# LLM-GUIDED MULTI-AGENT REINFORCEMENT LEARNING FOR EMERGENCY TRAFFIC SIGNAL CONTROL

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Emergency vehicle (EMV) signal preemption requires coordinated multi-agent decisions under uncertain arrival information and changing traffic environments. We propose a large language models (LLM)-guided multi-agent reinforcement learning (MARL) framework integrating probabilistic perception, preference formation, and action grounding. A probabilistic long short-term memory (LSTM) estimates EMV arrival times and uncertainty, and an LLM converts this temporal context and multi-agent traffic state into phase preference distributions that modulate learned Q-values. An uncertainty-aware grounded action transformation (UGAT) reduces the impact of environment-dependent dynamics mismatch on action selection. Experiments on three real-world traffic networks show consistent improvements over MARL and traffic-control baselines. Under heavy congestion, the method reduces overall and EMV travel time by 24.96% and 44.88%, respectively, while improving traffic stability. Cross-environment experiments demonstrate robustness to changes in network structure and traffic demand.

## 1 INTRODUCTION

Urban traffic signal control must balance network-wide traffic efficiency and emergency response under dynamic conditions. Emergency vehicle (EMV) preemption is challenging because rapid EMV passage must be achieved without excessive queue spillback or secondary congestion. Early rule-based strategies trigger priority phases upon EMV detection, but often perform poorly under dense and changing traffic, motivating adaptive learning-based approaches.

Multi-agent reinforcement learning (MARL) has shown strong potential for large-scale signal coordination by learning distributed policies through interactions with dynamic traffic environments. However, existing EMV-oriented MARL methods face three challenges. First, reliance on instantaneous observations limits anticipation of EMV arrivals, whose trajectories depend on temporal interactions among queues, signal phases, and traffic dynamics, potentially causing priority actions to activate too early or too late. Second, learned policies may struggle to balance EMV priority and regular traffic efficiency under changing conditions, while macroscopic traffic context and EMV operational states are difficult to encode directly in conventional RL observations. This motivates large language models (LLMs)-based high-level phase preferences. Third, policies trained in one simulation environment may degrade under different traffic dynamics because changes in vehicle interactions and state transitions create training-deployment discrepancies. Existing transfer methods often rely on indirect adjustments without explicitly addressing these transition mismatches.

To address these challenges, a hierarchical MARL framework is proposed for EMV-oriented traffic signal control. The framework integrates three tightly coupled components: (i) probabilistic LSTM-based state encoding for EMV arrival estimation; (ii) MARL-based signal optimization with low-frequency LLM-generated phase preferences for balancing EMV priority and regular traffic efficiency; and (iii) uncertainty-aware grounded action transformation (UGAT) for correcting action deviations under cross-environment dynamics mismatch.

The main contributions of this work are summarized as follows: (1) A probabilistic temporal encoding mechanism is developed to estimate EMV arrival states from historical trajectories, enabling proactive signal control under uncertain arrivals. (2) An LLM-guided preference reasoning frame-



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

work is developed to improve coordination between EMV priority and regular traffic efficiency in multi-objective traffic control. (3) A UGAT module is designed to mitigate cross-environment dynamics mismatch and improve transfer robustness. (4) Extensive experiments conducted across three real-world traffic networks demonstrate the effectiveness and transferability of the proposed framework under diverse traffic conditions.

## 2 RELATED WORK

**EMV-oriented traffic signal control.** MARL has become a prominent framework for large-scale urban traffic signal control because it supports decentralized decision-making and inter-intersection coordination (Katzilieris et al., 2026; Ren et al., 2025). Recent studies further enhance spatial coordination using graph neural networks (GNNs), enabling neighboring intersections to exchange traffic information through structured message passing (Zhai et al., 2025). These methods provide a foundation for network-level control under regular traffic conditions.

EMV-oriented control introduces additional temporal challenges, as EMV trajectories are affected by queue dissipation, phase switching, and priority constraints under congestion (Bairi et al., 2025; Yang et al., 2025). Existing studies often encode EMVs using current positions, intersection distances, or lane-level indicators as GNN features. While capturing spatial proximity, these features provide limited temporal information, leaving movement evolution and arrival uncertainty insufficiently represented. This limits proactive anticipation of EMV arrivals.

EMV preemption also requires balancing emergency priority with regular traffic efficiency. Existing MARL methods mainly address this trade-off through reward shaping, such as penalizing EMV delay or increasing emergency-objective weights (Benhamza et al., 2024; Cai et al., 2025; Nie et al., 2025; Su et al., 2022; 2023). However, fixed weights impose a predefined priority structure that may not adapt to changing traffic conditions, potentially causing excessive or insufficient preemption. This motivates adaptive high-level guidance for multi-objective signal coordination.

**LLM-Guided Reinforcement Learning for Traffic Control.** Recent studies have explored LLMs as contextual guidance for sequential decision-making, including traffic control and related optimization problems (Huang et al., 2026; Jovanović et al., 2026; Yao et al., 2025). This is particularly relevant to EMV-oriented signal control, where decisions must jointly consider local traffic states and changing emergency priorities. Early studies used LLMs as direct controllers by converting traffic observations into text and generating signal phases or green-time allocations (Lai et al., 2025; Movahedi & Choi, 2025). Despite their flexibility, high inference latency and output inconsistency limit their suitability for real-time EMV preemption.

Recent work therefore have turned to LLM-guided RL frameworks (Pternea et al., 2024; Wang et al., 2025). Existing methods mainly follow two directions. The first uses LLMs only during training, such as for reward refinement (Choi & Lim, 2025), injecting prior knowledge without online inference cost but removing LLM guidance during deployment. The second retains LLMs online for feature interpretation (Chen et al., 2025; Da et al., 2024), translating numerical traffic states into text and then numerical embeddings for policy learning. While this enriches state representation, action selection remains dominated by conventional neural policies, limiting the direct influence of high-level semantic preferences.

More tightly coupled LLM-RL architectures therefore combine fast RL decision-making with slower semantic guidance: RL agents perform real-time signal optimization, while LLMs provide preferences reflecting EMV behavior and network traffic conditions. However, their integration with EMV-oriented MARL remains limited, and robustness under changing traffic environments has received insufficient attention.

**Cross-environment transfer for RL-based traffic control.** Practical deployment of RL-based traffic signal control requires robustness to changing traffic dynamics. For EMV-oriented control, differences in car-following, lane-changing, and simulator mechanisms can alter the state-transition process encountered by learned policies (Da et al., 2023; Tiwari et al., 2026). Cross-simulator transfer therefore provides a useful test of robustness under environment shifts (Bokade & Jin, 2025; Mei et al., 2024).

Domain randomization improves transferability by exposing agents to diverse source-domain traffic conditions (Müller & Sabatelli, 2023). However, extensive randomization increases sample complexity, while independent perturbations may fail to preserve nonlinear dependencies in realistic traffic dynamics, leaving a mismatch between training and deployment transitions.

Safe RL provides another approach by constraining actions or penalizing unsafe behavior (Zhou et al., 2025). Existing methods mainly use execution-time action filtering, such as rule-based masking (Müller & Sabatelli, 2022), or training-time Lagrangian regularization in constrained MDPs (Gao et al., 2025). Although these methods improve policy stability, they do not explicitly model source-target transition mismatch and may remain vulnerable to substantial environment shifts (Gu et al., 2024). This motivates transfer mechanisms that incorporate environment-dependent transition dynamics into action selection.

## 3 Problem formulation

The road network is modeled as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $i \in \mathcal{V}$, with $\mathcal{V} = \{1, \dots, N\}$, is a signalized intersection, $(i, j) \in \mathcal{E}$ a directed road segment. A control agent is associated with each intersection, and neighboring intersections $\mathcal{N}(i) = \{j \mid (i, j) \in \mathcal{E} \vee (j, i) \in \mathcal{E}\}$ are those directly connected through traffic flows. At each discrete decision step, every agent selects one feasible signal phase. Multiple EMVs enter from different approaches along predefined routes; at each intersection only the active EMV whose route passes through it with the shortest remaining distance is considered. Each agent takes one of three roles encoded by $\rho_{i,t} \in \{0, 1, 2\}$: normal, primary pre-emption, or secondary and further. A normal agent serves no active EMV route and focuses on local congestion, a primary agent is the intersection currently occupied by an EMV, and a secondary agent lies downstream on the EMV route and must clear the corridor in advance.

The objective trades EMV progression against regular traffic efficiency: timely pre-emption cuts EMV delay, whereas premature or excessive pre-emption disrupts regular traffic and propagates congestion to adjacent intersections. Under partial local observations and spatially coupled dynamics, we cast the task as a networked decentralized partially observable Markov decision process (Networked Dec-POMDP) $\langle \mathcal{G}, \mathcal{S}, \mathcal{O}, \mathcal{A}, Z, R, \Omega, \gamma \rangle$, where $\mathcal{S}$ is the global state, $\mathcal{O}_i$ and $\mathcal{A}_i$ the local observation and action spaces, $Z$ the transition function, $R$ the reward function, $\Omega$ the observation function and $\gamma$ the discount factor. At step $t$, agent $i$ conditions only on $\mathcal{O}_i$, and the joint action induces the next state and reward. The goal is a set of decentralized policies that minimize EMV delay without degrading network-wide efficiency. Full tuple details are given in Appendix A.

## 4 Methodology

The framework couples four modules in a closed loop (Figure 1). A probabilistic LSTM estimates EMV arrival time together with its predictive uncertainty and injects the resulting temporal features into each agent's observation. A MADQN backbone then selects signal phases at every decision step, while a low-frequency LLM converts congestion and EMV-progression context into phase-level preferences that modulate the learned Q-values by residual fusion. Before execution, a UGAT module grounds the selected action in the target environment's transition dynamics. Perception, preference formation and action grounding are therefore coupled end-to-end rather than trained in isolation.

