This CVPR Findings paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the accepted version; the final published version of the proceedings is available on IEEE Xplore.

C^{2}T: Captioning-Structure and LLM-Aligned Common-Sense Reward Learning for Traffic–Vehicle Coordination

Yuyang Chen[1], Kaiyan Zhao[2], Yiming Wang[4], Ming Yang[3], Bin Rao[1], Zhenning Li[1*]

[1]The State Key Laboratory of Internet of Things for Smart City, University of Macau
[2]Wuhan University [3]Hong Kong Polytechnic University
[4]Computer and Information Science, University of Macau

zhenningli@um.edu.mo

Abstract

State-of-the-art (SOTA) urban traffic control increasingly employs Multi-Agent Reinforcement Learning (MARL) to coordinate Traffic Light Controllers (TLCs) and Connected Autonomous Vehicles (CAVs). However, the performance of these systems is fundamentally capped by their handcrafted, myopic rewards (e.g., intersection pressure), which fail to capture high-level, human-centric goals like safety, flow stability, and comfort. To overcome this limitation, we introduce C2T, a novel framework that learns a common-sense coordination model from traffic–vehicle dynamics. C2T distills "common-sense" knowledge from a Large Language Model (LLM) into a learned intrinsic reward function. This new reward is then used to guide the coordination policy of a cooperative multi-intersection TLC MARL system on CityFlow-based multi-intersection benchmarks. Our framework significantly outperforms strong MARL baselines in traffic efficiency, safety, and an energy-related proxy. We further highlight C2T's flexibility in principle, allowing distinct "efficiency-focused" versus "safety-focused" policies by modifying the LLM prompt.

1. Introduction

Urban traffic control has long been a central topic in intelligent transportation. Reinforcement-learning (RL)–based traffic light controllers typically optimize local efficiency by minimizing queue length, intersection pressure, or average delay [18, 19]. While such handcrafted objectives can deliver short-term gains, they often overlook network-level dynamics. When each intersection pursues its own local optimum, coordination across the city breaks down, producing congestion waves, unstable flows, and frequent braking. In practice, controllers that aggressively clear volume may underweight smoothness, safety, and comfort, yielding traffic that looks efficient on paper but feels brittle and unpredictable in deployment.

Coordinated control aims to overcome these limitations by letting signals and vehicles shape one another's behavior: signals create openings and platoons, while vehicles adjust approach speeds and lane usage to match those plans. CoTV-style systems [4] and related work on coordinated signal control [19] show that such coordination can improve travel time and energy in simulation, even when vehicles themselves are not explicitly modeled as learning agents. Yet the underlying reward recipe is usually unchanged: signals are scored by pressure or queue imbalance, and vehicles by speed tracking and acceleration penalties. These proxies favor aggressive clearing but say little about stability, safety, or comfort, leading to oscillatory phase changes, abrupt braking, and unsafe spacing even when average efficiency improves. What is missing is a notion of *traffic quality* that reflects human judgment and anticipates longer-horizon effects (e.g., platoon formation), and that can be learned offline while remaining compatible with standard multi-intersection RL training.

We take traffic quality itself as supervision and gather it before training. Evidence from other domains shows that large language models (LLMs) and related systems can make consistent pairwise judgments when asked to compare well-structured state descriptions [2, 9]. Building on this observation, we propose **C^{2}T**, a framework that renders each observation into a deterministic, unit-aware caption; samples caption pairs across time and approaches; and queries an LLM under a fixed prompt to decide which state better reflects good traffic. We retain only decisive labels and fit a lightweight preference scorer that outputs an intrinsic reward, with optional heads emphasizing safety, efficiency, or energy. All supervision is produced offline and cached. At training time, the intrinsic signal is added to a standard multi-intersection PPO pipeline as an additional reward stream—no simulator changes and no online LLM calls. In this paper, we instantiate **C^{2}T** on CityFlow-based

benchmarks [24] where each intersection-level traffic-light controller (TLC) is a learning agent and vehicles follow the built-in microscopic dynamics, but the same design extends naturally to settings with explicitly modeled connected autonomous vehicles (CAVs).

A key design choice is to mix the intrinsic reward *asymmetrically*: TLCs receive a blend of environment reward and the learned score, whereas vehicles remain environment driven in our instantiation. This places the human-aligned guidance where phase selection most strongly shapes platoons and network rhythm, and avoids extra non-stationarity that would arise if both sides were driven by a changing intrinsic signal. Safety and stability are enforced by a lightweight risk mask that suppresses the intrinsic signal under low time-to-collision (TTC) percentiles, clusters of harsh deceleration, or red-light risk flags parsed from the caption. For optimization, each reward stream is normalized and softly clipped before advantage estimation, and the mixing weight is scheduled so the policy first satisfies environment constraints and then gradually absorbs common-sense preferences. Because the scorer is trained offline, switching prompts or enabling safety/efficiency/energy heads only changes the cached scores, leaving the RL loop unchanged.

