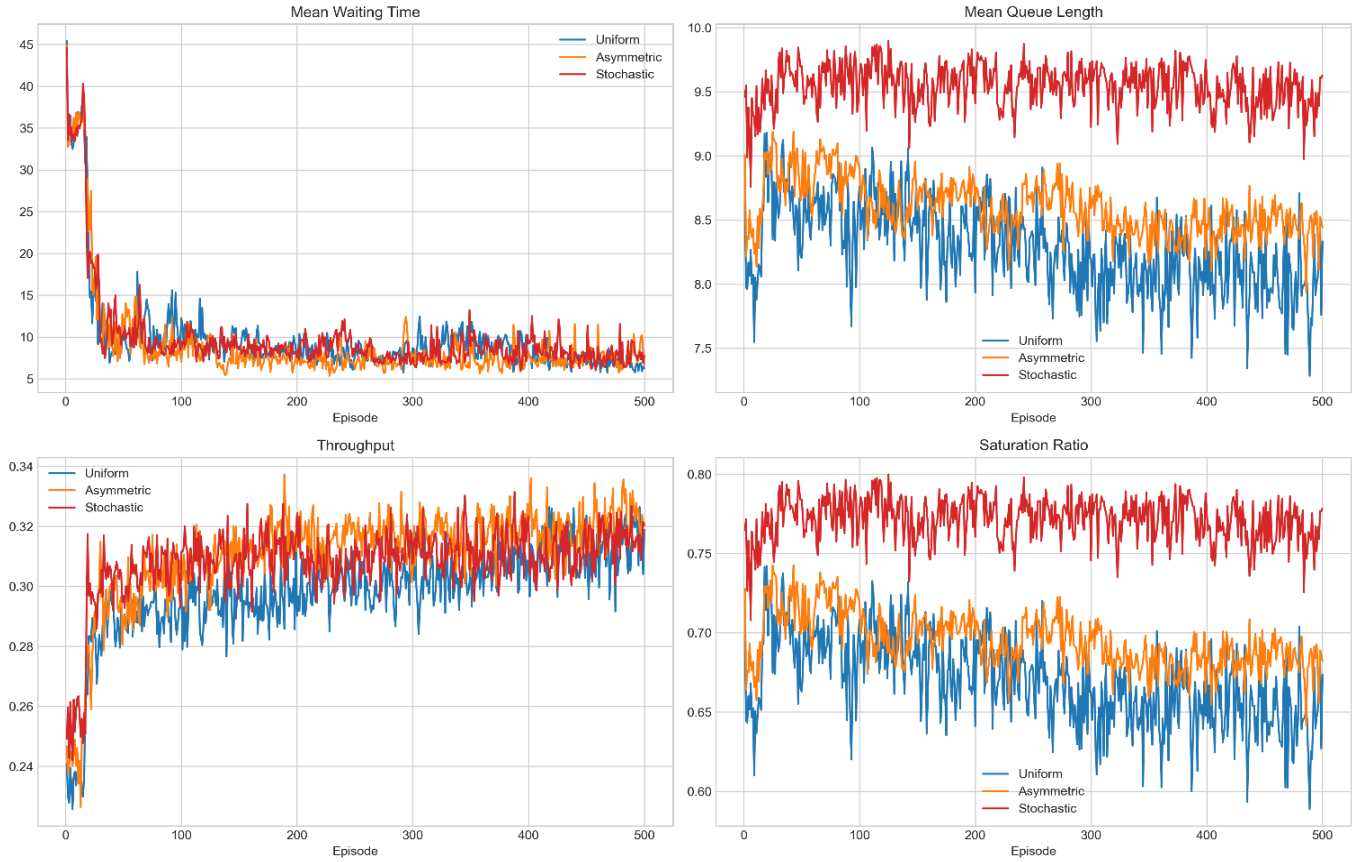


Figure 5. Traffic summary

Architectural analysis and diagnostic insight: The multi-head structure enables an implicit functional decomposition of the control policy (Figure 4). Different heads exhibit distinct activation patterns across scenarios, indicating differentiated behavioral roles that are further analyzed in the attention-traffic alignment study (Figure 6). The attention mechanism dynamically combines these components, resulting in smooth transitions and reduced oscillatory control. While these attention weights do not provide feature attribution or causal explanation, they offer a structured and behaviorally grounded diagnostic view of how policy components are activated over time.

Attention–traffic alignment analysis: To investigate structural interpretability, we analyze the relationship between traffic dynamics and attention allocation (Figure 6). The dominant attention head is identified as the one exhibiting the highest Pearson correlation with normalized queue length.

The results reveal a consistent alignment between congestion dynamics and internal policy weighting. Periods of queue buildup coincide with increased activation of the dominant attention head (Figure 6(a)-(c)), with positive correlations across all scenarios (r = 0.58 for asymmetric demand, r = 0.47 for stochastic demand, and r = 0.43 for uniform demand).

This relationship is further supported by the correlation heatmap (Figure 6(d)), confirming a structured and non-random association between traffic states and attention weights across heads and scenarios.

To assess robustness, a permutation test was conducted by randomly shuffling the queue signals. The resulting correlations collapse to near zero (mean $\approx$ 0, p = 0.001), demonstrating that the observed alignment is not due to spurious temporal correlations.

