Discovering Programmatic Policies from Reinforcement Learning-Based Traffic Signal Controllers

Lindong Xie[1,2] Yang Zhang[1] Beiyu Song[1]
Zeren Xing[1] Flora D. Salim[2] Edward Chung[1,*]

[1]The Hong Kong Polytechnic University

[2]University of New South Wales

* Corresponding author

Motivation: explicit traffic control rules

Traffic signal control (TSC) is essential for efficient urban transportation.

Deep reinforcement learning (DRL) is widely used for TSC, learning signal-control policies through interaction with the traffic environment.

Deployment challenges

- **Interpretability:** Neural decisions are difficult to explain.
- **Generalization:** Performance can degrade on new flows or networks.
- **Efficiency:** Neural inference requires more storage and computation.

Our goal

Distill DRL controllers into compact, readable control rules.

Student policy: score each signal phase

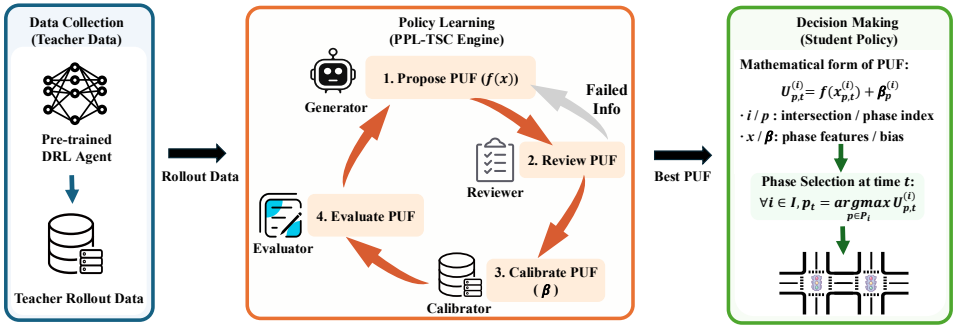
Input: Traffic features for each candidate phase.

Output: The signal phase with the highest utility.

$$U_{p,t}^{(i)} = \underbrace{f(\mathbf{x}_{p,t}^{(i)})}_{\text{shared symbolic rule}} + \underbrace{\beta_p^{(i)}}_{\text{local bias}}, \qquad p_t^{(i)} = \arg\max_p U_{p,t}^{(i)}$$

- Traffic features describe the movements served by each phase.
- One explicit rule captures common control logic.
- Calibrated biases adapt the rule to individual intersections.

Learning the symbolic policy



1. **Generate** symbolic candidates with an LLM.
2. **Review** validity and repair candidates.
3. **Calibrate** local biases using teacher rollouts.
4. **Evaluate** fidelity and traffic performance; evolve.

Traffic performance and teacher fidelity

CityFlow: 7 networks, 5 urban areas, 5 teacher families.

Student	Mean ATT reduction	ATT wins	Fidelity
Tailored	**7.62%**	113/175	0.64
Unified	4.91%	95/175	0.58

- **Tailored:** one student per teacher–scenario pair.
- **Unified:** one student per teacher across scenarios.

ATT: average travel time; lower is better. Wins include ties.
Fidelity: agreement with teacher actions; higher is better.

Generalization beyond the source conditions

Transfer setting	Mean ATT reduction	ATT wins
Different flow, same scenario	9.76%	423/700
Different scenario	**48.37%**	2787/3150

- Improvements are relative to the corresponding transferred teachers.
- Cross-scenario results cover EcoLight, FRAP, and MPLight.

Comparison scope follows the original report slides; EcoLight uses phase selection.

Readable rules with compact deployment

Example: MPLight student on Hangzhou3

$$U_p = \underbrace{\sum_{d \in \mathcal{R}_p} \hat{p}_{d,t}}_{\text{aggregate pressure}} + \underbrace{\max_{d \in \mathcal{R}_p} \hat{p}_{d,t}}_{\text{bottleneck pressure}} + \underbrace{0.15\,\mathbf{1}[p = p_{\text{cur}}]}_{\text{phase retention}} + \beta_p$$

Higher aggregate and bottleneck pressures increase a phase's priority. A small bonus favors keeping the current phase, while calibrated biases adapt priorities to the intersection.

ATT: 86.01 $\rightarrow$ **84.04** seconds Fidelity: **0.93**

PPL-TSC

Explicit control rules, competitive traffic performance, and compact inference.