Finally, we evaluate **C^{2}T** on CityFlow-based benchmarks following the LLMLight protocol. We consider three real-city networks (Jinan, Hangzhou, and New York) and two stress-test scenarios (extreme demand surges and 24-hour diurnal cycles), and we reuse the same training pipeline and random seed protocol for fair comparison. Across demand levels and seeds, **C^{2}T** reduces average travel time and increases throughput while improving safety, as indicated by higher TTC percentiles and fewer harsh-braking events; an energy-related proxy also improves. Ablations show that the intrinsic signal, the safety mask, and the combination of per-stream normalization with a simple schedule are all necessary for stable gains. Captions alone already help, and adding matched numeric slots yields further (smaller) improvements—underscoring the value of structured descriptions for learning traffic quality offline and integrating it into existing multi-intersection RL training without altering the simulator or incurring online LLM calls.

In summary, this work delivers:

- **Structured captioning for traffic RL:** a deterministic, unit-aware schema that makes state quality comparable by LLMs and reproducible across runs.
- **Offline preference learning for an intrinsic reward:** a lightweight scorer trained from decisive LLM preferences on caption pairs, with optional heads for safety, efficiency, and energy, and no online LLM calls.
- **Safe, asymmetric integration in multi-intersection RL:** intrinsic reward mixed only into the TLC objective, combined with a risk mask, per-stream normalization, soft clipping, and a simple schedule—fully compatible with standard PPO.
- **Evidence across benchmarks:** gains on CityFlow-based Jinan, Hangzhou, and New York networks, plus stress scenarios and ablations that clarify the role of the intrinsic signal, safety mask, and optimization choices.

2. Related Work

2.1. Multi-Agent Reinforcement Learning for Traffic Signal Control

RL-based traffic signal control started from single-intersection agents with handcrafted efficiency rewards such as delay or queue length, e.g., IntelliLight and the max-pressure-based PressLight [18, 19]. Later work focuses on coordination and scalability: CoLight propagates information via graph attention, FRAP models phase competition, and MPLight scales pressure-based control to city networks [1, 20, 27], with further refinements through attention architectures, meta-learning, and improved pressure features in AttendLight, MetaLight, Efficient-CoLight, and Advanced-CoLight [8, 21, 23, 25]. CoTV additionally couples TLC and CAV control under a composite reward over travel time, fuel, and emissions [4], but all these methods ultimately optimize fixed metric-based scalar rewards, whereas C^{2}T learns a commonsense intrinsic reward from LLM preferences over structured TLC–CAV descriptions. More broadly, recent RL work has improved learning via better experience replay [26], latent exploration [14, 15, 17], and goal-conditioned generalization [16]. These methods, however, focus on generic sample efficiency, exploration, or transfer rather than human-aligned traffic-quality reward design.

2.2. Reward Modeling and Preference Learning in Reinforcement Learning

Handcrafted rewards are fragile proxies for designer intent and often induce reward hacking in complex domains such as urban traffic. IRL methods like GAIL and AIRL infer rewards from expert demonstrations [3, 5], while preference-based RL learns a reward model from pairwise human comparisons [2], but both are hard to scale to large multi-agent traffic systems due to data demands, ambiguity, and computational cost. C^{2}T instead adopts an AI-assisted preference-learning view: it queries an LLM on deterministic, unit-aware captions and distills its judgments into an intrinsic reward tailored to TLC–CAV coordination.

2.3. Large Language Models for Decision Making and Reward Alignment

Recent work uses LLMs as high-level decision makers by mapping states to text and decoding actions, e.g., LLM-

Light for traffic signals and holistic vision–language driving with counterfactual reasoning in OmniDrive [7, 13], but directly placing LLMs in the control loop raises latency, reliability, and scalability issues in multi-agent settings. A complementary line treats LLMs as judges: DriveLM evaluates driving decisions via graph-based QA [12], and RLHF/RLAIF-style approaches train policies against reward models learned from human or AI feedback [9, 22], with Poutine-style systems showing that vision–language models plus light preference tuning can yield strong driving performance [22]. Existing LLM-as-judge work, however, typically operates on free-form descriptions or QA pairs for single agents, whereas C^{2}T uses deterministic schema-based captions to obtain stable pairwise preferences that can be distilled into a reusable intrinsic reward for MARL.

2.4. Structured Scene Representation and Driving Captions

Graph-based and object-centric representations factor traffic scenes into entities and relations, and are widely used in RL for phases, lanes, and movements, as in FRAP and CoLight [20, 27], but these formats are designed for neural nets rather than LLMs. Vision-based driving captioning and VQA datasets, such as nuScenes caption/QA benchmarks and DriveLM, provide natural language descriptions of driving scenes, and OmniDrive extends them to counterfactual reasoning [6, 12, 13]; yet free-form captions are often stylistic and under-specified, lacking precise quantitative details needed for consistent reward evaluation. C^{2}T bridges this gap by constructing deterministic, unit-aware, schema-constrained traffic captions that enumerate key variables (e.g., queues, delays, TTC, violations) with explicit semantics, making them both faithful to the underlying state and well-suited for stable LLM-based preference learning.