These findings suggest that the policy exhibits a structured functional organization, where specific heads specialize in congestion-sensitive regimes. While this does not provide a causal explanation or feature attribution, it offers a diagnostic characterization of policy behavior, supporting the notion of structural interpretability.

Interestingly, the strength and nature of this alignment vary across demand scenarios. Under asymmetric demand, the alignment is stronger and more stable, reflecting predictable congestion patterns. In contrast, stochastic demand induces more dynamic attention reallocation, while uniform demand results in more evenly distributed attention across heads. This indicates that the learned policy adapts its internal representation to the statistical structure of the environment.

Statistical significance analysis: Paired t-tests were conducted across five independent runs for key performance metrics, including waiting time, queue length, throughput, and saturation. The results show consistent improvements in waiting time and queue length across all scenarios, although these differences do not reach statistical significance at the 5% level, likely due to limited statistical power.

However, effect sizes are moderate to large in several cases (e.g., Cohen's $d > 1$), indicating meaningful and consistent performance differences that are not fully captured by p-values alone.

In contrast, statistically significant differences are observed in throughput when compared to DQN ($p < 0.05$), reflecting a trade-off between throughput maximization and traffic stability.

Overall, these results indicate that the observed performance trends are consistent across runs and not attributable to random variability, while highlighting the inherent stability–efficiency trade-off in TSC.

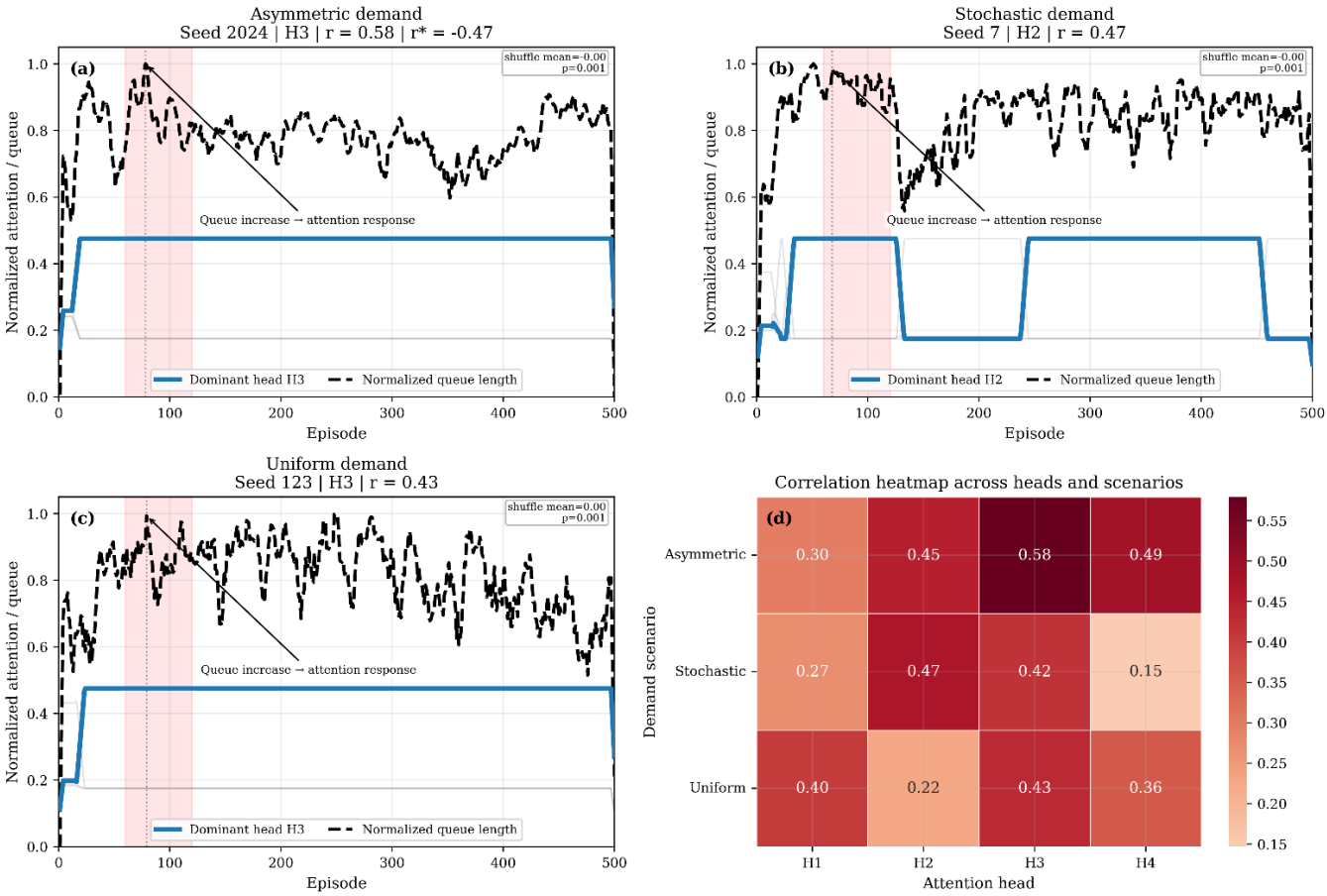


Figure 6. Attention–traffic alignment across demand scenarios: (a) Asymmetric, (b) Stochastic, (c) Uniform demand showing normalized queue length (black dashed) and the dominant attention head (blue), with others in gray, (d) Correlation heatmap of queue length vs. attention weights across heads and scenarios, confirming a structured, non-random relationship between traffic states and policy weighting

Note: Shaded regions indicate congestion intervals. The dominant head is selected via the highest Pearson correlation (reported per subplot), showing consistent alignment between queue buildup and attention allocation.

