CuraLight: Debate-Guided Data Curation for LLM-Centered Traffic Signal Control

Qing Guo*, Xinhang Li*, Junyu Chen*, Zheng Guo*,
Shengzhe Xu*, Lin Zhang*,†, Lei Li*
* School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China
† Beijing Big Data Center, Beijing, China

***Abstract*—As a core component of intelligent transportation systems (ITS), traffic signal control (TSC) mitigates congestion, reduces emissions, and improves travel-time reliability. Recent TSC methods based on rules, reinforcement learning (RL), and large language models (LLMs) have improved adaptivity across urban networks. Yet key limitations persist: rule-based methods and RL methods provide limited interpretability of timing actions; LLM-based methods face challenges in selecting high-dimensional timings at heterogeneous intersections; and because LLMs are typically trained with far less task-aligned interactive data than RL, intersection-specific operational knowledge remains limited. This paper presents CuraLight, an LLM-centered framework in which an RL agent assists the fine-tuning of an LLM agent. RL exploration gathers traffic states and actions, enhanced prompting constructs imitation pairs for fine-tuning the base model, timing decisions are structured to elicit interpretable preferences, and a multi-LLM ensemble deliberation system conducts adversarial debates that defend each RL-filtered phase action and consolidates a consensus outcome for RL-assisted fine-tuning. Evaluation in SUMO across heterogeneous networks from Jinan (17 intersections), Hangzhou (19), and Yizhuang (177) shows improvements over recent state of the art: ATT by about 5.34%, AQL by about 5.14%, and AWT by about 7.02%. Code and configurations are available at https://anonymous.4open.science/r/CuralightCode-6437/.**

***Index Terms*—traffic signal control, imitation fine-tuning, multi-agent debate, heterogeneous intersections**

I. Introduction

Recent advances in AI and networked sensing are accelerating Intelligent Transportation Systems, expanding the capacity to observe traffic states and coordinate operations at scale. Traffic signal control is central to allocating intersection capacity and improving travel-time reliability. However, many methods still focus on standard intersections or fixed-time settings, and dynamic timing is often studied at limited scales. Compared with reinforcement learning, LLM-based TSC methods typically rely on far fewer environment interactions, leading to insufficient scene-specific learning and brittle decisions under demand shifts and mild heterogeneity. Addressing these issues is crucial for scalable, adaptive, and interpretable urban TSC.

Reinforcement learning (RL) has been widely applied to traffic signal control (TSC). Jiang et al. proposed UniComm for multi-intersection coordination via universal communication [1]. Gu et al. introduced π-Light, representing policies as programmatic rules for interpretable control and improved transfer [2]. Yu et al. developed a decentralized RL framework that couples signal control with bus priority and supports model reuse across networks [3]. Despite progress, action-level interpretability and cross-scenario generalization remain fragile.

LLM-based methods leverage language-model reasoning for transparency and transfer. Da et al. proposed PromptGAT to inject LLM-derived dynamics knowledge for grounded action transformation [4]. Lai et al. introduced LLMLight, using a specialized backbone to perceive textualized states and decide signal phases via prompted chain-of-thought [5]. Yuan et al. presented CoLLMLight for cooperative network-wide control with spatiotemporal graph reasoning [6]. However, evaluations often assume non-heterogeneous intersections and predefined phase sets, leaving green durations fixed and heterogeneous phasing requirements insufficiently addressed.

Multi-agent deliberation and judge-ensemble methods improve reliability through cross-checking. Liang et al. proposed Multi-Agent Debate (MAD) [7]; Estornell and Liu et al. studied theory-grounded debate with stabilizing interventions [8]; Jung et al. introduced Trust-or-Escalate with cascaded judges [9]. Existing studies mainly target single-turn tasks, whereas traffic signal control requires per-timestep selection among many phase/timing candidates under real-time constraints, motivating deliberation mechanisms tailored to sequential control.

This paper presents CuraLight, a traffic signal control agent based on a large language model trained with a diffusion-based RL assistant and a multi-LLM ensemble deliberation system (Fig. 1). The RL assistant explores heterogeneous networks and provides reference actions and preference signals, while the deliberation system evaluates candidate timing actions and yields priority-aware supervision for adapter fine-tuning. Our contributions are:

- An RL-assisted data generation pipeline is developed, where a diffusion-based RL agent explores the phase and timing space and collaborates with GPT to generate interaction trajectories, which are converted into prompt and answer pairs for LoRA-based imitation fine-tuning.
- A multi-LLM ensemble deliberation system is introduced to evaluate and decompose candidate signal timing actions, enabling priority-aware collection, ranking, filtering, and adapter training on scenario-specific datasets.

Corresponding author: Lei Li (leili@bupt.edu.cn).