3. Preliminaries

We briefly review the reinforcement learning formulation for traffic signal control, from the standard single-agent MDP view to a cooperative multi-agent setting with traffic lights (TLCs) and connected autonomous vehicles (CAVs). This section only fixes notation and highlights the limitations of handcrafted rewards that motivate C^{2}T.

3.1. Traffic Signal Control as an MDP

We model a single intersection as a Markov Decision Process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r_{\text{ext}}, \gamma)$. At time step t, the TLC observes a local traffic state $s_t \in \mathcal{S}$ containing the current phase, its elapsed time, and lane-level statistics such as queue length and flow. It then chooses an action $a_t \in \mathcal{A}$ (e.g., `keep` or `switch` phase), and the microscopic simulator applies $\mathcal{P}(s_{t+1} \mid s_t, a_t)$.

The external reward $r_{\text{ext}}(s_t, a_t)$ is a handcrafted congestion proxy. A common choice is negative intersection pressure:

$$r_{\text{ext}}(s_t) = -\sum_{i=1}^{K} \left(n_i^{in} - n_i^{out}\right), \tag{1}$$

where n_i^{in} and n_i^{out} are the inflow and outflow vehicle counts on approach i. The TLC learns a policy $\pi(a_t \mid s_t)$ that maximizes the discounted return $\mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r_{\text{ext}}(s_t, a_t)]$.

This formulation has been widely used in RL-based traffic signal control, but it implicitly treats vehicles as passive entities and only optimizes short-term efficiency metrics defined by r_{ext}.

3.2. Multi-Agent Traffic Control and Its Limitations

Modern intelligent transportation systems increasingly coordinate both infrastructure and vehicles. We therefore consider a multi-agent setting with TLC agents $M = \{1, \ldots, |M|\}$ and (potential) CAV agents $N = \{1, \ldots, |N|\}$. At time t, each agent $i \in M \cup N$ receives a local observation o_t^i, selects an action a_t^i, and the shared traffic state evolves as

$$s_{t+1} = f\left(s_t, \{a_t^i\}_{i \in M \cup N}\right).$$

TLC agents aim to reduce local congestion via external rewards such as

$$r_{\text{ext}}^{\text{TL}}(s_t) = -\sum_{i=1}^{K} \left(n_i^{in} - n_i^{out}\right) - \lambda \cdot \text{delay}_t,$$

while CAV agents, when explicitly controlled, are typically encouraged to track a desired speed with smooth accelerations, e.g.,

$$r_{\text{ext}}^{\text{CAV}}(s_t) = -\left|v_t - v^*\right| - \beta \left\|a_t\right\|.$$

The cooperative objective is to optimize all agents jointly:

$$\max_{\pi_M, \pi_N} \mathbb{E}\left[\sum_{t=0}^{T} \sum_{i \in M \cup N} \gamma^t r_t^i\right], \tag{2}$$

where π_M and π_N are the (shared) policies for TLCs and CAVs.

However, both the single-agent and multi-agent formulations above are still driven by simple, handcrafted rewards $r_{\text{ext}}^{\text{TL}}$ and $r_{\text{ext}}^{\text{CAV}}$ that focus on local pressure, delay, or speed tracking. They provide limited guidance about broader notions of safety, comfort, and flow stability, and may therefore produce policies that are numerically optimal yet misaligned with human common sense.

This motivates our central goal: to learn a *human-aligned* intrinsic reward r_ϕ from preferences expressed by large language models, and to integrate it with the external rewards in multi-agent traffic control. The next section introduces the C^{2}T framework that realizes this idea.

4. Method

4.1. Overview

We propose **C^{2}T** (*Captioning-Structure and LLM-Aligned Common-Sense Reward Learning*). C^{2}T has three stages: (**Stage 1**) render simulator observations into structured captions and build preference pairs; (**Stage 2**) learn a common-sense intrinsic reward r_ϕ from LLM preferences; (**Stage 3**) inject r_ϕ into the RL loop via asymmetric reward mixing with explicit safety masks. We instantiate C^{2}T on CityFlow-based multi-intersection benchmarks where each traffic-light controller (TLC) is a learning agent while vehicles follow the built-in microscopic dynamics. In this work we focus on TLC-only learning and discuss extensions to explicitly modeled connected autonomous vehicles (CAVs) in the appendix.

Figure 1 summarizes the end-to-end pipeline. We next detail each stage.