6.6 Ablation study

A controlled ablation study was conducted to quantify the contribution of each architectural component of the proposed HMH-TD3 controller under the three traffic demand scenarios described in Section 6.2. In each experiment, one component was removed while all other training conditions and hyperparameters were held constant, enabling controlled attribution of performance variations. The analysis is based on learning dynamics (Figure 3) and internal policy diagnostics (Figure 4), and aggregated traffic-level metrics (Figure 5).

Hybrid action parameterization: The effect of continuous duration control was evaluated by comparing the full model with a discrete-only TD3 variant using fixed green time. Removing the duration modulation led to slower convergence and degraded traffic performance across all scenarios, with increased delays and reduced throughput. Degradation was most pronounced under stochastic demand, where limited temporal flexibility constrained adaptation to rapid traffic fluctuations. These results indicate that continuous green-time modulation contributes to responsiveness and performance stability under non-stationary conditions.

Multi-head actor: To assess policy diversification, the multi-head actor was replaced with a single-head TD3 architecture while maintaining identical critics and action spaces. The single-head variant exhibited a higher variance in the learning curves and reduced robustness under asymmetric and stochastic demands. This variability was reflected in less stable traffic-level performance, with larger oscillations in the queue lengths and waiting times. These observations show that multi-head decomposition supports behavioral specialization and stabilizes policy optimization.

Attention-based fusion: Replacing the attention mechanism with uniform averaging across policy heads resulted in lower cumulative returns and weaker adaptation to asymmetric traffic patterns. Without adaptive weighting, behavioral responses become less differentiated across traffic contexts, reducing the ability to prioritize critical maneuvers, such as clearing saturated approaches. This highlights the role of attention-based fusion in context-aware policy coordination and diagnostic behavioral weighting.

Stabilization mechanisms: Joint removal of TD3 stabilization components (twin critics, delayed updates, and policy smoothing) and domain-specific anti-oscillation constraints caused unstable learning dynamics, increased gradient variance, excessive phase switching, and reduced throughput. Under high demand, the resulting control behavior becomes operationally unstable, degrading the intersection performance. These findings indicate that stabilization mechanisms are essential for maintaining a realistic and robust signal control.

Summary of findings: Overall, the ablation results indicate that the proposed components are complementary rather than additive. As summarized in Table 1, their joint integration is

required to achieve stable learning dynamics and balanced traffic performance across diverse demand regimes.

Table 1. Ablation study: Component-level insights

Removed Component	Primary Failure Mode	Dominant Traffic Impact	Key Insight
Continuous duration control	Reduced adaptability	Higher delay, lower throughput	Fine-grained timing control is essential under non-stationary demand
Multi-head actor	Increased learning variance	Performance degradation under asymmetry	Policy diversification improves robustness
Attention-based fusion	Loss of context awareness	Poor scenario-specific adaptation	Adaptive weighting enables diagnostic strategy selection
Stabilization mechanisms	Unstable learning dynamics	Excessive switching, throughput loss	Stabilization is required for safe and deployable control

Model capacity discussion: The multi-head actor introduces a larger number of parameters compared to the single-head TD3 baseline. While this increased capacity may contribute to performance improvements, the observed gains are not limited to aggregate performance metrics but are consistently reflected in stability-related indicators.

In particular, the multi-head architecture consistently exhibits reduced variance in learning dynamics, improved stability, and more robust and consistent control behavior under stochastic and asymmetric demand conditions. Importantly, increasing model capacity alone does not inherently induce structured behavioral specialization or context-aware decision mechanisms. A purely capacity-driven improvement would not explain the observed reduction in variance and the emergence of structured attention patterns.

These properties are closely linked to structured policy decomposition and attention-based fusion, which enable specialization and context-aware decision-making.

While a strictly controlled parameter-matched comparison, e.g., scaling the single-head architecture to match the number of parameters, would provide a more rigorous evaluation of the relative contributions, it is left for future work due to its high computational cost and expanded experimental scope.

6.7 Comparative evaluation across all controllers

This section compares the proposed HMH-TD3 controller with four baselines: single-head TD3, DQN, VA, and FTC, under uniform, asymmetric, and stochastic demand scenarios. The analysis combines quantitative metrics (Table 2), bar-chart summaries (Figure 7), radar-based multi-metric profiles (Figure 8), and a normalized performance heatmap (Figure 9).

