# X-Light: Cross-City Traffic Signal Control Using Transformer on Transformer as Meta Multi-Agent Reinforcement Learner

**Haoyuan Jiang**[1*], **Ziyue Li**[2,†], **Hua Wei**[3], **Xuantang Xiong**[4*], **Jingqing Ruan**[4*], **Jiaming Lu**[5*], **Hangyu Mao**[6] and **Rui Zhao**[6]
[1]Baidu Inc., China
[2]University of Cologne, Germany
[3]Arizona State University, U.S.A
[4]Institute of Automation, Chinese Academy of Sciences, China
[5]Fudan University, China
[6]SenseTime Research, China
[*]*Research mainly done when with SenseTime Research,* [†]*Corresponding Author*
jianghaoyuan@zju.edu.cn, zlibn@wiso.uni-koeln.de.

## Abstract

The effectiveness of traffic light control has been significantly improved by current reinforcement learning-based approaches via better cooperation among multiple traffic lights. However, a persisting issue remains: how to obtain a multi-agent traffic signal control algorithm with remarkable transferability across diverse cities? In this paper, we propose a **T**ransformer **on T**ransformer (TonT) model for *cross*-city meta multi-agent traffic signal control, named as X-Light: We input the full Markov Decision Process trajectories, and the Lower Transformer aggregates the states, actions, rewards among the target intersection and its neighbors *within a city*, and the Upper Transformer learns the general decision trajectories *across different cities*. This dual-level approach bolsters the model's robust generalization and transferability. Notably, when directly transferring to unseen scenarios, ours surpasses all baseline methods with **+7.91%** on average, and even **+16.3%** in some cases, yielding the best results. The code is here.

## 1 Introduction

An effective traffic signal control (TSC) system is the key to alleviating traffic congestion. In recent years, Reinforcement Learning (RL) has been widely used in the field of TSC. It can interact with the environment, explore, and exploit, which helps agents discover better policies without artificial priors and assumptions. Numerous studies [Zheng *et al.*, 2019; Chen *et al.*, 2020; Wei *et al.*, 2019; Lu *et al.*, 2023; Du *et al.*, 2024] have demonstrated its noteworthy enhancements over conventional rule-based methods [Hunt *et al.*, 1982; Lowrie, 1990; Roess *et al.*, 2004; Smith *et al.*, 2013].

[1]The extended version of this paper with appendices is available at https://arxiv.org/abs/2404.12090.

However, most of the existing methods are *scenario-specific*, meaning that the training and testing should be in the same scenario (A scenario is a simulation environment, e.g., a virtual city, with a set of intersections). When deploying on a new scenario, rebuilding the environment and re-training are needed, which involves significant costs. As a result, to the best of our knowledge, all the cities still use rule-based methods such as SCATS or SCOOT for a very fundamental reason: they can be easily reused in a new district/region/city.

This raised a critical problem that **how to orchestrate the multiple intersections for various scenarios/cities with strong generalizability.**

There are some solutions for single-agent settings: MetaLight [Zang *et al.*, 2020] and GESA [Jiang *et al.*, 2024] used one single agent to control all the intersections in a scenario and achieve transferability via gradient-based meta RL and multi-city co-training, respectively. Their drawbacks are obvious, i.e., neglecting the cooperation among the multiple intersections. Yet, learning a general Multi-Agent RL model for various scenarios is non-trivial, given various scenarios could have various road networks and traffic dynamics, rendering various environment states for each agent and collaboration patterns for multiple agents. To the best of our knowledge, only a few existing works target the same challenge: MetaVIM [Zhu *et al.*, 2023] and MetaGAT [Lou *et al.*, 2022]. They both utilized the Markov Decision Processes (MDPs) trajectories to help the agents learn and distinguish scenario context. However, they still display limitations, such as an unstable training process and large performance drops when encountering quite dissimilar scenarios (details in Sec. **??**).