4.2. Stage 1: Structured Captioning and Preference Pairs

Let $o \in \mathcal{O}$ be an observation at time t for a given intersection. We define a deterministic captioner

$$c \;=\; \mathcal{C}(o;\, \Xi), \tag{3}$$

where Ξ is a fixed schema (fields and units). The schema concatenates structured slots (no free-form prose):

$$c = \Big[\, \underbrace{\text{phase},\ \text{elapsed}}_{\text{signal}},\ \underbrace{\{q_i,\, p_i\}_{i=1}^{K}}_{\text{queues/pressure}},\ \underbrace{\bar{d},\ \bar{\tau}}_{\text{delay/throughput}}\ ,\ \underbrace{\mathrm{TTC}_{\mathrm{p10}},\ \mathrm{TTC}_{\mathrm{p50}},\ h_{\text{brake}},\ \rho_{\text{red}}}_{\text{safety}},\ \underbrace{v_{\text{near}},\ a_{\text{near}},\ d_{\text{stop}}}_{\text{nearest vehicle}}\,\Big]. \tag{4}$$

Here q_i and p_i denote queue and pressure on approach i, $\bar{d}$ average delay, $\bar{\tau}$ throughput, $\mathrm{TTC}_{\mathrm{p}\alpha}$ the α-percentile time-to-collision, h_{brake} harsh braking count, ρ_{red} a red-light risk flag, and $(v_{\text{near}}, a_{\text{near}}, d_{\text{stop}})$ the kinematics of the nearest vehicle to the stop line.

From a pool $\mathcal{D}_{\text{obs}} = \{o\}$, we construct caption pairs $\mathcal{P} = \{(c_1, c_2)\}$ via a distribution prioritizing congestion/phase contrast, safety contrast, and short time shifts at the same junction:

$$\begin{aligned} p_{\text{pair}}\big((o_a, o_b)\big) \;\propto\; & \alpha\, \Delta_{\text{cong}}(o_a, o_b) \;+\; \beta\, \Delta_{\text{safety}}(o_a, o_b) \\ & +\; \gamma\, \mathbb{I}\big\{ j(o_a){=}j(o_b),\ |t_a{-}t_b| \in [\delta_1, \delta_2] \big\}, \end{aligned} \tag{5}$$

where $j(\cdot)$ maps an observation to its intersection index and $\Delta_{\text{cong}}, \Delta_{\text{safety}}$ are standardized contrasts (e.g., pressure gap, TTC-percentile gap). In our experiments, Δ_{cong} aggregates gaps in average queue length and delay, while Δ_{safety} aggregates gaps in TTC percentiles and harsh braking counts. We set $(c_1, c_2) = \big(\mathcal{C}(o_a),\, \mathcal{C}(o_b)\big)$.

Algorithm 1 Stage 1: Captioning and Pair Sampling

Require: Simulator $\mathcal{E}$, captioner $\mathcal{C}$, pairing distribution p_{pair}, episodes K
Ensure: Observation buffer $\mathcal{D}_{\text{obs}}$, pair set $\mathcal{P}$

```
D_obs ← ∅, P ← ∅
for k = 1 to K do
    Roll out episode k in E to obtain {o_t}_{t=1}^T
    for t = 1 to T do
        Append (o_t, C(o_t; Ξ)) to D_obs
    end for
end for
for each sampled pair (o_a, o_b) ~ p_pair(D_obs) do
    Append (C(o_a), C(o_b)) to P
end for
return D_obs, P
```

With a fixed prompt template $\mathfrak{P}$, the LLM outputs $y \in \{1, 2, \varnothing\}$ for (c_1, c_2). We keep decisive labels:

$$\mathcal{D}_{\text{pref}} \;=\; \big\{(c_1, c_2, y) \;:\; y \in \{1, 2\}\big\}. \tag{6}$$

To reduce template-frequency bias, we assign importance $w(c_1, c_2) \propto 1/\nu\big(\text{tpl}(c_1, c_2)\big)$, where ν counts occurrences of the rendering template.

4.3. Stage 2: Learning a Common-Sense Intrinsic Reward

We learn a scalar quality score $r_\phi : \mathcal{O} \to \mathbb{R}$ via an encoder f_ϕ over caption tokens and auxiliary numerics:

$$\begin{aligned} & r_\phi(o) = f_\phi\big(\text{tok}(c),\, x(o)\big), \\ & x(o) \subseteq \{\, p_i,\, \bar{d},\, \mathrm{TTC}_{\mathrm{p10}},\, \mathrm{TTC}_{\mathrm{p50}},\, h_{\text{brake}} \,\}. \end{aligned} \tag{7}$$

While our architecture supports multiple prompt-specific heads combined linearly, in our experiments we use a single scalar head for r_ϕ.