Table 2. Mean post-convergence traffic performance metrics across demand scenarios

Method	Scenario	Average Waiting Time (s/veh)	Average Queue Length (veh)	Throughput (veh/h)	Saturation Ratio (%)	Phase Switching (/h)
HMH-TD3	Uniform	8.53	7.69	1067.15	62.19	24
TD3		9.53	6.08	865.01	49.10	37
DQN		11.60	9.24	910	74.72	58
VA		64.76	9.94	874.29	80.36	52
FTC		68.19	9.84	880.00	79.54	18
HMH-TD3	Asymmetric	5.24	6.92	1031.22	55.96	28
TD3		4.72	9.38	927.17	75.81	41
DQN		9.18	8.94	1290.91	72.28	66
VA		64.64	9.84	822.86	79.51	50
FTC		67.92	9.86	982.86	79.71	15
HMH-TD3	Stochastic	8.33	8.14	1034.01	65.81	29
TD3		5.27	10.02	1173.08	80.97	44
DQN		12.05	9.98	1209.84	80.66	72
VA		64.76	9.78	900.00	79.08	56
FTC		67.87	9.97	902.86	80.58	21

Note: DQN: Deep Q-Network; HMH-TD3: Hybrid Multi-Head TD3; FTC: fixed-time control; VA: vehicle-actuated.

Uniform demand: Under a uniform demand, HMH-TD3 achieves the most balanced operational performance. As shown in Figure 7, it attains the lowest average waiting time (8.53 s/veh) and the highest throughput (1067 veh/h) while maintaining a moderate phase-switching frequency. In contrast, the Single-Head TD3 exhibits less efficient green-time utilization, and the rule-based controllers (VA and FTC) incur delays exceeding 60 s owing to limited adaptability. This balance is confirmed by the radar profile in Figure 8 (Uniform), where HMH-TD3 maintains consistently low values across the delay, queue length, saturation, and switching metrics.

Asymmetric demand: Under directional imbalance, the performance disparities increase markedly. HMH-TD3 effectively limits congestion spillback by maintaining low average queue lengths (6.92 veh) and controlled saturation levels (≈ 56%), as reported in Table 2 and Figure 7. In contrast, the Single-Head TD3 and DQN exhibit higher saturation ratios and increased phase-switching frequencies, indicating reduced stability. These trends are reflected in Figure 8 (Asymmetric), where HMH-TD3 consistently achieves superior performance across stability-related dimensions. This robustness is attributable to the attention mechanism, which emphasizes the heads specialized in heavy-flow dissipation during directional demand peaks.

Stochastic demand and stability–throughput trade-off: Under stochastic demand, a clear stability–throughput trade-off emerges. As shown in Figure 7, the DQN achieves a higher instantaneous throughput (≈ 1200 veh/h), but this is accompanied by saturation levels exceeding 80% and frequent

phase switching, indicative of oscillatory control. From a traffic engineering perspective, sustained saturation above approximately 75–80% corresponds to critical operating conditions associated with spillback amplification and gridlock risk, as documented in the Highway Capacity Manual [43]. In contrast, HMH-TD3 maintains bounded saturation (≈ 66%), favoring stable queue evolution and predictable signal behavior over short-term discharge maximization. This explains why higher throughput alone does not translate into better overall traffic performance in stochastic conditions, as further reflected in Figures 8 and 9.

The proposed controller is not designed to maximize throughput at all costs, but rather to maintain a balanced and sustainable operating regime by controlling saturation and limiting oscillatory behavior. Consequently, HMH-TD3 deliberately sacrifices a portion of short-term throughput to ensure long-term stability, robustness, and operational safety. This reflects a deliberate design choice prioritizing stability and operational robustness over throughput maximization, rather than a limitation of the proposed method.