To enhance cooperation and generalizability, we will incorporate the full MDP trajectories, including the observations, rewards, and actions $(o, a, r)$ of both the target intersection and its neighbors, into the method. Given the sequence nature of the trajectories from various MDPs, two natural questions are: **Q1: Can Transformer utilize the** $o, a, r$ **sequences of multiple intersections for better collaboration? Q2: Furthermore, can we Transformer learn the high-level cross-scenario MDPs dynamics for better transfer-**

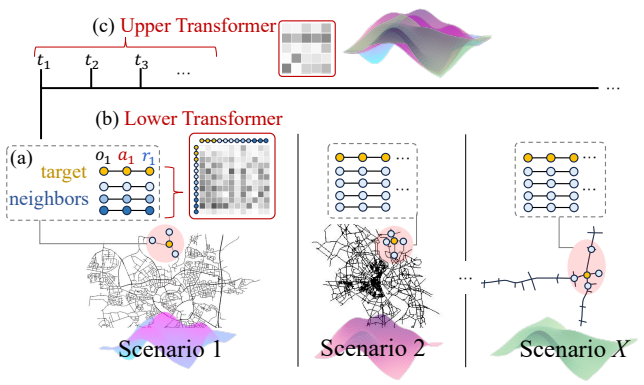


Figure 1: (a) X-Light takes the MDP $o, a, r$ trajectories of the target and its neighbors: (b) the Lower Transformer learns the attention for all the $o, a, r$-s, so that, e.g., one intersection's $o$ may have high attention with another intersection's $a$; (c) Upper Transformer learns the attention over the time through all different scenarios.

**ability.** This forms our solution as in Figure 1: a Transformer on Transformer (TonT) model for Meta Multi-Agent Reinforcement Learners. The **Lower Transformer** extracts single-step trajectory information from the target intersection and its neighbor intersections to encourage cooperation. The **Upper Transformer** learns from the historical multi-scenario multi-modal MDPs distributions and makes actions.

The main contributions are three-fold:

- In the domain of TSC, we propose the first-ever Transformer-on-Transformer (TonT) framework for meta MARL, which solves both multi-intersection collaboration and cross-city transferability/generalizability.
- Specifically, the **Lower Transformer** aggregates the target and its neighbors' fine-grained $o, a, r$ information and achieves better collaboration than other solutions such as the Graph Neural Network (GNN)-based methods [Wei *et al.*, 2019; Lou *et al.*, 2022], which only aggregates traffic states $o$; The **Upper Transformer**, together with a *dynamic prediction* pretext task and *multi-scenario co-training scheme*, learns the scenario-agnostic decision process and achieves better cross-city decision. Additionally, a *residual link* is added before inputting to actor-critic for better decision.
- We conduct rigid experiments with various scenarios and zero-shot transfer to each, and our method is constantly the best performer. In non-transfer settings, we also achieve the most top one results.

## 2 Methodology

This section delineates our proposed method, X-Light, a general cross-city multi-agent traffic signal control method.

As shown in Fig. 2, our method is trained using intersections across various scenarios within a batch. The $i$-th intersection and its neighbors $\mathcal{N}_i$ are selected, and their MDP trajectories $(o, a, r)$ from time frame $[t-K+1, t]$ are sampled and fed into TonT Encoder module. It is worth noting that to allow the model to handle various intersections, we use the GPI module proposed in [Jiang *et al.*, 2024], which maps various intersections' structure into a unified one, followed by an MLP to obtain $o_t^{\{i,\mathcal{N}_i\}}$.

As shown in Fig. 2, the TonT Encoder employs two types of transformers: the Lower Transformer and the Upper Transformer. The primary role of the Lower Transformer is to integrate the target and its neighbors' MPD information at time step $t$. It enhances agent collaboration compared with GNN-based collaboration by 6%-13%. The Upper Transformer utilizes historical trajectory information as context to infer the current task, thereby achieving improved transferability.

The output of the TonT Encoder is then utilized by both the Actor and Critic to output the policy $\pi$ for executing the action and to estimate the state value for training.

Similar to [Wei *et al.*, 2021], we choose five features from the observations as states: queue length, current phase, occupancy, flow, and the number of stopping cars. The action is to choose the eight pre-defined phases for the next time interval, as shown in Fig 2. At each time step $t$, the agent $i$ can choose to execute action $a_t^i$ from available action set $\mathcal{A}^i$ in the next $\Delta t$ seconds. In our experiments, we set $\Delta t$ as 15 seconds. The reward $r$ is defined as the weighted sum of queue length, wait time, delay time, and pressure.