Bradley–Terry likelihood and loss. Let $\Delta r_\phi = r_\phi(o_1) - r_\phi(o_2)$ for $(c_1, c_2, y) \in \mathcal{D}_{\text{pref}}$ with source observations o_1, o_2. With sigmoid $\sigma(z) = 1/(1 + e^{-z})$ and temperature $\tau_{\text{BT}} > 0$, we define

$$\begin{aligned} & p_\phi(y{=}1 \mid c_1, c_2) = \sigma\left(\frac{\Delta r_\phi}{\tau_{\text{BT}}}\right), \\ & p_\phi(y{=}2 \mid c_1, c_2) = 1 - p_\phi(y{=}1 \mid c_1, c_2). \end{aligned} \tag{8}$$

The weighted negative log-likelihood with ℓ_2 regularization and score centering is

$$\begin{aligned} \mathcal{L}(\phi) = & -\sum_{(c_1, c_2, y) \in \mathcal{D}_{\text{pref}}} w(c_1, c_2)\, \log p_\phi(y \mid c_1, c_2) \\ & +\; \eta\, \|\phi\|_2^2 \;+\; \zeta\, \big(\mathbb{E}[r_\phi(o)]\big)^2, \end{aligned} \tag{9}$$

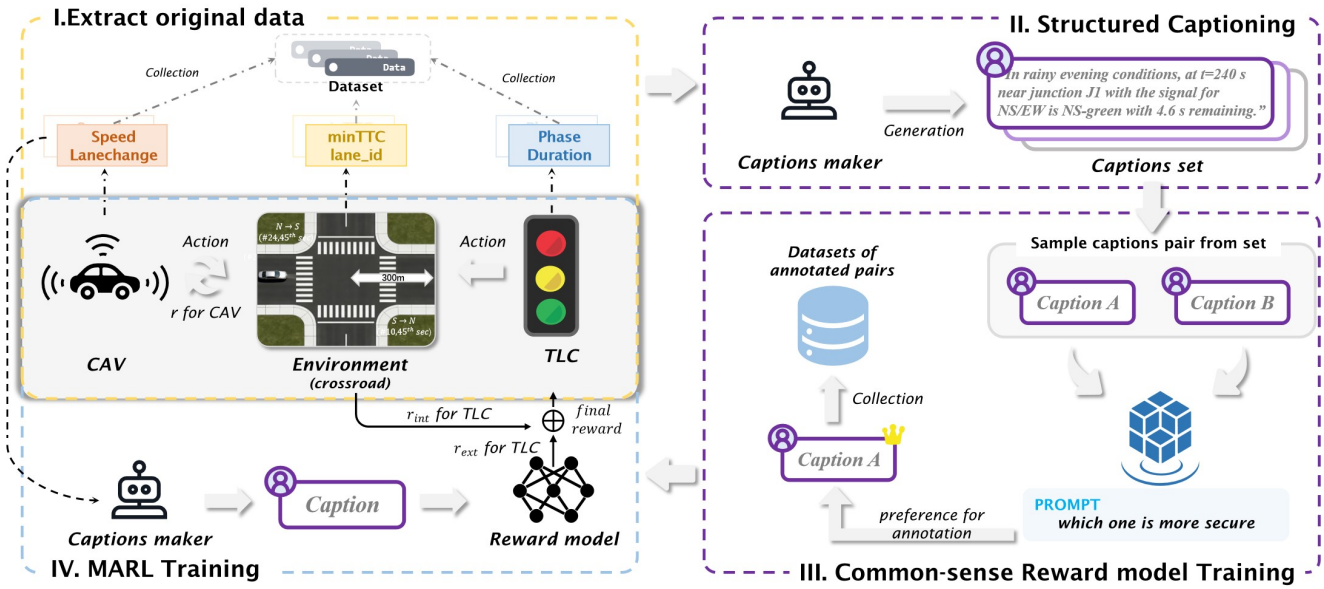

Figure 1. C^{2}T pipeline. Stage 1 converts raw observations into schema-constrained captions and samples high-contrast/safety-contrast pairs. Stage 2 queries an LLM to label pairwise preferences and trains a scalar scorer r_ϕ with a Bradley–Terry loss (with frequency reweighting and score centering). Stage 3 freezes r_ϕ and injects it as an intrinsic component for TLC steps under a safety mask, mixing with external rewards; all streams are normalized before GAE+PPO. In our CityFlow instantiation, only TLCs are learning agents and vehicles remain part of the environment dynamics.

Algorithm 2 Stage 2: Offline Reward Model Training

Require: LLM $\mathcal{L}$, prompt $\mathfrak{P}$, observation buffer $\mathcal{D}_{\text{obs}}$, pair budget M, epochs N
Ensure: Reward model r_ϕ

```
D_pref ← ∅
for j = 1 to M do
    Sample (o_1, o_2) ~ p_pair(D_obs)
    y ← L(P, C(o_1), C(o_2))
    if y ∈ {1, 2} then
        Append (o_1, o_2, y) to D_pref
    end if
end for
Initialize r_φ
for epoch = 1 to N do
    Sample minibatch from D_pref and update φ via (9)
end for
return r_φ
```

where $p_\phi(y \mid c_1, c_2)$ is given by (8). The centering term discourages degenerate shifts where r_ϕ drifts uniformly without improving pairwise discrimination.