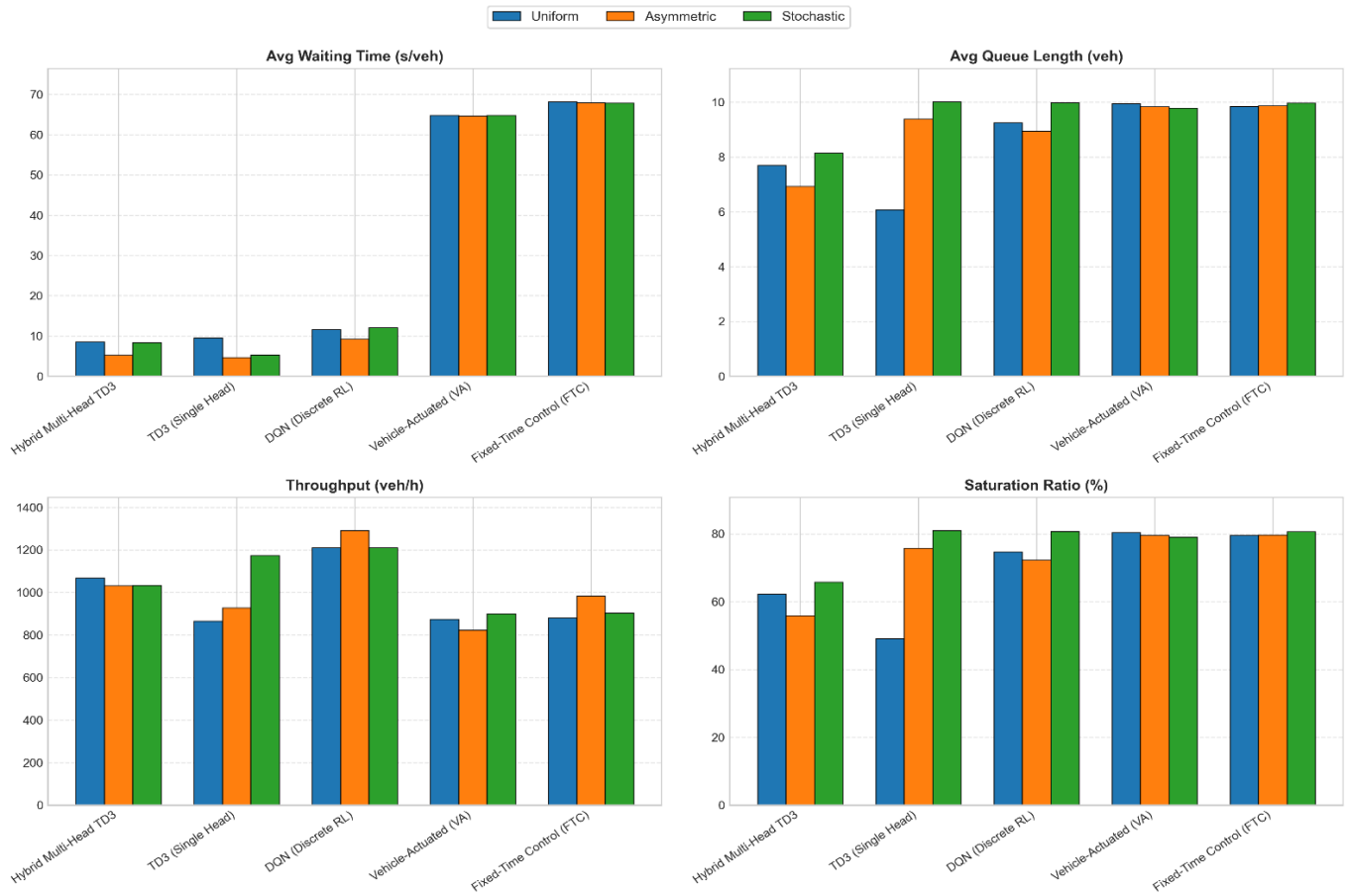


Figure 7. Traffic performance metrics across controllers and demand scenarios

Note: Subplots evaluate average waiting time, queue length, throughput, and saturation ratio. Bar groups represent control methods (HMH-TD3, TD3, DQN, VA, FTC) across uniform, asymmetric, and stochastic demand. HMH-TD3 achieves lower delay and improved saturation control.

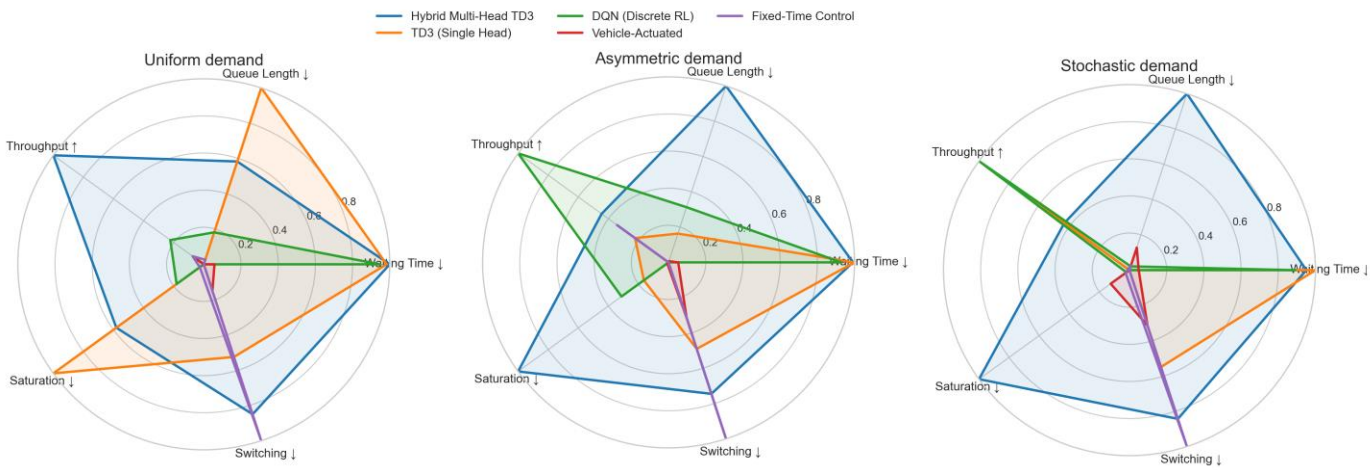


Figure 8. Normalized performance radar plots under varying demand

Note: Comparison of efficiency, stability, and switching behavior for (left) uniform, (center) asymmetric, and (right) stochastic scenarios. Metrics are scaled to [0,1] (higher is better); congestion-related indicators are inversely normalized.

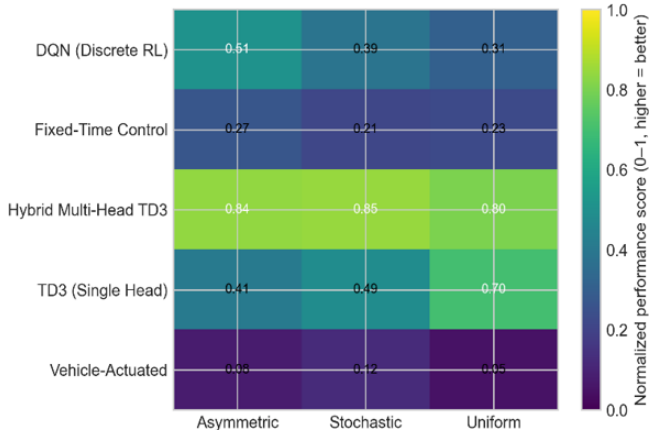


Figure 9. Composite performance heatmap across controllers and scenarios

Note: Rows denote control methods and columns represent demand types. Each cell displays a composite score in [0,1], computed as the mean of normalized metrics (waiting time, queue length, throughput, saturation, and switching frequency). Higher values indicate better overall performance.