## 3 To Recap Main Results

Our model achieves the best transfer results in all scenarios. **(1) Cooperation is needed:** Compared with single agent methods, e.g., MetaLight, cooperation is needed for better transferability. **(2) TonT is better for Meta MARL:** compared to the second-best MetaGAT, ours achieved a **+7.91%** improvement on average and **+16.3%** in *Grid5×5*, mainly because our unified Transformer on Transformer design captures collaborators' $o, a, r$ interdependency for better local cooperation, and global cross-scenario dynamics through multi-scenario co-training, respectively. Yet, MetaGAT only focuses on the neighbors' $o$ interdependency and lacks a scenario-agnostic training scheme. *In some cases (Grid5×5, Ingostadt21), MetaGAT cannot beat MetaLight, which means that when cooperation is not as well designed as ours, it can even worsen the transferability when dealing with multi-scenarios.* **(3) Lower Transformer is better for cooperation:** ours$_{GNN}$ flattens $o, a, r$, thus losing the $o, a, r$ interdependency, further highlighting the necessity of the Lower Transformer.

## References

[Chen *et al.*, 2020] Chacha Chen, Hua Wei, Nan Xu, Guanjie Zheng, Ming Yang, Yuanhao Xiong, Kai Xu, and Zhenhui Li. Toward a thousand lights: Decentralized deep reinforcement learning for large-scale traffic signal control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3414–3421, 2020.

[Du *et al.*, 2024] Xinqi Du, Ziyue Li, Cheng Long, Yongheng Xing, S Yu Philip, and Hechang Chen. Felight: Fairness-aware traffic signal control via sample-efficient reinforcement learning. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