### 4.1 Probabilistic EMV arrival perception

Reliable estimated time of arrival (ETA) is essential for fine-grained pre-emption. Distance–speed heuristics assume constant motion and underestimate arrival time whenever upstream congestion persists, because queue blocking and phase transitions add interaction delay. We therefore predict ETA with an LSTM over a sliding window of traffic features — EMV speed and remaining distance to the stop line, headway and leader speed, and lane queue length with the current phase — and adopt heteroscedastic regression to output both the mean ETA $\mu_t$ and its predictive variance $\sigma_t^2$, so that predictive uncertainty grows with traffic variability. Feature construction detail are in Appendix B. A dual-head output layer predicts the mean $\mu_t$ and logarithm of the variance $\log(\sigma_t^2)$ of the

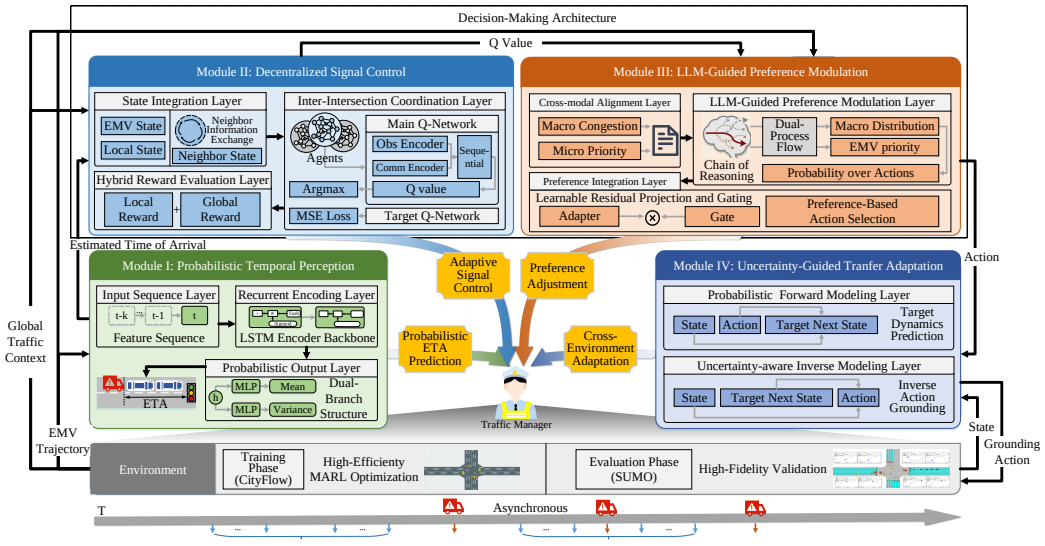



Figure 1: Overall architecture of the proposed EMV-oriented traffic signal control framework.

arrival-time distribution:

$$\left[\mu_t,\, \log(\sigma_t^2)\right] = \mathbf{MLP}(\mathbf{h}_t) \tag{1}$$

The model is trained using the Gaussian negative log-likelihood (GNLL):

$$L_{\text{GNLL}} = \frac{1}{B}\sum_{b=1}^{B}\left(\frac{1}{2}\log(\sigma_b^2) + \frac{(\tau_b^* - \mu_b)^2}{2\sigma_b^2}\right) \tag{2}$$

where $B$ is the batch size, $\tau_b^*$ is the ground-truth arrival time, and $\mu_b$, $\sigma_b^2$ are the predicted mean and variance for the $b$-th instance. The objective jointly optimizes ETA accuracy and uncertainty calibration, allowing the model to represent lower confidence under more variable traffic conditions.

The resulting probabilistic representation is incorporated into the MADQN observation space as an additional temporal feature, enabling downstream agents to incorporate both predicted arrival time and its uncertainty into decision-making.

## 4.2 MADQN-BASED TRAFFIC SIGNAL CONTROL

### 4.2.1 OBSERVATION SPACE AND COMMUNICATION MECHANISM

For intersection $i$, the observation $\mathbf{o}_{i,t}$ at time step $t$ consists of a local observation $\mathbf{o}_{i,t}^{\text{local}}$ and an aggregated neighbor representation $\mathbf{M}_{i,t}^{\mathcal{N}}$. The local observation contains lane-wise vehicle counts, the signal phase executed at the previous time step, and EMV-related features:

$$\mathbf{o}_{i,t}^{\text{local}} = \left[v_{i,t}^{(1)}, \ldots, v_{i,t}^{(\mathcal{K})},\, \varsigma_{i,t-1},\, \rho_{i,t},\, d_{i,t},\, \mu_{i,t},\, \log(\sigma_{i,t}^2)\right] \tag{3}$$

where, $v_{i,t}^{(k)}$ denotes the vehicle counts on the $k$-th lane, and $\varsigma_{i,t-1}$ represents the one-hot encoding of the previous signal phase. For EMV Features, $d_{i,t}$ represents the normalized relative distance of the EMV to the intersection's stop line.

Because each agent observes only its local state, neighboring information is incorporated through an attention-based communication encoder. The observations of adjacent intersections $j \in \mathcal{N}(i)$ are first projected into latent representations and then aggregated using multi-head self-attention. The resulting features are processed with residual connections and layer normalization, followed by mean pooling and concatenation with the local observation. The detailed attention formulation is provided in Appendix C.1.

#### 4.2.2 ROLE-ADAPTIVE REWARD DESIGN

The reward is designed according to the agent's topological role defined in Section 3. To balance regular traffic efficiency and EMV progression, the local traffic state is characterized by traffic pressure $\mathcal{P}_{i,t}$ and lane occupancy $U_{i,t}$. Their detailed definitions are provided in Appendix C.3.

The local reward is determined by the agent role:

$$r_{i,t}^{\text{local}} = \begin{cases} -\mathcal{P}_{i,t}, & \rho_{i,t} = 0, \\ -1, & \rho_{i,t} = 1, \\ -\big(\alpha \cdot \mathcal{P}_{i,t} + \beta \cdot U_{i,t}\big), & \rho_{i,t} = 2. \end{cases} \tag{4}$$

A normal agent minimizes local congestion through the negative traffic-pressure reward. A primary pre-emption agent receives a fixed large penalty during EMV occupancy. For secondary and further agents, the reward interpolates between EMV-route clearance and regular traffic efficiency according to the predefined role weight. The coefficients are determined by parameter $\chi$ and $\omega$: $\alpha = 1-\omega(1-\chi)$ and $\beta = \omega(1-\chi)$. The resulting local rewards are normalized and averaged to obtain the global reward:

$$r_t^{\text{global}} = \frac{1}{N} \sum_{j \in \mathcal{V}} r_{j,t}^{\text{local}} \tag{5}$$

The final reward of agent $i$ is the weighted combination of its local and global rewards:

$$r'_{i,t} = \lambda_1 \, r_{i,t}^{\text{local}} + \lambda_2 \, r_t^{\text{global}} \tag{6}$$

where $\lambda_1, \lambda_2$ represent the predefined weighting coefficients.

#### 4.2.3 NETWORK OPTIMIZATION AND TRAINING OBJECTIVE

For each intersection agent $i$, the Q-network $\big(Q_i(\mathbf{o}_{i,t}, \cdot; \theta_i)\big)$ estimates the values of all feasible signal phases from the observation defined above. The agents are trained using independent experience replay and a target network, with the temporal-difference objective optimized by mini-batch gradient descent. The complete TD target, loss function, RMSprop update, and target-network synchronization are provided in Appendix C.4.

### 4.3 LLM-GUIDED PREFERENCE MODULATION FOR MULTI-OBJECTIVE COORDINATION

The LLM-guided preference module provides contextual guidance to the numerical MADQN policy without replacing its real-time decision process, as illustrated in Figure 2. The module operates at a lower frequency and follows three steps: numerical traffic observations are converted into structured textual descriptions, the LLM generates phase-level preference signals, and the preferences are integrated into MADQN action values through residual fusion.

#### 4.3.1 TRAFFIC-TO-TEXT ALIGNMENT

To provide the LLM with structured traffic context, lane-level queues are aggregated into phase-level congestion indicators, while each active EMV is represented by its speed, remaining distance, the speed of the preceding vehicle and the distance from the preceding vehicle, current intersection, planned route and the feasible phase sets at the current and downstream intersections. These variables are converted into structured prompts, with downstream route information included when applicable. The detailed phase-level aggregation is provided in Appendix D.

#### 4.3.2 PHASE-LEVEL PREFERENCE GENERATION

Given the structured prompts, the LLM generates a phase-level preference vector $\boldsymbol{\pi}_{i,t}^{\text{sem}} = \big[\pi_{i,t}^{(1)}, \ldots, \pi_{i,t}^{(P)}\big]$, where $P$ is the number of candidate phases and larger values indicate stronger relative preference.

Two control mechanisms are incorporated into the preference generation. First, phase-level queue distributions are considered jointly to avoid concentrating green time on a single congested movement. When queue imbalance is large, non-dominant phases retain non-zero preference values, producing a more distributed phase selection under uneven demand.

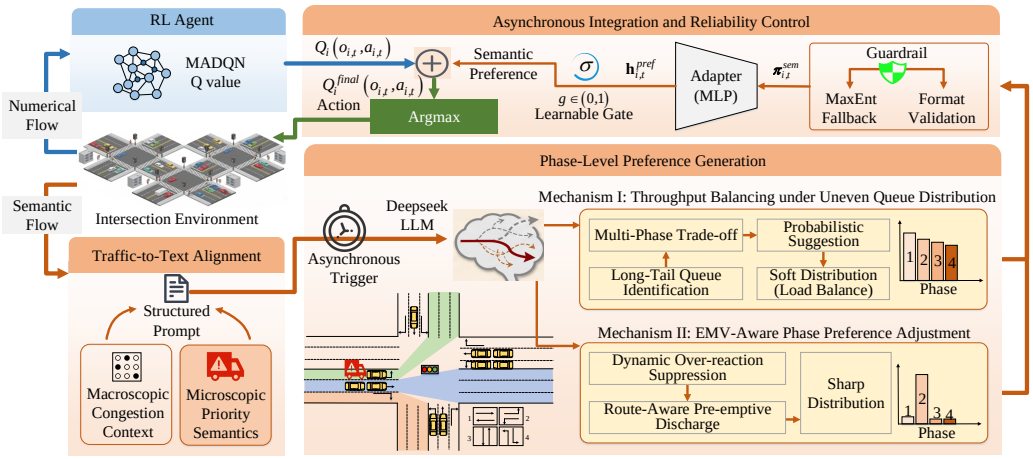



Figure 2: LLM-guided preference modulation framework.

Second, EMV priority is adjusted according to both arrival status and route position. When an EMV remains distant and its approach is not severely blocked, the preference for EMV-related phases is reduced to avoid premature pre-emption. As the EMV approaches, stronger priority is assigned to the corresponding phases. Downstream intersections along the planned EMV route are also considered, with higher preference assigned to phases that facilitate downstream clearing and reduce congestion propagation along the EMV path.