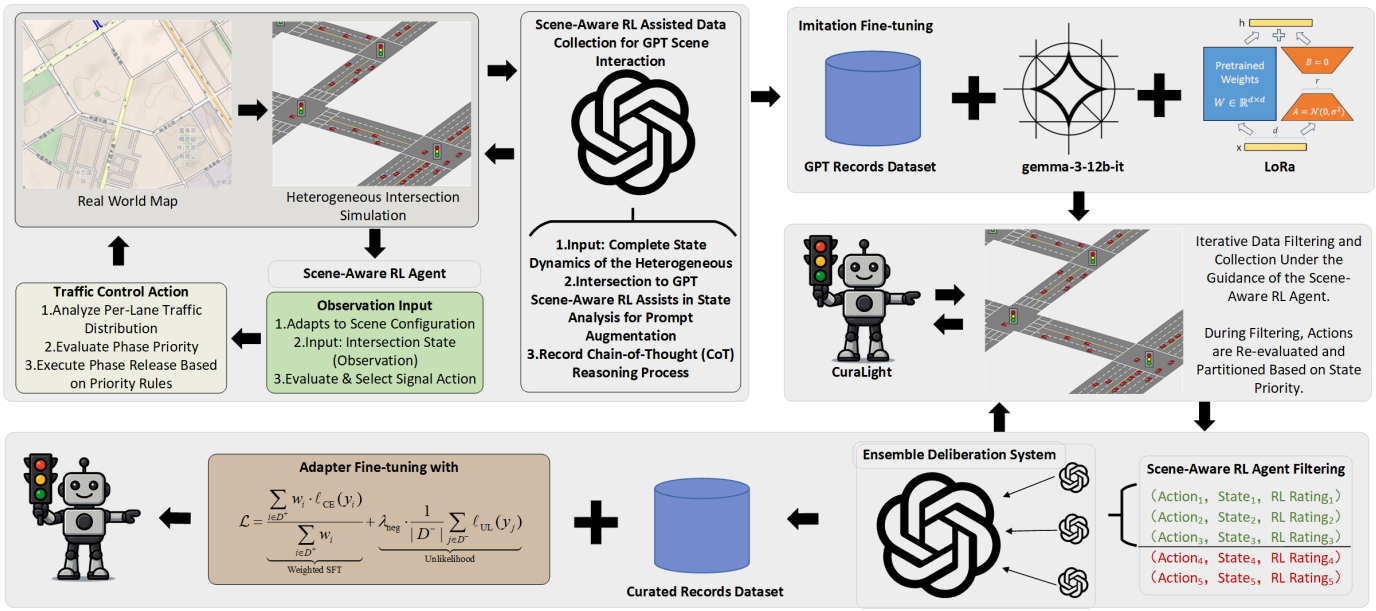


Fig. 1. Overview of CuraLight. RL agent assists GPT in exploring urban networks and collecting interaction trajectories for LoRA-based imitation fine-tuning of the CuraLight LLM agent. An RL-based selector and a multi-LLM ensemble deliberation system then evaluate and rank candidate signal-timing actions, providing prioritized supervision for scenario interaction data collection, filtering, and adapter training.

- Experiments in SUMO on heterogeneous real-world networks from Jinan (17 intersections), Hangzhou (19), and Yizhuang (177) demonstrate that CuraLight outperforms recent methods, improving ATT by about 5.34%, AQL by about 5.14%, and AWT by about 7.02%, while providing interpretable LLM-driven timing decisions.

The rest of this paper is organized as follows: Section II presents heterogeneous intersection modeling and the RL-assisted training pipeline with multi-LLM ensemble deliberation for priority-aware fine-tuning. Section III reports experimental results. Section IV concludes the paper.

II. Heterogeneous Intersection Modeling

As shown in Fig. 2, real-world urban road networks contain intersections with diverse geometries and lane configurations, leading to heterogeneous lane-level control requirements. We adopt a unified lane-based model for subsequent state and action design.

Definition 1 (Road network). The urban road network is modeled as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, 2, \dots, I\}$ is the set of intersections and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of directed roads. Each road $(i, j) \in \mathcal{E}$ consists of a lane set $\mathcal{L}_{ij} = \{\ell_{ij}^1, \ell_{ij}^2, \dots\}$ with a fixed travel direction from i to j.

Definition 2 (Traffic movements). At intersection $i \in \mathcal{V}$, incoming and outgoing lanes are denoted by $\mathcal{U}_i = \{u_i^1, \dots, u_i^{n_i^{\text{in}}}\}$ and $\mathcal{O}_i = \{o_i^1, \dots, o_i^{n_i^{\text{out}}}\}$, where $n_i^{\text{in}} = |\mathcal{U}_i|$ and $n_i^{\text{out}} = |\mathcal{O}_i|$ vary across intersections. A movement is a feasible connection $m = (u_i^j, o_i^k)$ with $u_i^j \in \mathcal{U}_i$ and $o_i^k \in \mathcal{O}_i$. The movement set is $\mathcal{M}_i \subseteq \mathcal{U}_i \times \mathcal{O}_i$, and $M_i = |\mathcal{M}_i|$.