4.4. Stage 3: Reward Shaping and RL Integration

After Stage 2, we freeze the learned scorer r_ϕ and integrate it into a multi-intersection TLC RL system on CityFlow, where each intersection-level TLC is modeled as an agent and vehicles follow the built-in microscopic simulator.

Safety mask and intrinsic shaping. Define a mask $m : \mathcal{O} \to \{0, 1\}$ that disables intrinsic rewards in unsafe regions:

$$m(o') = \mathbb{I}\{\, \text{TTC}_{\text{p10}}(o') \geq \tau_{\text{ttc}},\ |a_{\text{near}}(o')| \leq a_{\max},\ \rho_{\text{red}}(o') = 0 \,\}. \tag{10}$$

Given a transition (o, a, o') for a TLC, the intrinsic component is

$$r_{\text{int}}(o, o') \;=\; m(o')\, r_\phi(o'). \tag{11}$$

The mixed reward for a TLC agent is

$$r^{\text{TL}}_{\text{mix}} = r^{\text{TL}}_{\text{ext}} + \lambda(t)\, m(o')\, r_\phi(o') - \kappa_{\text{unsafe}}(1 - m(o'))\,, \tag{12}$$

where $\lambda(t) \in [0, \lambda_{\max}]$ is a slowly varying schedule (warm-up/decay or performance-gated) that gradually increases the weight of the intrinsic shaping signal.

Per-stream normalization and clipping. We treat the external reward, the intrinsic component, and the safety penalty as three separate streams for TLCs:

$$r_t^{(1)} = r^{\text{TL}}_{\text{ext},t}, \quad r_t^{(2)} = \lambda(t)\, r_{\text{int},t}, \quad r_t^{(3)} = -\kappa_{\text{unsafe}}(1 - m_t).$$

For each stream $k \in \{1, 2, 3\}$ we maintain running mean $\mu^{(k)}$ and standard deviation $\sigma^{(k)}$, and apply z-score normal-

Algorithm 3 Stage 3: Online Training with C^{2}T Intrinsic Reward (TLC-only)

Require: Simulator $\mathcal{E}$, trained scorer r_ϕ, initial TLC policy π_{TL}, iterations I, episodes E, horizon H
Ensure: Updated TLC policy π_{TL}

```
for iter = 1 to I do
    B ← ∅
    for episode = 1 to E do
        Reset E and observe initial states for all TLCs
        for h = 0 to H − 1 do
            for each TLC agent i do
                Sample action a_{h,i} ~ π_TL(· | o_{h,i})
            end for
            Step E with joint TLC actions, obtain o_{h+1,i}
            and external rewards r^TL_{ext,h,i}
            for each TLC agent i do
                Compute mask m_{h+1,i} ← m(o_{h+1,i})
                using (10)
                r^int_{h,i} ← m_{h+1,i} r_φ(o_{h+1,i})
                r_{h,i} ← r^TL_mix via (12)
                Store (o_{h,i}, a_{h,i}, r_{h,i}, o_{h+1,i}) into B
            end for
        end for
    end for
    Normalize and clip streams from B using (13) to
    obtain shaped rewards r̂^TL_t
    Compute advantages A_t from B with GAE
    Update π_TL and V_ψ with PPO
end for
return π_TL
```

ization with soft clipping:

$$\hat{r}_t^{(k)} = \text{clip}\left(\frac{r_t^{(k)} - \mu^{(k)}}{\sigma^{(k)} + \epsilon}, -c, c\right). \tag{13}$$

We then form a single step-wise shaped reward for TLCs by summing the normalized streams:

$$\hat{r}_t^{\text{TL}} = \sum_{k=1}^{3} \hat{r}_t^{(k)}.$$

Advantage estimation and policy optimization. Let $V_\psi(o)$ be the value function for TLCs and π_θ the shared TLC policy. We estimate advantages from $\hat{r}_t^{\text{TL}}$ using generalized advantage estimation (GAE) and optimize (θ, ψ) with the standard clipped-surrogate PPO objective [10, 11]. For completeness, we summarize one training iteration in Algorithm 3; implementation details and hyper-parameters follow the underlying CityFlow PPO setup and are deferred to the appendix.

4.5. Compatibility and Reproducibility

Unless otherwise noted, we keep the base RL setup (agent sets, shared TLC policy design, observation/action spaces, and PPO schedules) from the underlying CityFlow configuration. C^{2}T adds: (i) the captioner $\mathcal{C}$ with schema Ξ, (ii) the learned scorer r_ϕ via (8)–(9), and (iii) the mixed reward (12) with normalization (13) and safety mask (10). While our experiments only train TLC policies, the same shaping mechanism can be applied to cooperative TLC–CAV settings by applying the intrinsic component to TLC rewards and keeping vehicle agents either environment-driven or separately optimized; we discuss this extension in the appendix.