#### 4.3.3 Asynchronous Integration and Reliability Control

MADQN selects signal actions at every decision step, whereas the LLM is invoked only at predefined intervals or EMV-related events, avoiding dependence on LLM inference latency. The generated preference vector is first constrained to the feasible phase set; invalid outputs are regenerated, with a uniform distribution used as the fallback. The validated preference is projected into the Q-value space through an MLP:

$$\mathbf{h}_{i,t}^{\text{pref}} = \mathbf{W}_2\,\text{ReLU}\big(\mathbf{W}_1\boldsymbol{\pi}_{i,t}^{\text{sem}} + \mathbf{b}_1\big) + \mathbf{b}_2 \tag{7}$$

A gating factor $g \in (0,1)$ controls the contribution of the semantic signal, and the final action values are obtained by residual fusion:

$$Q_i^{\text{final}}(\mathbf{o}_{i,t}, a_{i,t}) = Q_i(\mathbf{o}_{i,t}, a_{i,t}; \theta_i) + g \cdot \mathbf{h}_{i,t}^{\text{pref}}\big[a_{i,t}\big] \tag{8}$$

Thus, the LLM provides phase-level guidance through an auxiliary residual term, while the MADQN Q-values remain the basis for action selection.

### 4.4 Uncertainty-Guided Action Grounding for Cross-Environment Transfer

The source environment $\mathcal{M}_{\text{src}}$ and target environment $\mathcal{M}_{\text{tgt}}$ may exhibit inconsistent state transition functions $Z_{\text{src}}(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t) \neq Z_{\text{tgt}}(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)$, leading to divergent next-state distributions under the same policy actions. UGAT addresses this mismatch by grounding policy actions using target-environment dynamics.

A forward-inverse model is constructed from source- and target-environment interaction data. A probabilistic forward model predicts the target-domain next observation from the current observation and action:

$$\hat{\mathbf{o}}_{i,t+1}^{\text{tgt}} = f_\phi\big(\mathbf{o}_{i,t}, a_{i,t}\big) \tag{9}$$

While an inverse model estimates the categorical distribution of actions that can produce the predicted transition:

$$\hat{Z}_{\text{inv}}\big(a \mid \mathbf{o}_{i,t}, \hat{\mathbf{o}}_{i,t+1}^{\text{tgt}}\big) = \text{Softmax}\big(h_\phi(\mathbf{o}_{i,t}, \hat{\mathbf{o}}_{i,t+1}^{\text{tgt}})\big) \tag{10}$$

The most probable action is selected as the grounded action:

$$\hat{a}_{i,t}^{g} = \arg\max_{a \in \mathcal{A}_i} \hat{Z}_{\text{inv}}\big(a \mid \mathbf{o}_{i,t}, \hat{\mathbf{o}}_{i,t+1}^{\text{tgt}}\big) \tag{11}$$

The reliability of this mapping is measured by the Shannon entropy:

$$u_{i,t} = \mathcal{H}\big(\hat{Z}_{\text{inv}}\big) = -\sum_{a \in \mathcal{A}_i} \hat{Z}_{\text{inv}}\big(a \mid \mathbf{o}_{i,t}, \hat{\mathbf{o}}_{i,t+1}^{\text{tgt}}\big) \log \hat{Z}_{\text{inv}}\big(a \mid \mathbf{o}_{i,t}, \hat{\mathbf{o}}_{i,t+1}^{\text{tgt}}\big) \tag{12}$$

The procedures and the selective grounding strategy are provided in Appendix E.

The complete framework is trained using a progressive strategy. MADQN is first pretrained as a stable baseline, after which the temporal perception and LLM-guided preference modules are sequentially integrated and fine-tuned. UGAT is subsequently trained from source–target interaction data and introduced during source-environment retraining, followed by deployment in SUMO for cross-environment evaluation. Detailed training procedures are provided in Appendix F.

# 5 Experiments

## 5.1 Experimental setup

CityFlow is used for training and SUMO for cross-environment evaluation. Five scenarios are constructed from three real-world networks: Jinan Dongfeng (4×3, 12 intersections), Hangzhou Gudang (4×4, 16 intersections), and Manhattan Upper East Side (16×3, 48 intersections). Jinan further includes low-, medium-, and high-demand settings (4365–6295 veh/h). EMV routes cover diverse intersections and turning movements under free-flow and congested conditions; details are given in Appendix G.1 and G.2.

Baselines include Fixed-time, MA2C (Chu et al., 2019), and HGAT (Zhai et al., 2025). Ablations include w/o ETA Refinement, Repl. Preference w/ State, and w/o UGAT Adaptation, removing or replacing the corresponding ETA, preference, and transfer modules, details are given in Appendix G.3. Overall ATT and EMV ATT are the primary metrics; additional traffic, stability, and environmental metrics are reported in Appendix G.4.

## 5.2 Overall performance and ablation analysis

Table 1[1] summarizes the performance of all evaluated methods across five traffic datasets. In the Jinan-high scenario, the proposed method reduces Overall ATT and AWC by 24.96% and 9.03%, respectively, relative to Fixed-time control. It also achieves strong spatial scalability, improving Overall ATT by 28.10% and 63.18% on the Hangzhou and New York datasets, respectively. Compared with advanced MARL baselines, it consistently reduces EMV ATT by 1.65%–26.71%, indicating that probabilistic ETA perception and LLM-guided preference modulation improve coordination between EMV priority and regular traffic efficiency. These gains are accompanied by 0.33%–7.50% lower fuel consumption and $CO_2$ emissions than the best-performing baselines. Ablation results further confirm the contribution of all core modules, with Overall and EMV ATT degrading by 2.08%–6.84% when key components are removed in Jinan-low and by 4.07%–29.91% in Jinan-high, demonstrating greater importance under high-demand conditions. Overall ATT distribution analysis are reported in Appendix G.6.

## 5.3 Spatial scalability and traffic demand adaptability

Figure 3 evaluates vehicle trajectories under low-, medium-, and high-demand conditions in the Jinan network. Under low and medium demand, the proposed method produces more continuous trajectories with fewer stop-and-go interruptions. Under high demand, the baseline exhibits substantial upstream queue spillback, whereas the proposed method maintains smoother discharge trajectories and limits congestion propagation along the corridor. Additional spatial heatmaps and environmental analyses are provided in Appendix G.8.

[1]Note: The best results are highlighted in bold and the second-best results are underlined. Improvement (%) indicates the relative improvement of the proposed method over the strongest baseline.

Table 1: Performance comparison of different traffic signal control methods across five datasets.

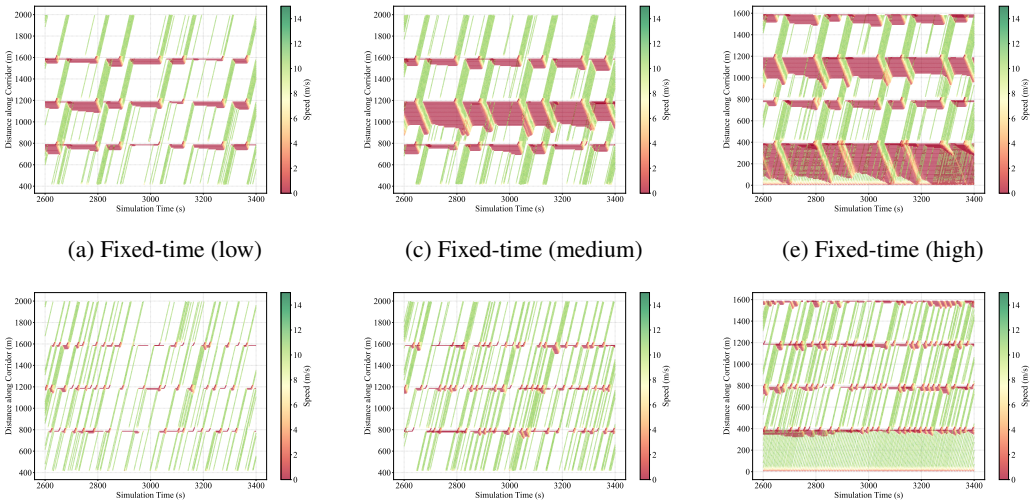
| Model | Metric | Jinan-Low | | Jinan-Medium | | Jinan-High | | Hangzhou | | New York | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Fixed-time | Overall ATT | 383.80 | | 383.11 | | 411.54 | | 513.01 | | 613.74 | |
| | EMV ATT | 198.00 | | 245.44 | | 302.56 | | 246.50 | | 403.33 | |
| | TP | 4048.00 | | 4864.00 | | 5480.00 | | 2549.00 | | 1334.00 | |
| | AWC | 2.22 | | 2.30 | | 2.77 | | 3.16 | | **5.11** | |
| | $CO_2$ | 883.04 | | 885.56 | | 959.71 | | 1153.25 | | 1559.76 | |
| | Fuel | 281.66 | | 282.46 | | 306.11 | | 367.84 | | 497.51 | |
| MA2C | Overall ATT | 372.15 | | 406.98 | | 458.13 | | 458.86 | | 536.98 | |
| | EMV ATT | 182.67 | | 277.22 | | 386.67 | | 200.00 | | 795.60 | |
| | TP | 4092.00 | | 4794.00 | | 5191.00 | | 2660.00 | | 761.00 | |
| | AWC | 5.53 | | 7.10 | | 8.88 | | 4.14 | | 7.89 | |
| | $CO_2$ | 844.39 | | 931.22 | | 1056.05 | | 1005.58 | | 1338.64 | |
| | Fuel | 269.33 | | 297.02 | | 336.84 | | 320.73 | | 426.98 | |
| HGAT | Overall ATT | 292.74 | | 300.90 | | 364.27 | | 383.96 | | 310.89 | |
| | EMV ATT | 153.11 | | 128.00 | | 212.67 | | 210.63 | | 179.50 | |
| | TP | 4137.00 | | 5005.00 | | 5435.00 | | 2693.00 | | 2543.00 | |
| | AWC | 2.45 | | 3.20 | | 6.13 | | 1.73 | | 7.32 | |
| | $CO_2$ | 643.47 | | 667.42 | | 822.06 | | 811.38 | | 794.80 | |
| | Fuel | 205.24 | | 212.88 | | 262.20 | | 258.79 | | 253.51 | |
| w/o ETA Refinement | Overall ATT | 297.75 | | 300.81 | | 323.31 | | 381.03 | | 243.34 | |
| | EMV ATT | 117.00 | | **125.56** | | 207.00 | | 193.25 | | 172.11 | |
| | TP | 4025.00 | | 4994.00 | | 5669.00 | | 2698.00 | | 2647.00 | |
| | AWC | 2.09 | | 2.34 | | 3.03 | | 1.71 | | 5.49 | |
| | $CO_2$ | 656.89 | | 667.96 | | 725.78 | | 801.77 | | 614.61 | |
| | Fuel | 209.52 | | 213.05 | | 231.49 | | 255.73 | | 196.04 | |
| Repl. Preference w/ State | Overall ATT | 311.62 | | 310.60 | | 327.92 | | 407.39 | | 254.35 | |
| | EMV ATT | 119.89 | | 130.00 | | 206.56 | | 191.25 | | 181.33 | |
| | TP | 4124.00 | | 4972.00 | | 5545.00 | | 2614.00 | | 2626.00 | |
| | AWC | 2.29 | | 2.50 | | 3.11 | | 2.37 | | 5.69 | |
| | $CO_2$ | 692.31 | | 693.43 | | 739.41 | | 873.34 | | 643.29 | |
| | Fuel | 220.82 | | 221.18 | | 235.84 | | 278.56 | | 205.18 | |
| w/o UGAT Adaptation | Overall ATT | 301.35 | | 301.11 | | 326.37 | | 384.26 | | 248.59 | |
| | EMV ATT | 114.89 | | 127.44 | | 216.67 | | 216.38 | | 176.22 | |
| | TP | 4133.00 | | 4988.00 | | **5670.00** | | 2693.00 | | 2636.00 | |
| | AWC | 2.12 | | 2.32 | | 3.11 | | **1.55** | | 5.55 | |
| | $CO_2$ | 665.12 | | 668.75 | | 734.24 | | 810.23 | | 628.24 | |
| | Fuel | 212.15 | | 213.30 | | 234.19 | | 258.42 | | 200.38 | |
| **Ours** | Overall ATT | **291.67** | 0.37% | **290.23** | 3.52% | **308.81** | 4.48% | **368.84** | 3.20% | **225.97** | 7.14% |
| | EMV ATT | **112.22** | 2.32% | 125.89 | -0.26% | **166.78** | 19.26% | **184.75** | 3.40% | **159.67** | 7.23% |
| | TP | **4142** | 0.12% | **5022** | 0.34% | 5628 | -0.74% | **2700** | 0.07% | **2659** | 0.45% |
| | AWC | **2.01** | 3.83% | **2.14** | 6.96% | **2.52** | 9.03% | 1.65 | -6.45% | 5.14 | -0.59% |
| | $CO_2$ | **639.59** | 0.60% | **639.36** | 4.20% | **686.93** | 5.35% | **799.12** | 0.33% | **568.50** | 7.50% |
| | Fuel | **204.00** | 0.60% | **203.93** | 4.20% | **219.10** | 5.35% | **254.88** | 0.33% | **181.33** | 7.50% |