Definition 3 (Signal phases and durations). A phase at intersection i is a subset $P_i^q \subseteq \mathcal{M}_i$ of mutually compatible movements. The phase set is $\mathcal{P}_i = \{P_i^1, \dots, P_i^{J_i}\}$, where J_i is determined by movement conflicts. Signal control proceeds in decision steps $t = 0, 1, 2, \dots$. At each step t, the controller selects a phase index $p_i(t) \in \{1, \dots, J_i\}$ and a green duration $D_i(t)$ with bounds $D_i^{\min} \le D_i(t) \le D_i^{\max}$. The pair $(p_i(t), D_i(t))$ defines the signal timing action.

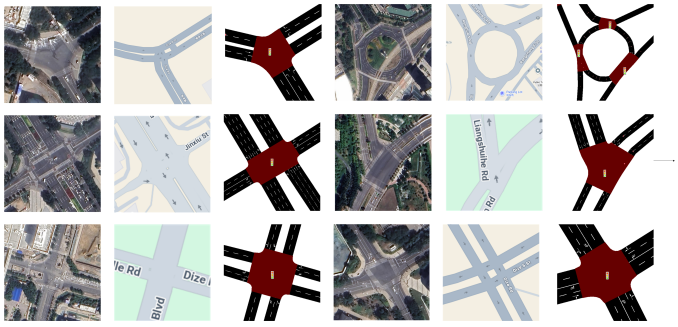


Fig. 2. Real-to-Simulation Modeling of Heterogeneous Intersections

A. RL Agent Assisted GPT Interaction

1) Diffusion based RL Assistant for TSC: Each signalized intersection i is assisted by a diffusion-based RL controller. At step t, a short history of lane-level observations is encoded into a compact state representation $h_i(t)$. The timing action is $a_i(t) = \big(p_i(t), D_i(t)\big)$, where $p_i(t) \in \{1, \dots, J_i\}$ is a phase index selected by a *Pressure-Based Selector* and $D_i(t) \in [D_i^{\min}, D_i^{\max}]$ is the green duration. Conditioned on $(h_i(t), p_i(t))$, a diffusion policy $\pi_\theta\big(D \mid h_i(t), p_i(t)\big)$ generates duration proposals, while a *History Experience Pool* retrieves a small set of previously executed durations as anchors.

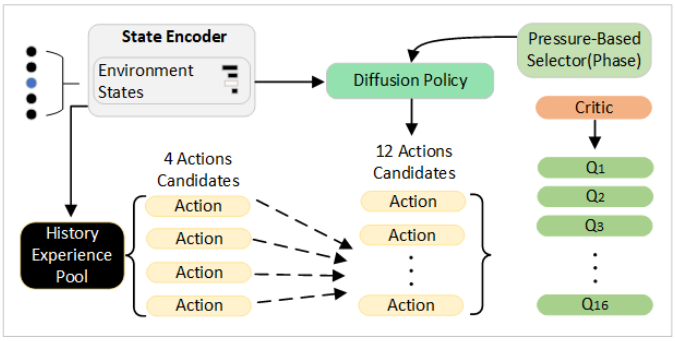


Fig. 3. Overview of the Diffusion-Based RL Assistant with Pressure-Based Phase Selection

During online control (Fig. 3), the selector first determines $p_i(t)$ from $h_i(t)$. A candidate set of durations is then formed as

$$\mathcal{D}_i(t) = \mathcal{D}_{i,\text{hist}}(t) \cup \mathcal{D}_{i,\text{diff}}(t), \tag{1}$$

where $\mathcal{D}_{i,\text{hist}}(t)$ contains several history anchors and $\mathcal{D}_{i,\text{diff}}(t)$ contains diffusion samples drawn from $\pi_\theta(\cdot \mid h_i(t), p_i(t))$. A distributional critic $Q_\psi\big(h_i(t), p_i(t), D\big)$ evaluates each candidate and selects

$$D_i^\star(t) = \arg\max_{D \in \mathcal{D}_i(t)} \mathbb{E}\big[Q_\psi\big(h_i(t), p_i(t), D\big)\big]. \tag{2}$$

The executed timing action is $\big(p_i(t), D_i^\star(t)\big)$, producing high-quality phase–duration trajectories over heterogeneous intersections.

2) RL Assisted Guidance for LLMs: The diffusion-based RL assistant also guides the LLM traffic signal controller. At step t and intersection i, it provides a reference proposal $a_i^{\text{RL}}(t) = \big(p_i^{\text{RL}}(t), D_i^{\text{RL}}(t)\big)$, where $p_i^{\text{RL}}(t)$ is selected by the *Pressure-Based Selector* and $D_i^{\text{RL}}(t)$ is generated/selected by the diffusion duration module with the critic. The proposal is injected into the LLM prompt. RL trajectories are further summarized by (i) a predictive rollout after applying $a_i^{\text{RL}}(t)$ yielding $\tilde{\mathbf{q}}_i(t+\Delta)$, and (ii) a historical time–queue mapping $D_i^*(\mathbf{q})$.