5. Experiments

5.1. Benchmarks and Simulation Settings

Datasets. Following the LLMLight protocol [7], we evaluate on **CityFlow** [24] using three real-city networks and two synthetic stress scenarios: **Jinan** (12 intersections; three time periods), **Hangzhou** (16 intersections; two periods), **New York** (196 intersections; two periods), **Extreme High-traffic** (arrival rate $\times 4$ within 5 minutes), and a **24-hour Cycle** (full diurnal fluctuation). Unless specified otherwise, each scenario runs for 1 hour.

Environment. Each intersection has four phases (E/W straight, E/W left-turn, N/S straight, N/S left-turn). Signal timing follows: **30s green → 3s yellow → 2s all-red**. Agents act only at phase-switches for fair comparison. Right-turns are permissive; lane-changing and car-following follow CityFlow defaults.

Metrics. We follow standard traffic-engineering metrics: **ATT** (Average Travel Time, ↓), **AQL** (Average Queue Length, ↓), **AWT** (Average Waiting Time, ↓). Safety/comfort metrics include **rear-end TTC percentiles** (P10/P25, ↑) and **Harsh-brake/km** (↓). Energy-related proxies (stops, speed variance) are reported in the appendix.

5.2. Compared Methods

Transportation baselines: Random, FixedTime, MaxPressure.

RL baselines: MPLight [1], AttendLight [8], PressLight [19], CoLight [20], Efficient-CoLight [21], Advanced-CoLight [25] and its *no-communication* variant.

LLM agents: GPT-, Qwen-, and Llama-family variants reported by LLMLight [7].

All RL baselines share LLMLight's training budget and hyperparameters.

Our method (C^{2}T). We transform simulator states into structured, unit-aware captions, collect offline pairwise

Table 1. **Jinan/Hangzhou overall performance.** Lower is better.

Dataset/Method	ATT↓	AQL↓	AWT↓
Jinan-1 (Baseline)	62.30	14.80	21.50
Jinan-1 (C^2T)	**56.10**	**13.10**	**19.40**
Jinan-2 (Baseline)	58.90	13.10	19.70
Jinan-2 (C^2T)	**53.00**	**11.80**	**17.90**
Hangzhou-1 (Baseline)	71.40	16.30	24.20
Hangzhou-1 (C^2T)	**65.10**	**14.90**	**22.00**
Hangzhou-2 (Baseline)	68.10	15.40	23.10
Hangzhou-2 (C^2T)	**62.40**	**13.90**	**21.10**

Table 2. **Large-scale New York (196 intersections).**

Method	ATT↓	AQL↓	AWT↓
Adv-CoLight (baseline)	95.0	22.6	38.4
C^2T (ours)	**87.9**	**20.3**	**34.5**

preferences, and train an intrinsic scorer r_ϕ. The intrinsic reward is mixed *asymmetrically* into TL controllers (TLCs) only; CAVs (if present) remain environment-driven. A risk mask suppresses r_ϕ when TTC percentiles drop below thresholds or harsh deceleration clusters appear. Reward streams are normalized and softly clipped prior to PPO advantage computation.

5.3. Implementation Details

RL. We adopt LLMLight's PPO settings [7]: learning rate 1×10^{-3}, replay buffer 12,000, sample size 3,000, hidden size 20, PPO clip 0.2, batch size 128, 16 mini-batches, entropy coefficient 10^{-3}, GAE($\lambda = 0.95$), $\gamma = 0.99$. Results are averaged over 5 random seeds.

Preference labeling. We use temperature 0 for proprietary LLMs and 0.1 for open-source models, with top-p=1.0. When fine-tuning open-source judges, we apply LoRA with rank 8 and $\alpha = 16$; all preference labels are collected offline, and RL training/inference require no LLM calls.

Safety proxies. TTC uses per-vehicle rear-end time-to-collision; harsh braking is $a < -3\,\mathrm{m/s^2}$. For stress tests we also track an *oscillation* index measuring normalized phase-switch frequency (higher means more frequent switching).

5.4. Main Results (RQ1: Effectiveness)

Safety side metrics. Across Jinan/Hangzhou, C^2T consistently improves TTC P10/P25 and reduces harsh braking, indicating smoother approaches and fewer stop–go cycles (see Appendix for full numbers).

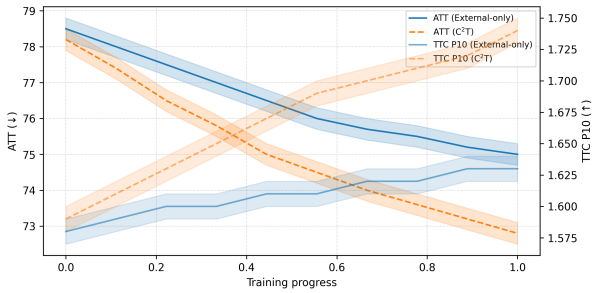

Figure 2. **Learning curves on CityFlow.** C^2T steadily improves ATT (↓) while also increasing TTC P10 (↑), showing that efficiency does not come at the cost of safety.