### 5.4 Temporal operational stability and congestion mitigation

Figure 4 shows that the proposed method maintains lower queue length, more stable speed, and steadier throughput in the Hangzhou network. These temporal patterns indicate more effective congestion dissipation and stable network operation under sustained demand. Additional intersection-level waiting distributions are provided in Appendix G.7.

## 6 Conclusion

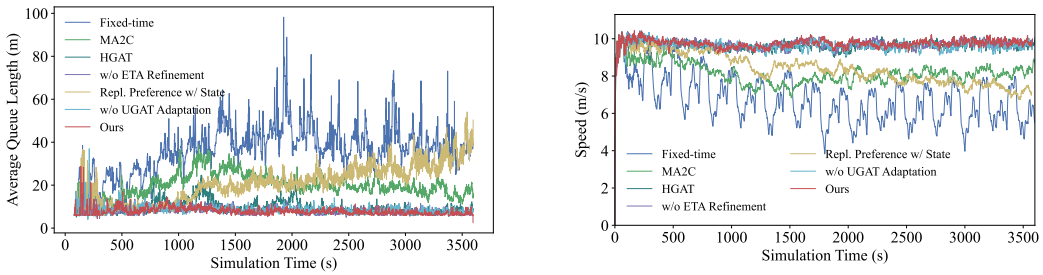
This study presented an LLM-guided MARL framework for EMV-oriented traffic signal control, integrating probabilistic ETA perception, LLM-based phase preferences, and uncertainty-guided action grounding to address temporal uncertainty, multi-objective coordination, and cross-environment dynamics mismatch. Experiments on three real-world traffic networks demonstrate consistent improvements in traffic efficiency and EMV response across demand levels and network scales. Ablation studies further validate the complementary contributions of temporal prediction, phase preference, and UGAT.



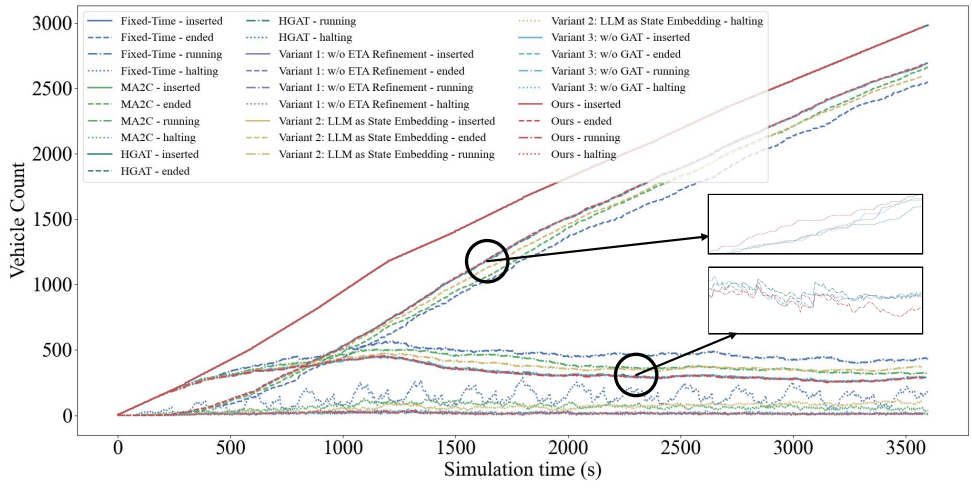

(a) Fixed-time (low) (c) Fixed-time (medium) (e) Fixed-time (high)

(b) Ours (low) (d) Ours (medium) (f) Ours (high)

Figure 3: Vehicle time-space trajectories under different traffic demand levels in the Jinan dataset.




(a) Temporal profile of AQL. (b) Temporal profile of AS.




(c) Vehicle kinematic dynamics including inserted, TP, running, and halting vehicles.

Figure 4: Temporal evolution of traffic operational metrics in the Hangzhou dataset.

### AI USE STATEMENT

In this work, an LLM is used as a component of the proposed method: it converts probabilistic estimates of emergency-vehicle arrival times and multi-agent traffic states into phase preference distributions that modulate learned Q-values. This LLM is the object of study rather than a tool used to conduct the research; its role is explicitly described and evaluated in the manuscript. Separately, generative AI tools were used during manuscript preparation for language editing and wording refinement only. We have not used generative AI tools for any of the tasks requiring disclosure: we did not generate synthetic datasets, develop theoretical or conceptual frameworks, formulate or help prove mathematical claims, propose or refine hypotheses, design or provide feedback on the research methodology or experiments, implement methods or code, clean or reformat datasets, or interpret results. All AI-assisted text was reviewed and verified by the authors. We take responsibility for the final content of this work, including all text, claims, analyses, and artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

We affirm that this work fully adheres to the ICLR Code of Ethics. We do not collect, use, or release any new human subject data. All experiments are conducted on established and publicly available datasets, which have been widely used in prior academic research.

### REPRODUCIBILITY STATEMENT

The implementation of the proposed framework, including the LLM-guided phase preference module, uncertainty-guided action grounding, and MARL-based traffic signal control, is provided as anonymous supplementary material. Details of the model architecture, training procedure, reward design, and experimental settings are provided in Sections 3 and 4, with additional implementation details and experimental configurations included in the Appendix and supplementary material.

## REFERENCES

P. Bairi, S. Swain, A. Bandyopadhyay, K. Aurangzeb, M. Alhussein, and S. Mallik. Intelligent VANET-based traffic signal control system for emergency vehicle prioritization and improved traffic management. *Egyptian Informatics Journal*, 30:100700, 2025. doi: 10.1016/j.eij.2025.100700.

K. Benhamza, H. Seridi, M. Agguini, and A. Bentagine. A multi-agent reinforcement learning based approach for intelligent traffic signal control. *Evolving Systems*, 15(6):2383–2397, 2024. doi: 10.1007/s12530-024-09622-4.

R. Bokade and X. Jin. PyTSC: A unified platform for multi-agent reinforcement learning in traffic signal control. *Sensors*, 25(5):1302, 2025. doi: 10.3390/s25051302.

S. Cai, J. Fang, and M. Xu. XLight: An interpretable multi-agent reinforcement learning approach for traffic signal control. *Expert Systems with Applications*, 273:126938, 2025. doi: 10.1016/j.eswa.2025.126938.

Z. Chen, H. Deng, Z. Li, H. Wen, G. Jin, R. Yu, and B. Leng. HCRMP: An LLM-hinted contextual reinforcement learning framework for autonomous driving. In *Proceedings of the 39th Conference on Neural Information Processing Systems*, 2025. doi: 10.52202/085713-3537.

Sujeong Choi and Yujin Lim. Optimizing traffic signal control using LLM-driven reward weight adjustment in reinforcement learning. *Journal of Information Processing Systems*, 21(1):43–51, 2025. doi: 10.3745/JIPS.04.0334.

T. Chu, J. Wang, L. Codecà, and Z. Li. Multi-agent deep reinforcement learning for large-scale traffic signal control. *IEEE Transactions on Intelligent Transportation Systems*, 21(3):1086–1095, 2019. doi: 10.1109/TITS.2019.2901791.

L. Da, H. Mei, R. Sharma, and H. Wei. Uncertainty-aware grounded action transformation towards sim-to-real transfer for traffic signal control. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pp. 1124–1129, 2023. doi: 10.1109/CDC49753.2023.10383645.

L. Da, M. Gao, H. Mei, and H. Wei. Prompt to transfer: Sim-to-real transfer for traffic signal control with prompt learning. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, pp. 82–90, 2024. doi: 10.1609/aaai.v38i1.27758.

Z. Gao, H. Hao, F. Gao, and R. Zhao. Constrained reinforcement-learning-enabled policies with augmented lagrangian for cooperative intersection management. *IEEE Internet of Things Journal*, 12(5):5396–5411, 2025. doi: 10.1109/JIOT.2024.3487854.

S. Gu, L. Yang, Y. Du, G. Chen, F. Walter, J. Wang, and A. Knoll. A review of safe reinforcement learning: Methods, theories, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):11216–11235, 2024. doi: 10.1109/TPAMI.2024.3457538.

H. Huang, W. He, Q. Liu, Q. Liu, C. Huang, A. P. P. Abdul Majeed, X. Dai, G. Fang, and X. Xu. Towards generalisable and explainable traffic signal control via deep reinforcement learning and large language models. *CAAI Transactions on Intelligence Technology*, 11(2):483–497, 2026. doi: 10.1049/cit2.70113.