The critic provides preference scores for LLM decisions: for each $a_i^{\text{LLM}}(t)$ under $h_i(t)$, it computes

$$Q_\psi\big(h_i(t), a_i^{\text{LLM}}(t)\big), \tag{3}$$

which is converted into a priority label or sampling weight. These priorities rank, filter, and subsample the LLM interaction dataset so that higher-quality, intersection-specific timing plans are retained more frequently for adapter fine-tuning.

3) Priority Aware LoRA Imitation Fine Tuning: Interaction data are organized as priority-scored prompt–answer pairs

$$\mathcal{D} = \{(x_n, y_n, s_n)\}_{n=1}^{N}, \tag{4}$$

where x_n is the structured prompt describing the heterogeneous intersection state and RL reference signals, y_n is the corresponding reasoning and phase–duration decision, and s_n is a scalar priority derived from the RL critic. A filtered subset $\mathcal{D}_{\text{high}} = \{(x_n, y_n, s_n) \mid s_n \geq \kappa\}$ is used first, and lower-priority samples in $\mathcal{D} \setminus \mathcal{D}_{\text{high}}$ are introduced later via score-based sampling.

The backbone language model is the instruction-tuned Gemma-3-12B-it model. Low-rank adaptation (LoRA) injects traffic signal control knowledge by updating only adapter matrices with a standard autoregressive loss on the filtered data. The adapted LLM is referred to as the CuraLight agent.

During environment interaction, the CuraLight agent is trained under the guidance of both the diffusion-based RL assistant and a multi-LLM ensemble deliberation system. At each decision step t and intersection i, a set of candidate phase–duration actions $\{a_i^{(k)}(t)\}_k$ is first evaluated by the RL assistant, and clearly inferior actions are removed by an RL-based pre-filter, leaving $\tilde{\mathcal{A}}_i(t)$.

B. Training CuraLight with RL Filtering and Multi-LLM Ensemble Review

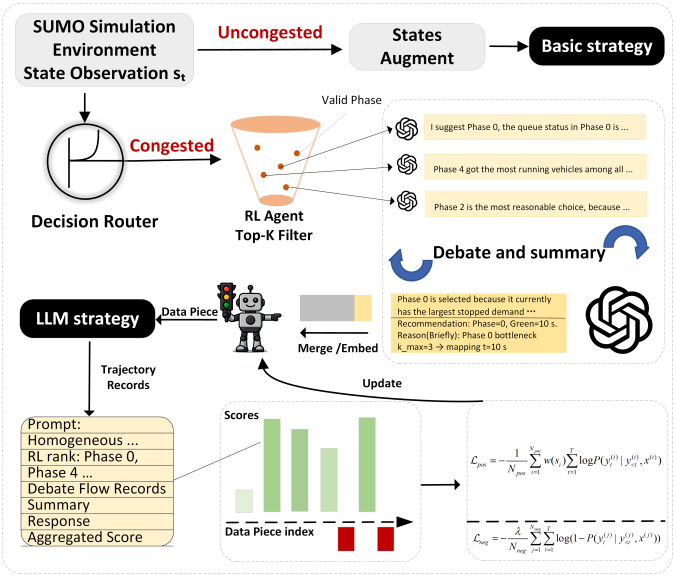


Fig. 4. Training Pipeline of CuraLight with RL Top-K Filtering and Multi-LLM Deliberation

Preference scoring module. The consensus LLM assigns a relative preference score to each $a \in \tilde{\mathcal{A}}_i(t)$, yielding $r_i^{\text{LLM}}(a)$. These scores are combined with RL critic values $q_i^{\text{RL}}(a)$ to obtain a unified priority

$$s_i(a) = \alpha\, f\big(q_i^{\text{RL}}(a)\big) + (1-\alpha)\, g\big(r_i^{\text{LLM}}(a)\big), \tag{5}$$

where $f(\cdot)$ and $g(\cdot)$ are normalized transformations and $\alpha \in [0,1]$ balances RL evaluation and ensemble preference. When CuraLight selects $a_i^{\text{CL}}(t)$, the transition is stored with priority $s_i\big(a_i^{\text{CL}}(t)\big)$. If an action has not passed through the ensemble deliberation, it is assigned a negative priority.

Second stage adapter fine-tuning. The unified priority scores are used as weights in a second stage of adapter fine-tuning.