Synthesis of findings: Three conclusions emerge from all scenarios. First, HMH-TD3 is the only controller that consistently maintains a stable and well-balanced performance envelope (Figure 8). Second, its advantage over rule-based controllers increases with scenario complexity, becoming the most pronounced under asymmetric and stochastic demands. Third, improvements over single-head TD3 indicate that the multi-head attention-based architecture enhances robustness and stability. While part of this gain may stem from increased model capacity, the consistent reduction in performance variance and improved behavioral stability suggest that structural policy decomposition plays a central role. These observations are further consolidated in Figure 8, where HMH-TD3 consistently achieves the highest composite performance across all demand scenarios.

Overall, HMH-TD3 is best interpreted as a deployment-oriented TSC framework that prioritizes stability, safety, and predictability over throughput maximization. This reflects a deliberate stability–throughput trade-off, where moderate throughput reductions are accepted in exchange for reduced congestion risk and more consistent performance under varying demand conditions.

7. CONCLUSION AND FUTURE WORK

This study presented HMH-TD3, a structured hybrid multi-head RL framework with intrinsic structural interpretability for adaptive TSC. Rather than proposing a new RL algorithm, this work addresses a structural limitation of continuous-control formulations: the absence of explicit behavioral decomposition at the action-selection level. By combining parameterized hybrid actions with a multi-head actor architecture and stabilized TD3 learning, the framework enables robust and diagnostically interpretable control under non-stationary traffic conditions.

The proposed semi-differentiable learning formulation reconciles discrete signal phase selection φ_t with continuous green-time duration Δ_t in a manner consistent with the physical constraints of traffic signal operation. By restricting gradient propagation to the continuous control component while evaluating discrete decisions through value-based selection, the proposed approach retains the benefits of continuous actor–critic optimization while preserving operational validity and safety.

Extensive experimental evaluation in a microscopic simulation environment demonstrated that HMH-TD3 achieves a well-balanced performance across multiple traffic metrics. The results highlight that maximizing a single metric, such as throughput, does not necessarily lead to desirable traffic behavior under dynamic conditions. In particular, higher throughput observed in some baselines under stochastic demand is associated with elevated saturation levels and increased control instability. In contrast, the proposed controller explicitly promotes a balanced optimization regime, prioritizing delay reduction, queue stabilization, and bounded saturation. These findings are consistently supported by the multi-metric analysis presented in Figures 7-9.

In congestion-prone and highly non-stationary scenarios, the proposed controller achieves substantial delay reductions—up to approximately 60%—while maintaining bounded saturation levels and stable phase-switching behavior.

These results reflect a deliberate stability–throughput trade-off, rather than a limitation of the proposed method, where spillback prevention and sustainable traffic operation are favored over aggressive short-term throughput maximization.

In addition to its empirical performance, the proposed architecture provides intrinsic structural interpretability, enabling diagnostic inspection of policy behavior. The multi-head actor facilitates the emergence of distinct behavioral primitives, while the attention mechanism reflects how these behaviors are weighted across varying traffic conditions. This property is structural rather than causal, providing a diagnostic insight into policy behavior in safety-critical traffic control applications. The consistency between internal policy behavior (Figure 6) and external performance (Figures 7-9) further supports the validity of the proposed framework.

This study focused on isolated intersections with complete observability. Future work will extend the framework to coordinated multi-intersection settings, incorporate connected-vehicle information, and investigate the integration of formal safety constraints. Validation through hardware-in-the-loop experimentation and field deployment is a necessary step toward real-world adoption. Finally, while the proposed multi-head architecture demonstrates clear advantages in stability and robustness, part of the observed performance gains may be influenced by increased model capacity. A parameter-matched comparison with scaled single-head architectures remains an important direction for future work to better isolate the contribution of architectural design from model capacity.

REFERENCES

[1] Zhao, H., Dong, C., Cao, J., Chen, Q. (2024). A survey on deep reinforcement learning approaches for traffic signal control. Engineering Applications of Artificial Intelligence, 133: 108100. https://doi.org/10.1016/j.engappai.2024.108100

[2] Michailidis, P., Michailidis, I., Lazaridis, C.R., Kosmatopoulos, E. (2025). Traffic signal control via reinforcement learning: A review on applications and innovations. Infrastructures, 10(5): 114. https://doi.org/10.3390/infrastructures10050114

[3] Liang, X., Du, X., Wang, G., Han, Z. (2019). A deep

reinforcement learning network for traffic light cycle control. IEEE Transactions on Vehicular Technology, 68(2): 1243-1253. https://doi.org/10.1109/TVT.2018.2890726

[4] Bálint, K., Tamás, T., Tamás, B. (2022). Deep reinforcement learning based approach for traffic signal control. Transportation Research Procedia, 62: 278-285. https://doi.org/10.1016/j.trpro.2022.02.035