A. Jovanović, A. Uzelac, D. Teodorović, and K. Kukić. Traffic signal control at displaced left-turn intersections using an LLM-based actor-critic algorithm. *Expert Systems with Applications*, 301: 130555, 2026. doi: 10.1016/j.eswa.2025.130555.

K. Katzilieris, E. Kampitakis, and E. I. Vlahogianni. A multi-agent reinforcement learning framework for integrated traffic signal control and dynamic bus lane access management. *Transportation Research Part C: Emerging Technologies*, 182:105391, 2026. doi: 10.1016/j.trc.2025.105391.

S. Lai, Z. Xu, W. Zhang, H. Liu, and H. Xiong. LLMLight: Large language models as traffic signal control agents. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, pp. 2335–2346, 2025. doi: 10.1145/3690624.3709379.

H. Mei, X. Lei, L. Da, B. Shi, and H. Wei. Libsignal: An open library for traffic signal control. *Machine Learning*, 113(8):5235–5271, 2024. doi: 10.1007/s10994-023-06412-y.

M. Movahedi and J. Choi. The crossroads of LLM and traffic control: A study on large language models in adaptive traffic signal control. *IEEE Transactions on Intelligent Transportation Systems*, 26(2):1701–1716, 2025. doi: 10.1109/TITS.2024.3498735.

A. Müller and M. Sabatelli. Safe and psychologically pleasant traffic signal control with reinforcement learning using action masking. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 951–958, 2022. doi: 10.1109/ITSC55140.2022.9922306.

A. Müller and M. Sabatelli. Bridging the reality gap of reinforcement learning based traffic signal control using domain randomization and meta learning. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 5271–5278, 2023. doi: 10.1109/ITSC57777.2023.10421987.

L. Nie, D. Qi, B. Liu, P. Li, H. Bao, and H. He. CMRM: Collaborative multi-agent reinforcement learning for multi-objective traffic signal control. *IEEE Transactions on Consumer Electronics*, 71(2):2793–2805, 2025. doi: 10.1109/TCE.2025.3563723.

M. Pternea, P. Singh, A. Chakraborty, Y. Oruganti, M. Milletari, S. Bapat, and K. Jiang. The RL/LLM taxonomy tree: Reviewing synergies between reinforcement learning and large language models. *Journal of Artificial Intelligence Research*, 80:1525–1573, 2024. doi: 10.1613/jair.1.15960.

Y. Ren, Y. Chang, Z. Cui, X. Chang, H. Yu, X. Li, and Y. Wang. Is cooperative always better? multi-agent reinforcement learning with explicit neighborhood backtracking for network-wide traffic signal control. *Transportation Research Part C: Emerging Technologies*, 179:105265, 2025. doi: 10.1016/j.trc.2025.105265.

H. Su, Y. D. Zhong, B. Dey, and A. Chakraborty. EMVLight: A decentralized reinforcement learning framework for efficient passage of emergency vehicles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 4593–4601, 2022. doi: 10.1609/aaai.v36i4.20383.

H. Su, Y. D. Zhong, J. Y. J. Chow, B. Dey, and L. Jin. EMVLight: A multi-agent reinforcement learning framework for an emergency vehicle decentralized routing and traffic signal control system. *Transportation Research Part C: Emerging Technologies*, 146:103955, 2023. doi: 10.1016/j.trc.2022.103955.

R. Tiwari, S. Khapre, and A. Singh. Reinforcement learning in robotic systems: A review on sim-to-real transfer. *Robotics and Autonomous Systems*, 198:105327, 2026. doi: 10.1016/j.robot.2025.105327.

Z. Wang, J. Li, Y. Liu, X. Li, C. Yan, and Y. Zhang. RL-Fusion: The large language model fusion method based on reinforcement learning for task enhancing. *Applied Sciences*, 15(4):2186, 2025. doi: 10.3390/app15042186.

K. Yang, Z. Wang, X. Meng, L. Li, Y. Shi, Y. Yu, H. Wang, and Z. Yao. Store-and-forward with graph attention: Enhanced multi-agent reinforcement learning for emergency-responsive traffic signal control. *Engineering Applications of Artificial Intelligence*, 159:111602, 2025. doi: 10.1016/j.engappai.2025.111602.

J. Yao, J. Li, X. Xu, C. Tan, K. H. Yap, and R. Su. Incorporating vision-based artificial intelligence and large language model for smart traffic light control. *Applied Soft Computing*, 179:113333, 2025. doi: 10.1016/j.asoc.2025.113333.

Z. Zhai, R. Hao, B. Cui, and S. Wang. HGAT and multi-agent RL-based method for multi-intersection traffic signal control. *IEEE Transactions on Intelligent Transportation Systems*, 26 (5):6848–6864, 2025. doi: 10.1109/TITS.2025.3526223.

R. Zhou, T. Nousch, L. Wei, and M. Wang. Constrained traffic signal control under competing public transport priority requests via safe reinforcement learning. *Expert Systems with Applications*, 284: 127676, 2025. doi: 10.1016/j.eswa.2025.127676.

## A Networked Dec-POMDP Formulation

At decision step $t$, the network occupies global state $\mathbf{s}_t \in \mathcal{S}$, while agent $i$ receives only its partial observation $\mathbf{o}_{i,t} = \Omega_i(\mathbf{s}_t, \mathcal{N}(i)) \in \mathcal{O}_i$ from local sensors and neighboring intersections. It selects an action $a_{i,t} \in \mathcal{A}_i$, and the joint action $\mathbf{a}_t = (a_{1,t}, \ldots, a_{N,t})$ determines the network transition $Z(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)$ and resulting reward. Under decentralized execution, each agent must make real-time decisions without access to the full global state while accounting for traffic dependencies through neighboring intersections.

## B Probabilistic EMV Arrival Perception

In Figure 5, EMV movement is continuously affected by queue blocking and discrete signal phase transitions, which introduce non-negligible interaction delays. As a result, distance–speed based formulations tend to underestimate the actual arrival time, particularly when upstream congestion persists.

Figure 6 presents the architecture of the probabilistic ETA prediction module, which consists of three stages: feature construction, temporal encoding, and heteroscedastic decoding. At each time step $t$, the input feature vector $\mathbf{x}_t \in \mathbb{R}^F$ contains three categories of information: (1) ego-dynamics, including the EMV speed $\zeta_e$ and remaining distance $\lceil_{\text{rem}}$ to the stop line; (2) interaction features, including the headway distance $\lceil_{\text{front}}$ to the leading vehicle and the leader speed $\zeta_{\text{front}}$; and (3) intersection context, including lane queue length and the current signal phase. A sliding temporal window $\mathbf{X}_{\text{seq}} = \{\mathbf{x}_{t-T+1}, \ldots, \mathbf{x}_t\}$ of length $T$ is used to represent the recent evolution of these variables. The resulting sequence is encoded by LSTM layers to obtain a hidden representation $\mathbf{h}_t$.

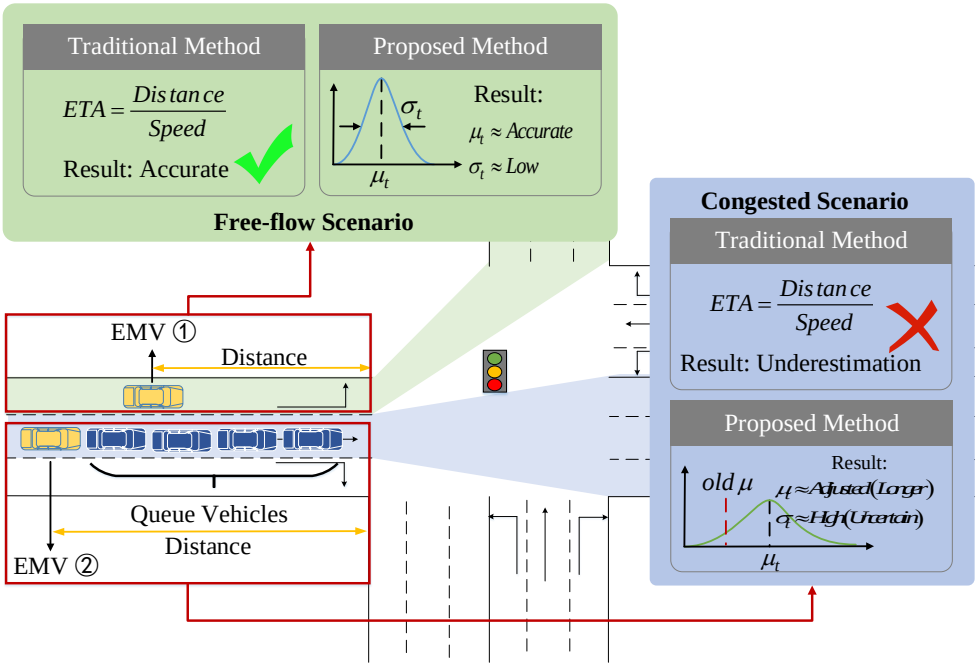



Figure 5: Comparison between heuristic ETA and probabilistic ETA modeling.

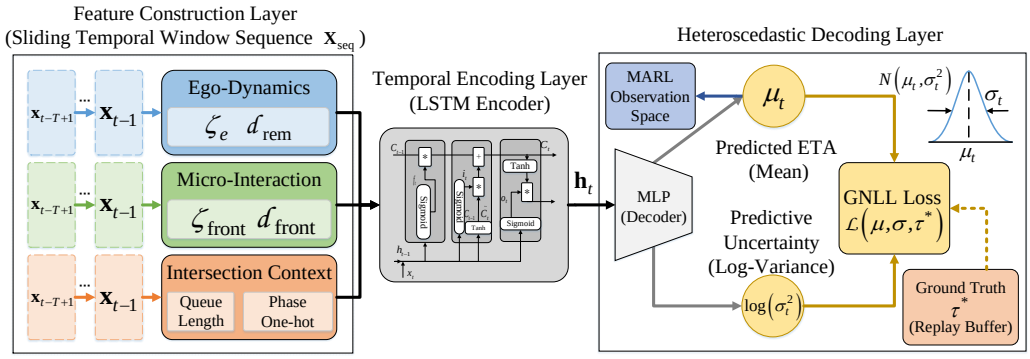



Figure 6: Architecture of the LSTM-based probabilistic ETA prediction module.