Trajectories are partitioned into D^+ and D^- by the sign of their priority scores. Following [10], we optimize

$$\mathcal{L} = \frac{\sum_{i \in D^+} w_i \cdot \ell_{\mathrm{CE}}(y_i)}{\sum_{i \in D^+} w_i} + \lambda_{\mathrm{neg}} \cdot \frac{1}{|D^-|} \sum_{j \in D^-} \ell_{\mathrm{UL}}(y_j), \quad (6)$$

where $w_i \geq 0$ is obtained from the unified priority score, $\ell_{\mathrm{CE}}(\cdot)$ is the cross-entropy loss on y_i, and $\ell_{\mathrm{UL}}(\cdot)$ is the unlikelihood loss on D^-. λ_{neg} controls the repulsion strength. After convergence, the updated adapters are merged into the CuraLight agent, completing the second training stage that combines RL filtering, multi-LLM ensemble review, and priority-aware imitation for traffic signal control.

RL based pre-filtering. For each $a_i^{(k)}(t)$, the RL critic computes an approximate value and discards actions dominated under the current congestion pattern. The remaining set $\tilde{\mathcal{A}}_i(t)$ is forwarded to the ensemble deliberation system.

Defender LLMs. Each $a_i^{(k)}(t) \in \tilde{\mathcal{A}}_i(t)$ is assigned to a defender LLM, which receives the structured intersection state, competing actions, and historical summaries, and outputs a concise justification focusing on queue dissipation, balance across approaches, and anticipated network-level effects.

Consensus LLM. A consensus LLM aggregates defender arguments with the shared traffic context and produces a condensed comparison of candidates. This text is appended to the CuraLight prompt as a hidden auxiliary field: it is accessible to the model but cannot be directly quoted in the final explanation, encouraging internalization rather than copying.

III. Experiments and Results

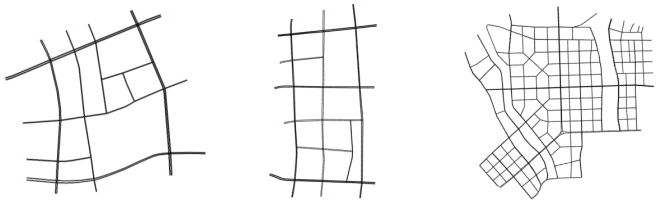
A. Experiments Settings

1) Performance metrics: ATT (s) denotes average travel time; AQL (m) denotes the network-wide average queue length; and AWT (s) denotes the average waiting time at intersections.

2) Compared Models: Baselines include transportation methods (Random, FixedTime [11], MaxPressure [12]), deep RL families (PressLight [13], MPLight [14], CoLight [15], Efficient-CoLight [16], Advanced-CoLight [17], AttendLight [18], π-Light [2], UniTSA [19]), Large-scale AI Models (Qwen3-32B, Llama-3-70B, Llama-4-Scout, Gemma-3-12B-it, Gemma-3-27B-it, DeepSeek-R1-14B, DeepSeek-R1-32B, GPT-5-mini), LLMLight [5], Traffic-R1 [20], and the proposed CuraLight.

3) Map traffic flow: Benchmarks are conducted on three real-world networks: **Jinan** (JN-1/2/3), **Hangzhou** (HZ-1/2), and **Yizhuang** (YZ-1/2), with traffic demands (vehicles/half hour) of 2150/2530/2814, 4000/4500, and 8000/10500, respectively.

4) Resource Usage: : Dual-socket Intel Xeon Platinum 8457C with a single NVIDIA L20 (48 GiB, CUDA 12.4). Under synchronous batching (avg. batch size $\approx$ 2.85; avg. input/output 960/356 tokens), a total token throughput of 242 tokens/s, and peak GPU memory usage of about 44,200 MiB.



(a) Jinan (b) Hangzhou (c) Yizhuang, Beijing

Fig. 5. Heterogeneous networks in three real-world scenarios.

B. Comparison Between CuraLight and Other TSC Methods

Table I shows that CuraLight delivers the most consistent gains across heterogeneous networks and ranks first on most scenario–metric combinations. Compared with FixedTime, MaxPressure, and deep RL baselines such as MPLight and CoLight variants, CuraLight achieves lower travel time, waiting time, delay, and queue length in most settings, indicating stronger adaptability to heterogeneous phase structures and demand patterns.

Compared with the strongest RL baseline UniTSA, CuraLight shows clearer advantages under heavy and heterogeneous traffic. On Hangzhou 2, it reduces ATT from 277.57 s to 262.14 s (5.56%) and AWT from 134.52 s to 118.80 s (11.69%), suggesting a more balanced allocation of service across approaches. An exception appears in Jinan 2: although CuraLight achieves the lowest AQL of 13.06 m, a 17.03% reduction from UniTSA (15.74 m), it does not obtain the best ATT or AWT, indicating a scenario-dependent trade-off between queue dissipation and end-to-end travel time.