[5] Chu, T., Wang, J., Codecà, L., Li, Z. (2020). Multi-agent deep reinforcement learning for large-scale traffic signal control. IEEE Transactions on Intelligent Transportation Systems, 21(3): 1086-1095. https://doi.org/10.1109/TITS.2019.2901791

[6] Yoon, J., Ahn, K., Park, J., Yeo, H. (2021). Transferable traffic signal control: Reinforcement learning with graph centric state representation. Transportation Research Part C: Emerging Technologies, 130: 103321. https://doi.org/10.1016/j.trc.2021.103321

[7] Zheng, G., Xiong, Y., Zang, X., Feng, J., Wei, H., Zhang, H., Li, Y., Xu, K., Li, Z. (2019). Learning phase competition for traffic signal control. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM'19), Beijing, China, pp. 1963-1972. https://doi.org/10.1145/3357384.3357900

[8] Fereidooni, Z., Palesi, L.A.I., Nesi, P. (2025). Multi-agent optimizing traffic light signals using deep reinforcement learning. IEEE Access, 13: 106974-106988. https://doi.org/10.1109/ACCESS.2025.3578518

[9] Bouktif, S., Cheniki, A., Ouni, A., El-Sayed, H. (2025). Parameterized-action based deep reinforcement learning for intelligent traffic signal control. Engineering Applications of Artificial Intelligence, 159(Part A): 111422. https://doi.org/10.1016/j.engappai.2025.111422

[10] Sekiyama, K., Nakanishi, J., Takagawa, I., Higashi, T., Fukuda, T. (2001). Self-organizing control of urban traffic signal network. In 2001 IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace (Cat. No. 01CH37236), Tucson, AZ, USA, pp. 2481-2486. https://doi.org/10.1109/ICSMC.2001.972930

[11] Gershenson, C. (2005). Self-organizing traffic lights. Complex Systems, 16(1): 29-53. https://doi.org/10.25088/ComplexSystems.16.1.29

[12] Gershenson, C., Rosenblueth, D.A. (2012). Self-organizing traffic lights at multiple-street intersections. Complexity, 17(4): 23-39. https://doi.org/10.1002/cplx.20392

[13] Gershenson, C., Helbing, D. (2015). When slower is faster. Complexity, 21(2): 9-15. https://doi.org/10.1002/cplx.21736

[14] Li, L., Lv, Y., Wang, F.Y. (2016). Traffic signal timing via deep reinforcement learning. IEEE/CAA Journal of Automatica Sinica, 3(3): 247-254. https://doi.org/10.1109/JAS.2016.7508798

[15] Cai, C., Wei, M. (2024). Adaptive urban traffic signal control based on enhanced deep reinforcement learning. Scientific Reports, 14: 14116. https://doi.org/10.1038/s41598-024-64885-w

[16] Wei, H., Xu, N., Zhang, H., Zheng, G., Chen, X.Y., Li, Z. (2019). CoLight: Learning network-level cooperation for traffic signal control. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM), Beijing, China, pp. 1913-1922. https://doi.org/10.1145/3357384.3357900

[17] Paul, S., Mitra, S. (2022). Deep reinforcement learning-based cooperative control of traffic signals for multi-intersection networks using edge computing. Transactions on Emerging Telecommunications Technologies, 33(11): e4588. https://doi.org/10.1002/ett.4588

[18] Zhu, L., Peng, P., Lu, Z., Tian, Y. (2023). MetaVIM: Meta-variationally intrinsic motivated reinforcement learning for decentralized traffic signal control. IEEE Transactions on Knowledge and Data Engineering, 35(11): 11570-11584. https://doi.org/10.1109/TKDE.2022.3232711

[19] Chen, W., Yang, S., Li, W., Hu, Y., Liu, X., Gao, Y. (2024). Learning multi-intersection traffic signal control via coevolutionary multi-agent reinforcement learning. IEEE Transactions on Intelligent Transportation Systems, 25(11): 15947-15963. https://doi.org/10.1109/TITS.2024.3410023

[20] Schreiber, L.V., Alegre, L.N., Bazzan, A.L.C., Ramos, G.O. (2022). On the explainability and expressiveness of function approximation methods in RL-based traffic signal control. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), Padua, Italy, pp. 1-8. https://doi.org/10.1109/IJCNN55064.2022.9892644

[21] Zhu, Y., Yin, X., Chen, C. (2022). Extracting decision tree from trained deep reinforcement learning in traffic signal control. IEEE Transactions on Computational Social Systems, 10(4): 1997-2007. https://doi.org/10.1109/ICCSI52319.2021.00009

[22] Jin, J., Xing, S., Ji, E., Liu, W. (2025). XGate: Explainable reinforcement learning for transparent and trustworthy API traffic management in IoT sensor networks. Sensors, 25(7): 2183. https://doi.org/10.3390/s25072183

[23] Korecki, M. (2023). Deep reinforcement meta-learning and self-organization in complex systems: Applications to traffic signal control. Entropy, 25(7): 982. https://doi.org/10.3390/e25070982

[24] Gong, T., Zhu, L., Yu, F. R., Tang, T. (2023). Edge intelligence in intelligent transportation systems: A survey. IEEE Transactions on Intelligent Transportation Systems, 24(9): 8919-8944. https://doi.org/10.1109/TITS.2023.3275741