# C MADQN-BASED TRAFFIC SIGNAL CONTROL

## C.1 ATTENTION-BASED NEIGHBOR COMMUNICATION

Each agent observes only its local state, neighboring information is incorporated through an attention-based communication encoder. For agent $i$, the raw observations of adjacent intersections $j \in \mathcal{N}(i)$ are assembled into a neighbor feature matrix $\mathbf{X}_{i,t}^{\mathcal{N}}$ and transformed into an initial embedding $\mathbf{H}_{i,t}^{0}$:

$$\mathbf{H}_{i,t}^{0} = \mathrm{ReLU}\left(\mathbf{X}_{i,t}^{\mathcal{N}}\mathbf{W}_e + \mathbf{b}_e\right) \tag{13}$$

where the linear transformation parameters are learnable. A multi-head self-attention mechanism then captures the interactions among neighboring intersections:

$$\mathbf{head}_m = \mathrm{Softmax}\left(\frac{\mathbf{H}_{i,t}^{0}\mathbf{W}_m^{Q}\left(\mathbf{H}_{i,t}^{0}\mathbf{W}_m^{\mathcal{K}}\right)^{\top}}{\sqrt{\xi_h}}\right)\mathbf{H}_{i,t}^{0}\mathbf{W}_m^{V} \tag{14}$$

where the attention computation is performed independently for each head $m \in \{1, 2, \ldots, M\}$ with the corresponding learnable projections $\mathbf{W}_m^Q, \mathbf{W}_m^K, \mathbf{W}_m^V$ and scaling factor $\xi_h$. The resulting head outputs are concatenated and linearly projected, followed by a residual connection and layer normalization:

$$\mathbf{H}_{i,t}^{\text{att}} = \text{LayerNorm}\big(\mathbf{H}_{i,t}^0 + \text{Concat}\big(\mathbf{head}_1, \ldots, \mathbf{head}_M\big)\mathbf{W}_{\mathcal{Z}}\big) \tag{15}$$

The refined neighbor features are further projected to the final representation $\mathbf{M}_{i,t}^{\mathcal{N}}$, and mean pooling is applied before concatenation with the local observation:

$$\mathbf{M}_{i,t}^{\mathcal{N}} = \mathbf{H}_{i,t}^{\text{att}}\mathbf{W}_\phi + \mathbf{b}_\phi, \mathbf{o}_{i,t} \quad = \Big[\mathbf{o}_{i,t}^{\text{local}} \,\|\, \text{Pool}\big(\mathbf{M}_{i,t}^{\mathcal{N}}\big)\Big] \tag{16}$$

### C.2 Action space

Each agent selects one feasible and conflict-free signal phase from its predefined local action space:

$$\mathcal{A}_i = \big\{a^{(1)}, a^{(2)}, \ldots, a^{(P)}\big\}, \tag{17}$$

where each action represents a legal combination of compatible traffic movements. The selected phase is executed at the corresponding intersection at decision step $t$.

### C.3 Traffic Pressure and Lane Occupancy

Two local traffic metrics are first calculated. Traffic pressure $\mathcal{P}_{i,t}$ aggregates vehicle count $v_{i,t}^{(k)}$, queue length $q_{i,t}^{(k)}$, and cumulative delay $\varpi_{i,t}^{(k)}$ across incoming lanes $\mathcal{L}_{\text{in}}$ and is normalized as

$$\mathcal{P}_{i,t} = \text{Norm}\left(\sum_{k=1}^{|\mathcal{L}_{\text{in}}|}\big(w_1 \cdot v_{i,t}^{(k)} + w_2 \cdot q_{i,t}^{(k)} + w_3 \cdot \varpi_{i,t}^{(k)}\big)\right) \in [0,1] \tag{18}$$

where $w_1$, $w_2$, and $w_3$ are the predefined weighting coefficients for the respective traffic metrics. Lane occupancy $U_{i,t}$ is additionally defined as the ratio between the total vehicle length and the lane length:

$$U_{i,t} = \frac{1}{|\mathcal{L}_{\text{in}}|}\sum_{k' \in \mathcal{L}_{\text{in}}} \frac{\sum_{v \in \psi_{k',t}} l_v}{l_{k'}} \in [0,1] \tag{19}$$

where $l_{k'}$, $l_v$ denote the physical length of lane $k'$ and vehicle $v$, $\psi_{k',t}$ defines the set of vehicles present on lane $k'$ at time step $t$.

### C.4 Network optimization and training objective

For each intersection agent $i$, the Q-network parameterized by $\theta_i$ takes the observation $\mathbf{o}_{i,t}$ as input and estimates the action-values $Q_i(\mathbf{o}_{i,t}, \cdot; \theta_i)$ for all feasible signal phases. During training, actions are selected using an $\varepsilon$-greedy strategy.

The primary objective of this main network is to approximate the optimal action-value function $Q_i^*$. To improve training stability, a target network parameterized by $\theta_i^-$ is introduced to construct the temporal difference (TD) target $y_{i,t}$:

$$y_{i,t} = r'_{i,t} + \gamma \max_{a' \in \mathcal{A}_i} Q_i\big(\mathbf{o}_{i,t+1}, a'; \theta_i^-\big). \tag{20}$$

To break the temporal correlations inherent in sequential traffic data, training is performed over a mini-batch of transitions sampled from the experience replay buffer $\mathcal{D}$. The main network parameters $\theta_i$ are then optimized by minimizing the expected mean squared error (MSE) loss over these samples:

$$L(\theta_i) = \mathbb{E}_{(\mathbf{o}_{i,t}, a_{i,t}, r'_{i,t}, \mathbf{o}_{i,t+1}) \sim \mathcal{D}}\Big[\big(y_{i,t} - Q_i(\mathbf{o}_{i,t}, a_{i,t}; \theta_i)\big)^2\Big]. \tag{21}$$

Finally, the network parameters $\theta_i$ are optimized using RMSprop with learning rate $\eta$:

$$\theta_i \leftarrow \theta_i - \eta \nabla_{\theta_i} L(\theta_i). \tag{22}$$

To ensure stable convergence, the target network parameters $\theta_i^-$ are periodically synchronized with the main network via hard updates ($\theta_i^- \leftarrow \theta_i$).

## D TRAFFIC-TO-TEXT ALIGNMENT FOR LLM PROMPTING

Numerical traffic observations are converted into structured descriptions through macroscopic congestion and microscopic EMV-priority information. For intersection $i$ and phase $p$, lane-level queue measurements are aggregated into a phase-level congestion indicator:

$$c_{i,p,t}^{\text{macro}} = \sum_{k \in \mathcal{L}'_{\text{in}}(p)} q_{i,t}^{(k)} \tag{23}$$

where $\mathcal{L}'_{\text{in}}(p)$ denotes the set of lanes associated with phase $p$. The resulting indicators describe the spatial distribution of congestion across competing phases.

For each active EMV $e$, a state tuple is constructed as:

$$\mathbf{o}_e = \left\langle \zeta_e,\ \lceil_{\text{rem}},\ \zeta_{\text{front}},\ \lceil_{\text{front}},\ I_{\text{curr}},\ \ell_{\text{fut}},\ \Phi_{\text{valid}},\ \Phi_{\text{valid}}^{\text{fut}} \right\rangle \tag{24}$$

where vehicle speed and distance variables characterize EMV urgency, while the current intersection $I_{\text{curr}}$ and planned route $\ell_{\text{fut}}$ describe its spatial context. The feasible phase sets at the current $\Phi_{\text{valid}}$ and downstream intersections $\Phi_{\text{valid}}^{\text{fut}}$ are also included to identify EMV-relevant movements. These numerical features are converted into structured prompts using conditional templates, with downstream routing information included only when applicable.

## E UNCERTAINTY-GUIDED ACTION GROUNDING

During retraining in the source environment, UGAT is embedded into the interaction loop. At each step, the policy first outputs an intended action $a_{i,t}$. The forward model predicts the corresponding next observation under target dynamics, and the inverse model maps this predicted transition back to a grounded action $\hat{a}_{i,t}^{g}$ with uncertainty $u_{i,t}$. An uncertainty-aware selection rule determines the executed action $\tilde{a}_{i,t}$. If $u_{i,t}$ falls below a preset threshold $\varepsilon$, the grounded action $\tilde{a}_{i,t} = \hat{a}_{i,t}^{g}$ is adopted. Otherwise, the system retains the original policy action $\tilde{a}_{i,t} = a_{i,t}$. Through this selective grounding strategy, action correction is performed only under reliable inverse mappings, allowing the retraining process to better approximate target-domain transition patterns and thereby improving transferability under unseen traffic dynamics.

## F PROGRESSIVE TRAINING AND DEPLOYMENT STRATEGY

Since the proposed framework contains multiple tightly coupled modules with different optimization objectives and update frequencies, directly training all components jointly may introduce unstable interactions during policy learning. To improve training stability and ensure reliable feature propagation across modules, a progressive training and transfer strategy is adopted, as illustrated in Figure 7.

Stage 1 begins with pre-training the MADQN in the CityFlow environment to establish a baseline policy under EMV scenarios. Initially, the ETA for EMVs is estimated using conventional distance-speed heuristics. After policy converges, the learned policy is frozen to stabilize the traffic interaction patterns, and the EMV spatiotemporal trajectories are collected to construct the datasets for subsequent temporal modeling.

Building upon the trajectory datasets collected, the LSTM-based heteroscedastic regression module is trained in Stage 2. After convergence, the LSTM parameters are frozen, and the probabilistic outputs are incorporated into the MADQN observation space. Since the augmented observation space changes the input distribution of the policy network, the MADQN is subsequently fine-tuned to adapt its decision-making behavior to the newly introduced temporal information.

After the temporal perception module becomes stable, the LLM-generated phase preference distribution is incorporated into the MADQN in Stage 3 through the residual fusion mechanism. The policy network is then further optimized to align its action selection behavior with the newly introduced semantic preferences.

In Stage 4, interaction datasets are collected in both CityFlow and simulation of urban mobility (SUMO) using the same baseline control policy. Based on these datasets, the forward and inverse

models of UGAT are trained. The UGAT module is then incorporated into the retraining loop in CityFlow. Finally, the learned policy is directly deployed in SUMO to evaluate transferability under different traffic dynamics.

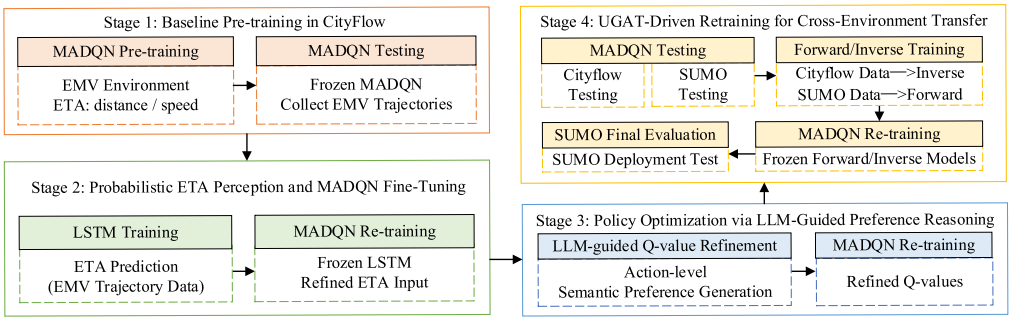



Figure 7: The proposed multi-stage training and transfer pipeline.