C. Generalization Performance in Yizhuang

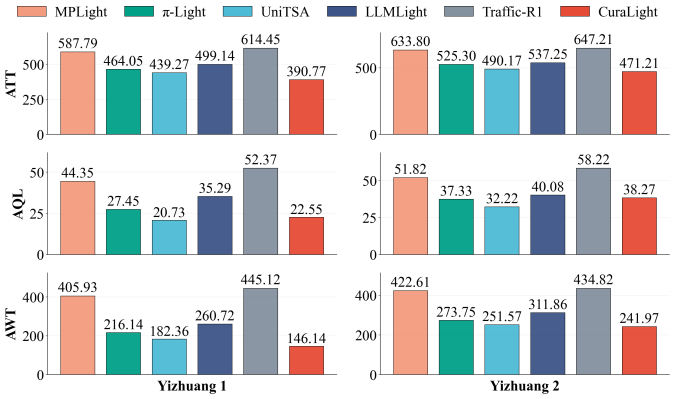


Fig. 6. Cross-Network Generalization from Hangzhou 2 to Yizhuang (177 Intersections)

Cross-network generalization on Yizhuang. Methods are trained on Hangzhou 2 and directly evaluated on the Yizhuang network with 177 intersections under two demand levels (Yizhuang 1/2), without further adaptation. CuraLight shows stronger zero-shot transfer, achieving lower travel time and intersection waiting than UniTSA. For instance, on Yizhuang 1, CuraLight reduces ATT by 11.04% and AWT by 19.86%

TABLE I

RESULTS ON **JINAN** (JN-1/2/3) AND **HANGZHOU** (HZ-1/2) ACROSS THREE METRICS. ALL VALUES ARE ROUNDED TO TWO DECIMAL PLACES. THE BEST RESULTS ARE IN **BOLD**, THE SECOND-BEST ARE UNDERLINED, AND THE THIRD-BEST ARE DOUBLE UNDERLINED.

Method	JN-1			JN-2			JN-3			HZ-1			HZ-2		
	ATT	AQL	AWT	ATT	AQL	AWT	ATT	AQL	AWT	ATT	AQL	AWT	ATT	AQL	AWT
Transportation Methods															
Random	538.07	45.10	431.61	500.11	56.63	403.28	520.62	56.89	415.73	549.15	79.56	412.62	579.36	83.66	432.14
FixedTime	354.94	27.71	245.06	363.13	33.13	252.00	371.43	36.47	259.37	473.00	70.32	311.45	425.75	66.47	267.11
MaxPressure	199.26	14.35	96.68	214.72	16.96	107.22	278.03	33.35	175.47	257.82	21.03	112.05	313.02	38.09	170.36
RL Methods															
MPLight	285.70	25.21	178.55	304.76	29.99	193.77	302.58	36.17	191.36	387.39	45.64	226.71	400.17	50.99	238.35
AttendLight	319.22	67.05	246.16	348.91	69.88	295.60	446.17	72.81	383.95	491.44	113.97	389.20	411.16	59.45	309.19
PressLight	333.63	27.06	220.03	335.98	36.79	225.83	331.65	38.12	220.51	378.67	45.46	222.43	420.83	63.49	263.62
CoLight	532.24	50.49	426.82	394.93	47.61	289.64	449.21	55.60	346.09	483.14	73.67	337.13	458.43	87.70	311.50
Efficient-CoLight	380.94	36.81	273.61	403.01	42.44	288.57	402.07	44.78	288.11	446.45	62.61	293.90	454.47	67.57	295.01
Advanced-CoLight	362.96	69.22	296.70	309.55	73.77	248.83	380.06	72.16	309.39	438.32	85.67	310.35	429.34	89.46	284.26
π-Light	164.64	13.96	68.04	202.89	24.71	87.28	226.37	30.07	104.64	328.95	33.08	153.07	384.65	44.38	189.86
UniTSA	171.45	12.82	71.59	186.75	15.74	84.45	200.75	18.18	96.27	255.48	23.80	115.13	277.57	29.18	134.52
Large-Scale AI Models															
Qwen3-32B	326.47	39.57	248.02	185.09	17.57	90.07	205.30	21.56	208.28	251.52	22.23	113.60	277.81	28.47	136.43
Llama-3-70B	158.61	11.49	64.28	200.51	18.42	104.30	223.17	22.86	125.90	247.14	21.76	110.12	279.79	27.92	136.96
Llama-4-Scout	156.93	12.38	61.64	184.10	17.97	84.59	215.57	25.11	122.03	244.95	22.04	106.35	286.87	30.70	141.55
Gemma-3-12b-it	163.58	13.49	70.23	242.15	20.35	94.45	273.66	37.21	150.28	269.42	30.13	130.64	315.42	39.78	165.25
Gemma-3-27b-it	158.46	12.78	65.45	223.06	16.95	84.41	197.32	19.78	95.79	267.44	24.52	126.67	300.93	31.36	152.32
DeepSeek-R1-14B	207.25	18.92	115.23	198.30	17.68	100.31	241.17	23.86	138.10	274.66	33.01	139.82	299.13	31.43	150.38
DeepSeek-R1-32B	169.56	12.88	73.55	199.82	18.05	103.37	280.32	31.02	172.80	248.72	21.69	111.84	289.11	30.15	144.62
GPT-5-mini	149.23	11.05	**53.16**	**176.52**	17.47	**81.75**	264.11	33.14	174.60	239.34	24.47	101.33	272.58	27.34	127.69
LLM-Based Methods															
LLMLight	235.01	18.13	134.20	343.01	34.20	250.47	450.40	51.91	369.22	328.34	34.17	190.71	334.85	38.10	191.67
Traffic-R1	308.88	33.65	239.46	323.01	27.92	263.98	287.78	35.43	228.91	343.72	40.52	222.25	405.43	55.16	280.87
CuraLight	**148.95**	**10.86**	54.60	212.19	**13.06**	116.60	**182.68**	**17.81**	**85.25**	**238.74**	**20.65**	**99.68**	**262.14**	**25.03**	**118.80**