[25] Chen, C., Wei, H., Xu, N., Zheng, G., Yang, M., Xiong, Y., Xu, K., Li, Z. (2020). Toward a thousand lights: Decentralized deep reinforcement learning for large-scale traffic signal control. In Proceedings of the AAAI Conference on Artificial Intelligence, pp. 3414-3421. https://doi.org/10.1609/aaai.v34i04.5744

[26] Elharoun, A., Abid, M., Chehri, A. (2025). Adaptive traffic signal control using deep reinforcement learning: A multi-objective approach for single and multi-intersection scenarios. Expert Systems with Applications, 227: 120381. https://doi.org/10.1016/j.eswa.2024.120381

[27] Fujimoto, S., Hoof, H., Meger, D. (2018). Addressing function approximation error in actor-critic methods. In Proceedings of the 35th International Conference on Machine Learning, PMLR 80, pp. 1587-1596. https://proceedings.mlr.press/v80/fujimoto18a.html.

[28] Gao, J., Shen, Y., Ito, M., Shiratori, N. (2024). Dual

experience replay-based TD3 for single intersection signal control. Engineering Applications of Artificial Intelligence, 128: 107302. https://doi.org/10.1016/j.engappai.2023.107302

[29] Prathiba, S.B., Raja, G., Dev, K., Kumar, N., Guizani, M. (2021). A hybrid deep reinforcement learning for autonomous vehicles smart-platooning. IEEE Transactions on Vehicular Technology, 70(12): 13340-13350. https://doi.org/10.1109/TVT.2021.3122257

[30] Feng, S., Li, X., Ren, L., Xu, S. (2023). Reinforcement learning with parameterized action space and sparse reward for UAV navigation. Intelligent Robotics, 3: 161-175. https://doi.org/10.20517/ir.2023.10

[31] Hausknecht, M., Stone, P. (2016). Deep reinforcement learning in parameterized action space. In Proceedings of the International Conference on Learning Representations (ICLR), San Juan, Puerto Rico, USA, pp. 1-12. https://www.cs.utexas.edu/~AustinVilla/details/ICLR16-hausknecht.html.

[32] Luo, H., Bie, Y., Jin, S. (2024). Reinforcement learning for traffic signal control in hybrid action space. IEEE Transactions on Intelligent Transportation Systems, 25(6): 5225-5241. https://doi.org/10.1109/TITS.2023.3344585

[33] Hu, Y., Du, L., Easa, S.M. (2025). Explainable reinforcement learning for improved traffic signal control. Computer-Aided Civil and Infrastructure Engineering, 40(24): 3911-3933. https://doi.org/10.1111/mice.70037

[34] Zambaldi, V., Raposo, D., Santoro, A., Bapst, V., Li, Y., Babuschkin, I., Tuyls, K., Reichert, D., Lillicrap, T., Lockhart, E., Shanahan, M., Langston, V., Pascanu, R., Botvinick, M., Vinyals, O., Battaglia, P. (2019). Deep reinforcement learning with relational inductive biases. In Proceedings of the International Conference on Learning Representations (ICLR), pp. 1-18.

[35] Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., Meger, D. (2018). Deep reinforcement learning that matters. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1). https://doi.org/10.1609/aaai.v32i1.11694

[36] Behrisch, M., Bieker, L., Erdmann, J., Krajzewicz, D. (2011). SUMO–simulation of urban mobility: An overview. In Proceedings of SIMUL 2011, the Third International Conference on Advances in System Simulation, Barcelona, Spain, pp. 55-60. https://elib.dlr.de/71460/.

[37] Lopez, P.A., Behrisch, M., Bieker-Walz, L., Erdmann, J., et al. (2018). Microscopic traffic simulation using SUMO. In 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, pp. 2575-2582. https://doi.org/10.1109/ITSC.2018.8569938

[38] Krajzewicz, D. (2010). Traffic simulation with SUMO–simulation of urban mobility. In Fundamentals of Traffic Simulation, pp. 269-293. https://doi.org/10.1007/978-1-4419-6142-6_7

[39] Han, Y., Wang, M., Leclercq, L. (2023). Leveraging reinforcement learning for dynamic traffic control: A survey and challenges for field implementation. Communications in Transportation Research, 3: 100104. https://doi.org/10.1016/j.commtr.2023.100104

[40] van Hasselt, H. (2010). Double Q-learning. Advances in Neural Information Processing Systems, 23: 2613-2621. https://proceedings.neurips.cc/paper_files/paper/2010/file/091d584fced301b442654dd8c23b3fc9-Paper.pdf.

[41] Agarwal, R., Schwarzer, M., Castro, P.S., Courville, A., Bellemare, M.G. (2021). Deep reinforcement learning at the Edge of the statistical precipice. Advances in Neural Information Processing Systems, 35: 29304-29320. https://dl.acm.org/doi/10.5555/3540261.3542505

[42] Osband, I., Blundell, C., Pritzel, A., Van Roy, B. (2016). Deep exploration via bootstrapped DQN. Advances in Neural Information Processing Systems, 29: 4026-4034.

[43] Transportation Research Board and National Academies of Sciences, Engineering, and Medicine. (2022). Highway Capacity Manual 7th Edition: A Guide for Multimodal Mobility Analysis. The National Academies Press, Washington, DC, USA. https://doi.org/10.17226/26432