# G Experimental Details and Additional Results

## G.1 Traffic datasets

The Jinan network is further configured with three traffic demand levels ranging from 4365 veh/h to 6295 veh/h. Together with the Hangzhou and New York networks, these configurations form five evaluation datasets covering a wide spectrum of cumulative traffic volumes (Figure 8b).

## G.2 EMV route configurations

Within these traffic scenarios, EMV routes are generated to cover different intersections and turning movements across the road networks, as illustrated in Figure 9a . The route lengths are distributed within the normal traffic travel range to avoid extreme trajectory settings. In addition, EMVs are dispatched throughout the evaluation period under different traffic conditions, including free-flow and congested states. The resulting configurations provide diverse spatiotemporal emergency traffic scenarios for performance evaluation.

## G.3 Baselines and Ablation Settings

To comprehensively evaluate the performance of the proposed method and analyze the contribution of its core components, three representative external baselines and three internal ablation variants are selected for comparative analysis.

The external baselines cover a range of traffic signal control approaches, from traditional rule-based methods to advanced data-driven approaches.

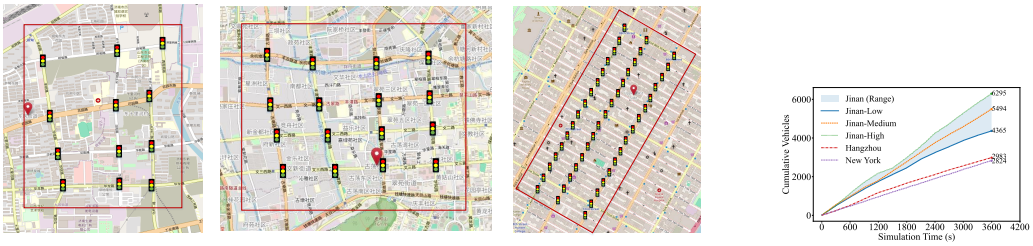
·Fixed-time: A traditional non-adaptive control method employing pre-configured cycle and phase schemes, serving as the baseline reference for deterministic control.

·MA2C: A decentralized advantage actor-critic algorithm that introduces neighbor fingerprints and spatial discount factors, representing the performance benchmark of mainstream MARL approaches.

·HGAT: A recent graph-based traffic signal control method integrating graph attention networks (GAT) to explicitly model road network topologies and capture spatiotemporal dependencies.

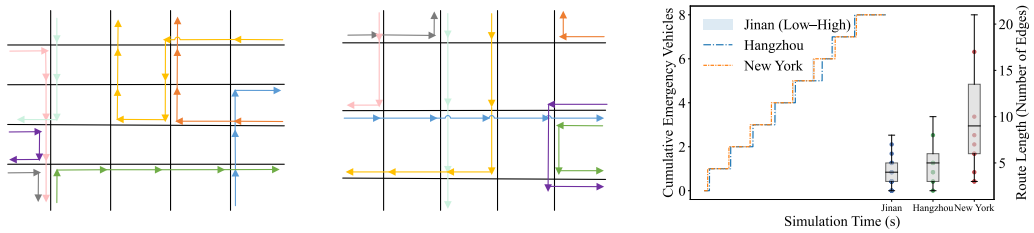
The internal ablation variants are designed to verify the effectiveness of key modules in the proposed method (denoted as Ours):

·Without ETA refinement (w/o ETA Refinement): Removes the LSTM-based prediction module, reverting to conventional distance-speed heuristics for EMV arrival estimation. This variant is used to evaluate the contribution of the probabilistic ETA prediction module.




(a) (b)

Figure 8: Road network structures and traffic volume distributions. (a) Dongfeng Sub-district, Jinan, China. Gudang Sub-district, Hangzhou, China. Upper East Side, Manhattan, New York, USA. (b) Cumulative number of inserted vehicles over the evaluation period for each dataset.




(a) Visualization of EMV trajectories in the Hangzhou and Jinan networks.

(b) EMV departure/arrival times and route lengths for each dataset.

Figure 9: Spatiotemporal profiles and trajectory distributions of EMV operations.

·LLM as feature encoder (Repl. Preference w/ State): Substitutes the active phase preference distribution with a semantic embedding pipeline. In this configuration, the LLM functions solely as a feature encoder: Raw traffic observations are converted into natural language descriptions, encoded by an embedding model, and integrated via an adapter-fusion network. This setup eliminates directional guidance on decision-making, aiming to contrast the contribution of semantic feature encoding and preference-guided decision adjustment.

·Without UGAT adaptation (w/o UGAT Adaptation): Removes the forward and inverse dynamics models of the UGAT mechanism, directly deploying the CityFlow-trained policy to SUMO. This serves to evaluate the impact of cross-environment traffic dynamics differences and validates the necessity of UGAT.

### G.4 Evaluation Metrics

To systematically quantify algorithmic performance across the comparative models, a comprehensive suite of evaluation metrics is employed. As detailed in Table 2, the evaluation metrics are grouped into three categories: traffic efficiency, intersection stability, and environmental sustainability.

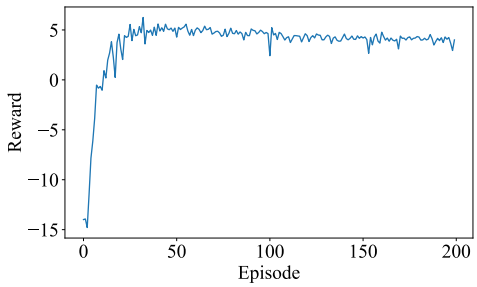
### G.5 Training convergence analysis

Figure 10 illustrates the learning dynamics and the evolution of key traffic performance metrics during the training phase. Specifically, Figure 10a depicts the convergence trajectory of the global cumulative reward. During the early training stage, the cumulative reward exhibits noticeable fluctuations due to policy exploration. As training progresses and the policy transitions from exploration to exploitation, the reward steadily increases, eventually reaching a stable plateau around episode 100, indicating stable policy convergence.

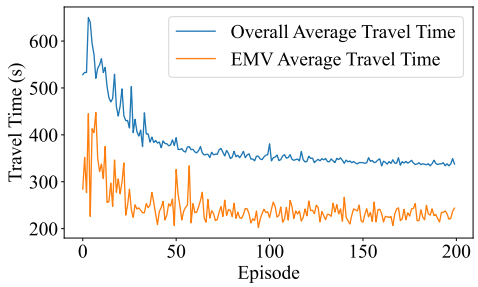
Along with the convergence of the reward, several traffic performance metrics exhibit consistent improvements. As shown in Figure 10b, the average travel times for both regular vehicles and

Table 2: Comprehensive evaluation metrics and corresponding operational dimensions.

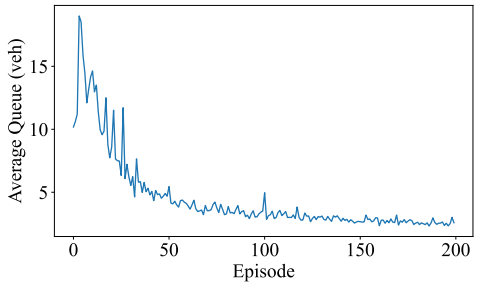
| Operational Dimension | Evaluation Metric | Abbr. | Unit |
|---|---|---|---|
| Travel Efficiency | Average Travel Time (Overall / EMV) | ATT | s |
| | Network Throughput | TP | veh/h |
| | Average Speed | AS | m/s |
| | Time Loss | TL | s |
| Intersection Stability | Average Queue Length | AQL | m |
| | Average Waiting Time | AWT | s |
| | Average Waiting Count | AWC | times |
| Environmental Sustainability | Fuel Consumption | Fuel | g |
| | Carbon Emissions | $CO_2$, CO | g |
| | Trace Pollutants | HC, $NO_x$, $PM_x$ | mg |
| | Noise Levels | Noise | dB |



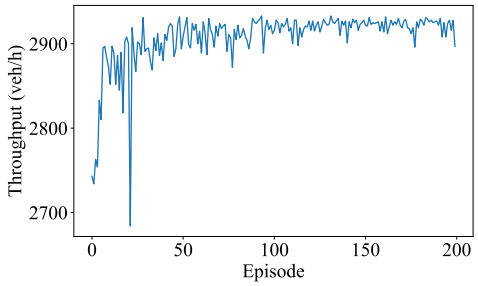

(a) Convergence profile of the global cumulative reward.




(b) Average travel time of regular vehicles and EMVs.




(c) Evolution of average queue vehicles across the network.




(d) Evolution of network throughput during training.

Figure 10: Training convergence and evolution of key traffic performance metrics.

EMVs display a consistent downward trend. The reduction suggests that the learned policy can improve emergency response efficiency without significantly degrading overall traffic operations.

To further explain the mechanism behind this macroscopic efficiency gain, Figure 10c and Figure 10d map the evolution of network congestion and capacity, respectively. As the average queue length progressively decreases, the network throughput shows a continuous upward trend. This inverse relationship suggests that the learned policy effectively improves traffic flow utilization at intersections.

### G.6 Overall ATT Distribution Analysis

Figure 11 further examines the distribution of Overall ATT across the five scenarios. The proposed method maintains a low median and relatively compact interquartile range across different demand levels and network sizes, indicating that the observed gains are accompanied by stable performance rather than isolated improvements.

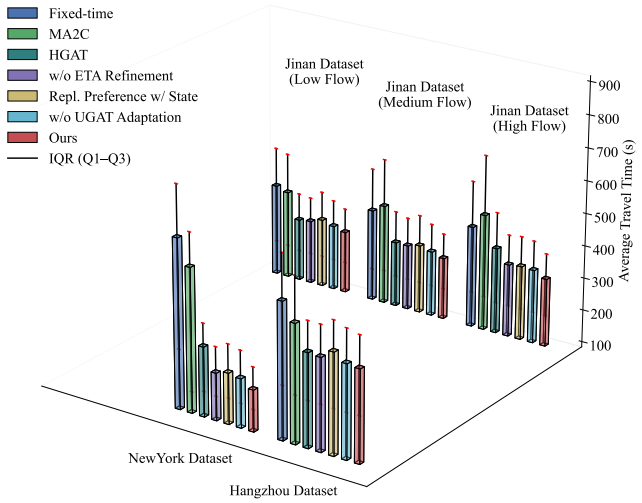



Figure 11: Overall ATT distributions across five traffic scenarios.