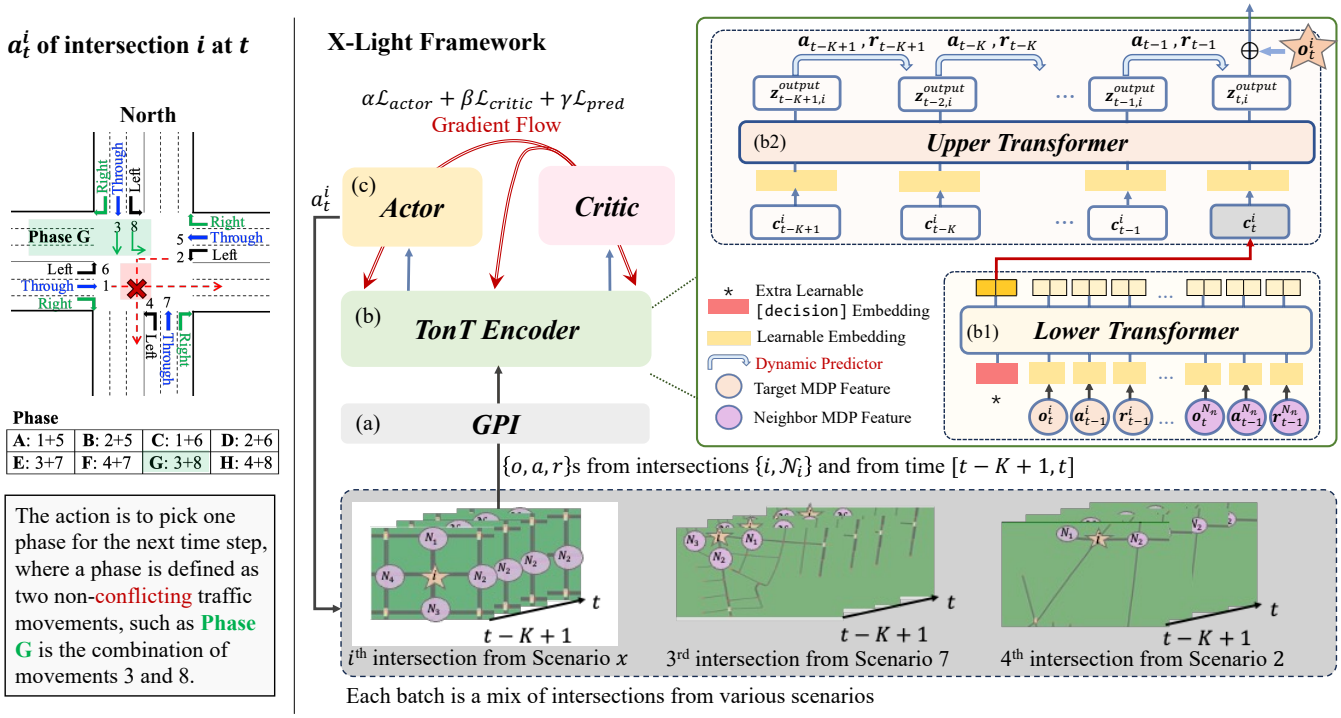


Figure 2: Our method is co-trained with intersections' MDPs from various scenarios: (a) a GPI module unifying all the scenarios, (b) the proposed TonT Encoder, and (c) an actor-critic to make a decision. The TonT Encoder contains (b1) a Lower Transformer aggregating the $o$, $a$, and $r$ among the target and its neighbors and (b2) an Upper Transformer learning general decisions from multi-scenario historical MDPs.


[Hunt *et al.*, 1982] PB Hunt, DI Robertson, RD Bretherton, and M Cr Royle. The SCOOT on-line traffic signal optimisation technique. *Traffic Engineering & Control*, 23(4), 1982.

[Jiang *et al.*, 2024] Haoyuan Jiang, Ziyue Li, Zhishuai Li, Lei Bai, Hangyu Mao, Wolfgang Ketter, and Rui Zhao. A general scenario-agnostic reinforcement learning for traffic signal control. *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[Lou *et al.*, 2022] Yican Lou, Jia Wu, and Yunchuan Ran. Meta-reinforcement learning for multiple traffic signals control. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4264–4268, 2022.

[Lowrie, 1990] PR Lowrie. SCATS, Sydney co-ordinated adaptive traffic system: A traffic responsive method of controlling urban traffic. 1990.

[Lu *et al.*, 2023] Jiaming Lu, Jingqing Ruan, Haoyuan Jiang, Ziyue Li, Hangyu Mao, and Rui Zhao. Dualight: Enhancing traffic signal control by leveraging scenario-specific and scenario-shared knowledge. *arXiv preprint arXiv:2312.14532*, 2023.

[Roess *et al.*, 2004] Roger P Roess, Elena S Prassas, and William R McShane. *Traffic engineering*. Pearson/Prentice Hall, 2004.

[Smith *et al.*, 2013] Stephen F Smith, Gregory Barlow, Xiao-Feng Xie, and Zachary B Rubinstein. Surtrac: Scalable urban traffic control. 2013.

[Wei *et al.*, 2019] Hua Wei, Nan Xu, Huichu Zhang, Guanjie Zheng, Xinshi Zang, Chacha Chen, Weinan Zhang, Yanmin Zhu, Kai Xu, and Zhenhui Li. Colight: Learning network-level cooperation for traffic signal control. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1913–1922, 2019.

[Wei *et al.*, 2021] Hua Wei, Guanjie Zheng, Vikash Gayah, and Zhenhui Li. Recent advances in reinforcement learning for traffic signal control: A survey of models and evaluation. *ACM SIGKDD Explorations Newsletter*, 22(2):12–18, 2021.

[Zang *et al.*, 2020] Xinshi Zang, Huaxiu Yao, Guanjie Zheng, Nan Xu, Kai Xu, and Zhenhui Li. Metalight: Value-based meta-reinforcement learning for traffic signal control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1153–1160, 2020.

[Zheng *et al.*, 2019] Guanjie Zheng, Yuanhao Xiong, Xinshi Zang, Jie Feng, Hua Wei, Huichu Zhang, Yong Li, Kai Xu, and Zhenhui Li. Learning phase competition for traffic signal control. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1963–1972, 2019.

[Zhu *et al.*, 2023] Liwen Zhu, Peixi Peng, Zongqing Lu, and Yonghong Tian. Metavim: Meta variationally intrinsic motivated reinforcement learning for decentralized traffic signal control. *IEEE Transactions on Knowledge and Data Engineering*, 2023.