relative to UniTSA. In contrast, AQL is slightly higher, indicating a mild trade-off that prioritizes end-to-end efficiency and waiting reduction over queue storage.

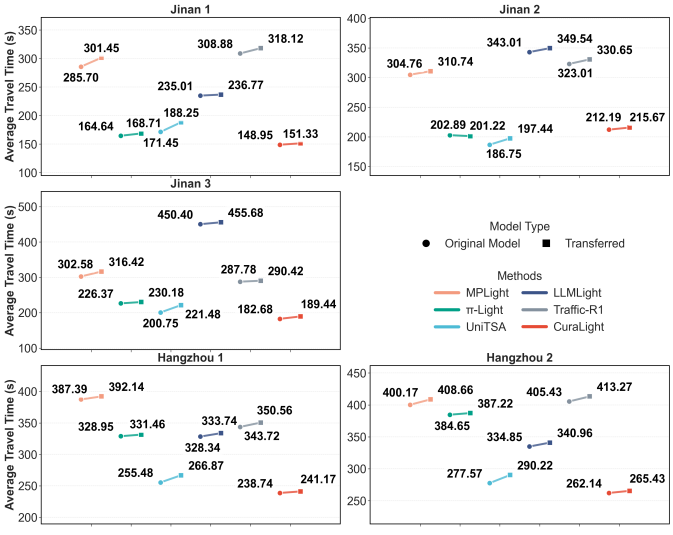


Fig. 7. Transferability Comparison on Jinan and Hangzhou

D. In-scenario transferability

As shown in Fig. 7, each *Original Model* (trained on the target scenario) is paired with a *Transferred* model reused from another setting and directly evaluated on the same scenario; the paired markers visualize the ATT change induced by transfer. Overall, reusing a model tends to increase ATT, reflecting the mismatch between the source training distribution and the target traffic dynamics and phase competition patterns.

CuraLight exhibits noticeably stronger robustness to this shift: its transferred performance degrades only mildly and it remains among the top methods across scenarios. For instance, on Hangzhou 2 the ATT increase of CuraLight is about 1.26%, while UniTSA rises by about 4.56%, indicating that CuraLight is less sensitive to distribution mismatch. This behavior is consistent with the design that emphasizes RL-filtered candidate timing actions and deliberation-based supervision, which help preserve stable timing preferences when the controller is reused rather than retrained.

E. Ablation Study

Table II compares the Gemma-3-12b-it backbone, an imitation-finetuned variant, and the multi-LLM ensemble deliberation system. Imitation fine-tuning reduces ATT, AQL, and AWT relative to the backbone, yielding about 19.3% average improvement, which suggests that RL-guided interaction

trajectories provide effective supervision for learning scenario-specific phase and duration patterns. Adding the multi-LLM ensemble deliberation system brings further gains across all metrics and achieves the best overall results, corresponding to about 28.5% average improvement over Gemma-3. This indicates that deliberation not only filters clearly inferior timing candidates, but also supplies more structured, consistency-oriented feedback that helps the agent avoid brittle choices under heterogeneous lane configurations. Overall, the ablation supports that imitation establishes a strong base policy, while deliberation-based screening and preference signals offer complementary benefits that further sharpen timing decisions.

TABLE II
ABLATION STUDY ON THE HANGZHOU 2 DATASET.

Model	ATT (s)	AQL (m)	AWT (s)
Gemma-3-12b-it	315.42	39.78	165.25
Imitation-Finetune	275.35	30.48	136.98
Multi-LLM Ensemble Deliberation System	**262.14**	**25.03**	**118.80**

IV. CONCLUSION

This paper presents CuraLight, an LLM-centered traffic signal control framework that combines a diffusion-based RL assistant with multi-LLM deliberation to provide high-quality supervision for heterogeneous intersections, improving the reliability and interpretability of phase and duration decisions. Experiments on real-world networks in Jinan, Hangzhou, and Yizhuang show consistent improvements in travel time, queue length, and waiting time over transportation, RL, and LLM-based baselines, with stronger transfer robustness; future work will reduce deliberation overhead for real-time deployment and extend the framework to emission-aware control and city-scale coordination.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62176024).