### G.7 Intersection-level Waiting and Environmental Analysis

To further evaluate the spatial distribution of traffic pressure, Figure 12 compares the AWT distribution across intersections in the Hangzhou dataset. The proposed method achieves the lowest median AWT together with a relatively compact distribution, indicating more balanced traffic conditions across intersections. In contrast, Fixed-time control exhibits a higher median and a wider distribution range, suggesting the existence of localized congestion hotspots. Although HGAT and the ablation variants reduce the overall waiting level to some extent, they still exhibit wider whisker ranges than the proposed method, indicating larger performance variation across intersections.

Figure 13 illustrates the correspondence between EMV arrival time and departure time across different traffic signal control methods in the Hangzhou dataset. In this visualization, the vertical deviation from the ideal reference diagonal (y=x) corresponds to the travel times of the EMVs. The proposed method remains closely distributed around the reference line, indicating relatively stable EMV travel times throughout the evaluation period. In contrast, the baseline methods exhibit progressively larger deviations as the simulation proceeds, suggesting that they experience increasing difficulty in dissipating accumulated congestion over time, leading to longer delays for subsequently dispatched EMVs.

Shifting the focus from emergency response to environmental impact, Figure 14 compares the energy consumption, pollutant emissions, and noise levels of different control methods, where all results are normalized against the Fixed-time baseline (100%). The results show that all adaptive control methods achieve lower environmental costs than Fixed-time control. Among them, the proposed method consistently maintains the lowest normalized values across all evaluated categories. In particular, CO and HC emissions are reduced to 31.3% and 36.9% of the Fixed-time baseline, respectively. Combined with the mobility improvements reported in previous sections, these results

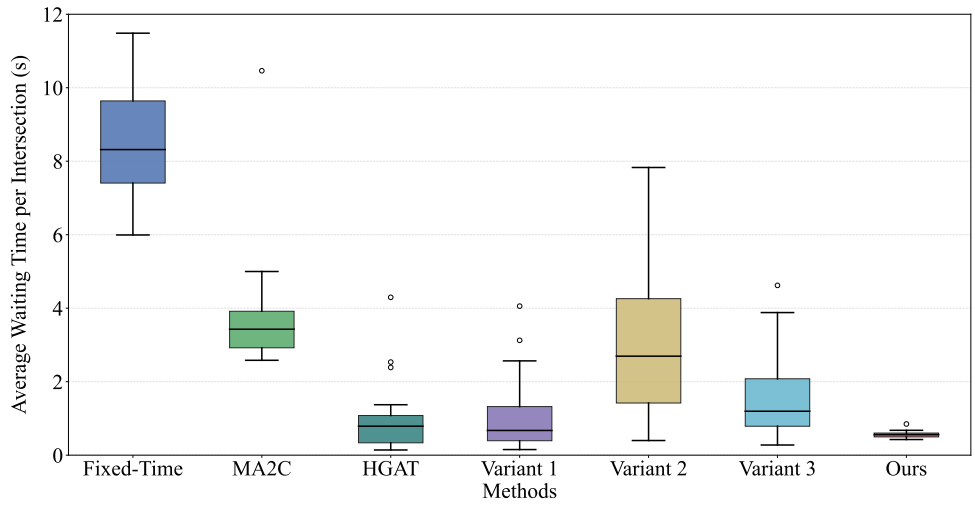



Figure 12: Distribution of intersection-level AWT under different control methods in the Hangzhou dataset.

indicate that the proposed method improves EMV traffic efficiency while simultaneously reducing environmental impact.

### G.8 Spatial Delay Heatmaps under Varying Demand

Figure 15 presents lane-level TL heatmaps across different traffic networks to evaluate the spatial scalability of the proposed method. Taking the large-scale New York dataset as an example, the Fixed-time baseline exhibits large high-delay regions along arterial corridors, indicating accumulated delays across consecutive intersections. In contrast, the proposed method alleviates these high-delay regions, with a larger proportion of the network shifting toward lower-delay states. Compared with Fixed-time control, the proposed framework exhibits a more spatially balanced distribution of lane-level delay across the network. Similar trends can also be observed in the Jinan and Hangzhou datasets. Across different network topologies, the proposed method consistently reduces the intensity and spatial extent of localized congestion regions, indicating stable traffic performance.

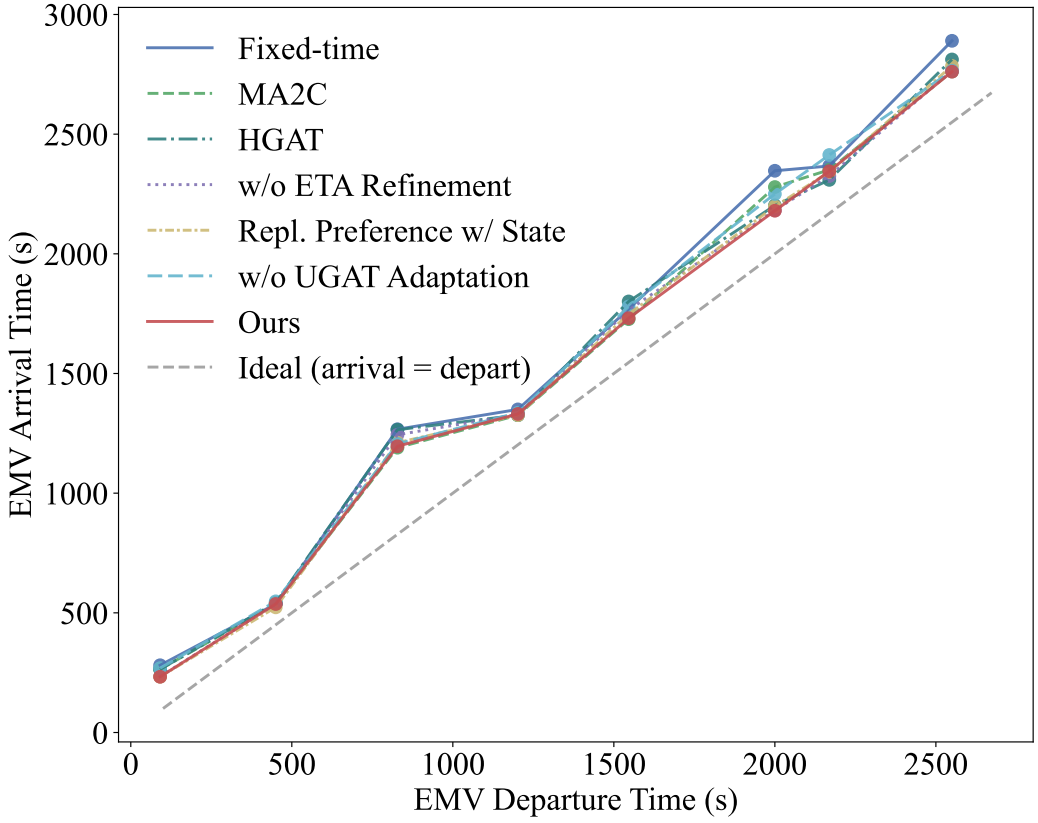



Figure 13: Relationship between EMV departure and arrival times under different control methods in the Hangzhou dataset.

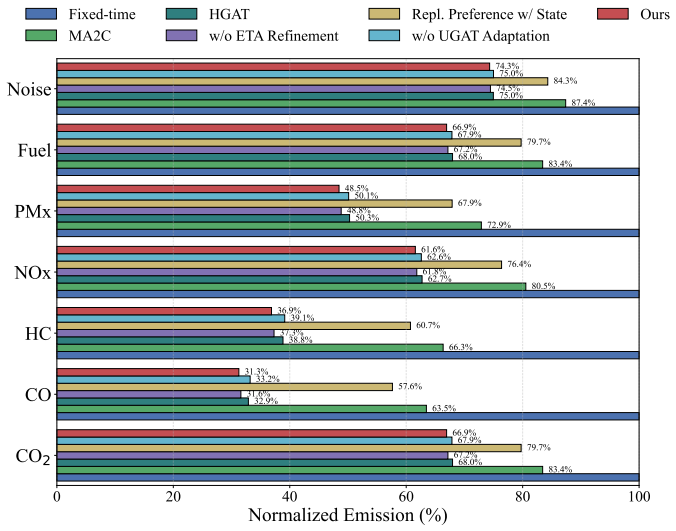



Figure 14: Normalized comparison of environmental metrics under different control methods in the Hangzhou dataset.

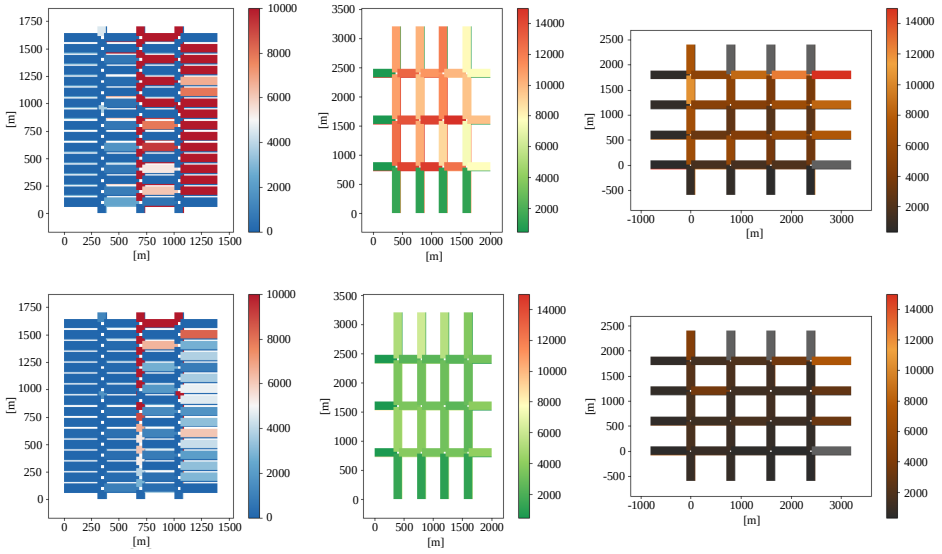



Figure 15: Lane-level TL heatmaps under Fixed-time control and the proposed method across different traffic networks. (a) New York: Fixed-time. (b) Jinan-Low: Fixed-time. (c) Hangzhou: Fixed-time. (d) New York: Ours. (e) Jinan-Low: Ours. (f) Hangzhou: Ours.