REFERENCES

[1] Q. Jiang, M. Qin, S. Shi, W. Sun, and B. Zheng, "Multi-agent reinforcement learning for traffic signal control through universal communication method," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, L. D. Raedt, Ed. International Joint Conferences on Artificial Intelligence Organization, 7 2022, pp. 3854–3860, main Track. [Online]. Available: https://doi.org/10.24963/ijcai.2022/535

[2] Y. Gu, K. Zhang, Q. Liu, W. Gao, L. Li, and J. Zhou, "π-light: Programmatic interpretable reinforcement learning for resource-limited traffic signal control," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 19, pp. 21 107–21 115, Mar. 2024. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/30103

[3] J. Yu, P.-A. Laharotte, Y. Han, and L. Leclercq, "Decentralized signal control for multi-modal traffic network: A deep reinforcement learning approach," *Transportation Research Part C: Emerging Technologies*, vol. 154, p. 104281, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X2300270X

[4] L. Da, M. Gao, H. Mei, and H. Wei, "Prompt to transfer: Sim-to-real transfer for traffic signal control with prompt learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 1, pp. 82–90, Mar. 2024. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/27758

[5] 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*, ser. KDD '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 2335–2346. [Online]. Available: https://doi.org/10.1145/3690624.3709379

[6] Z. Yuan, S. Lai, and H. Liu, "Collmlight: Cooperative large language model agents for network-wide traffic signal control," 2025. [Online]. Available: https://arxiv.org/abs/2503.11739

[7] T. Liang, Z. He, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, S. Shi, and Z. Tu, "Encouraging divergent thinking in large language models through multi-agent debate," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 17 889–17 904. [Online]. Available: https://aclanthology.org/2024.emnlp-main.992/

[8] A. Estornell and Y. Liu, "Multi-llm debate: Framework, principals, and interventions," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 28 938–28 964.

[9] J. Jung, F. Brahman, and Y. Choi, "Trust or escalate: LLM judges with provable guarantees for human agreement," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=UHPnqSTBPO

[10] S. Mukherjee, V. D. Lai, R. Addanki, R. A. Rossi, S. Yoon, T. Bui, A. Rao, J. Subramanian, and B. Kveton, "Offline RL by reward-weighted fine-tuning for conversation optimization," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. [Online]. Available: https://openreview.net/forum?id=WAFD6VYIEa

[11] P. Koonce, "Traffic signal timing manual," Tech Report FHWA-HOP-08-024, 2008.

[12] P. Varaiya, "Max pressure control of a network of signalized intersections," *Transportation Research Part C: Emerging Technologies*, vol. 36, pp. 177–195, 2013.

[13] H. Wei, C. Chen, G. Zheng, K. Wu, V. Gayah, K. Xu, and Z. Li, "Presslight: Learning max pressure control to coordinate traffic signals in arterial network," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1290–1298.

[14] C. Chen, H. Wei, N. Xu, G. Zheng, M. Yang, Y. Xiong, K. Xu, and Z. Li, "Toward a thousand lights: Decentralized deep reinforcement learning for large-scale traffic signal control," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, pp. 3414–3421, Apr. 2020.

[15] H. Wei, N. Xu, H. Zhang, G. Zheng, X. Zang, C. Chen, W. Zhang, Y. Zhu, K. Xu, and Z. Li, "Colight: Learning network-level cooperation for traffic signal control," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, ser. CIKM '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1913–1922.

[16] Q. Wu, L. Zhang, J. Shen, L. Lu, B. Du, and J. Wu, "Efficient pressure: Improving efficiency for signalized intersections," *ArXiv*, vol. abs/2112.02336, 2021.

[17] L. Zhang, Q. Wu, J. Shen, L. Lü, B. Du, and J. Wu, "Expression might be enough: representing pressure and demand for reinforcement learning based traffic signal control," in *International Conference on Machine Learning*. PMLR, 2022, pp. 26 645–26 654.

[18] A. Oroojlooy, M. Nazari, D. Hajinezhad, and J. Silva, "Attendlight: Universal attention-based reinforcement learning model for traffic signal control," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 4079–4090.

[19] M. Wang, X. Xiong, Y. Kan, C. Xu, and M.-O. Pun, "Unitsa: A universal reinforcement learning framework for v2x traffic signal control," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 10, pp. 14 354–14 369, 2024.

[20] X. Zou, Y. Yang, Z. Chen, X. Hao, Y. Chen, C. Huang, and Y. Liang, "Traffic-r1: Reinforced llms bring human-like reasoning to traffic signal control systems," 2025. [Online]. Available: https://arxiv.org/abs/2508.02344