Table 3. **City transfer: train on Jinan, evaluate on Hangzhou.**

Method	ATT↓	AQL↓	AWT↓	Steps↓
External-only	72.5	17.2	25.8	100k
C^2T zero-shot	68.9	16.4	24.6	80k
C^2T (1% prefs)	67.3	16.0	23.9	70k
Fusion (50/50)	**66.1**	**15.6**	**23.5**	**65k**

Table 4. **Stress test results.** Lower is better unless noted.

Setting / Method	ATT↓	AWT↓	Oscillation↓
Extreme (baseline)	105.0	44.2	1.00
C^2T	**96.8**	**40.1**	**0.86**
24h (baseline)	78.5	29.4	0.76
C^2T	**72.2**	**26.9**	**0.65**

5.5. Generalization and Scalability (RQ2)

City transfer. We train a source-city scorer r_ϕ^s on Jinan and evaluate zero-shot on Hangzhou. Few-shot preference relabeling (1–5%) further improves ATT/AWT. We also study a simple fusion of source- and target-city scorers, $r_\phi = \lambda r_\phi^s + (1-\lambda) r_\phi^t$, which accelerates convergence.

Scalability. Training on small networks (Jinan/Hangzhou) generalizes to New York (196 intersections), while several RL baselines degrade sharply; see the appendix for the full comparison.

5.6. Stress Tests (RQ3)

Extreme High-traffic. Arrival rates increase by $\times4$ over 5 minutes. We evaluate saturation robustness and the *oscillation* index (normalized phase-switch frequency).

24-hour Cycle. We replay a full-day traffic profile, tracking stability and frequency of risk-mask activation.

5.7. Ablation Studies (RQ4)

We conduct two groups of ablations to understand which components of C^2T are responsible for the gains; see Fig. 3 and Tab. 5 for a summary, and the appendix for full results and additional plots.

Component-wise ablations. Figure 3 removes, one at a time, the intrinsic reward r_ϕ, the safety mask, per-stream normalization, and the intrinsic-reward schedule. Dropping r_ϕ almost collapses performance back to the external-only PPO baseline, confirming that the learned reward signal, rather than extra computation, is the main source of improvement. Removing the safety mask slightly reduces ATT/AWT but noticeably worsens rear-end TTC percentiles and harsh-brake frequency, indicating that the mask is crucial for turning the intrinsic reward into *safe* guidance. Without per-stream normalization the mixed reward becomes highly skewed and training becomes less stable, leading to higher variance and degraded ATT. Finally, using a constant intrinsic weight in place of the curriculum schedule slows down convergence and yields weaker final performance, suggesting that gradually introducing r_ϕ is beneficial for optimization.

Caption-structure ablations. Table 5 studies the effect of our structured caption design. Using only raw numerics ("Numeric-only") or only text ("Caption-only") gives weaker offline agreement with LLM preferences (AUC/Spearman) and worse downstream ATT/AWT than the full fusion model. Shuffling slots or removing explicit units ("Shuffled / No-units") hurts both ranking metrics and RL performance, showing that schema consistency and unit-awareness matter. The full C^2T model, which combines numerics with ordered, unit-aware captions, achieves the best trade-off between preference prediction and control.

Interaction and robustness. We also explore how these components interact. For example, combining "no mask" with "no normalization" leads to occasional training instability and oscillatory phase-switch behaviour, even when ATT is comparable, whereas removing either component alone is less harmful. Conversely, keeping the mask and schedule but training a weaker reward model (Numeric-only head) still improves safety but yields smaller efficiency gains. Across all settings we observe a consistent trend: better offline preference alignment (higher AUC/Spearman) correlates with better RL performance, supporting our design choice of evaluating reward models with both offline metrics and online control. Extended sensitivity curves and cross-city ablations are provided in the appendix.

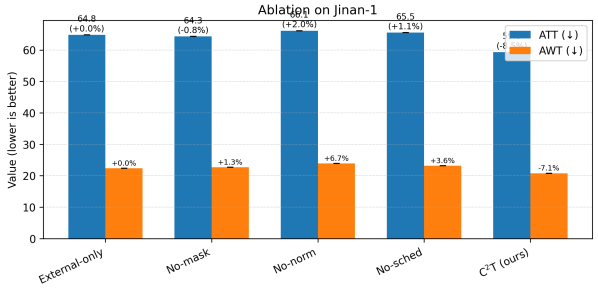

Figure 3. **Ablation on Jinan-1.** Removing mask/norm/schedule harms performance, while the full C^2T achieves the best ATT/AWT.

Table 5. **Caption-structure ablation**. Higher AUC/Spearman is better; lower ATT/AWT is better.

Model	AUC↑	Spearman↑	ATT↓	AWT↓
Numeric-only	0.67	0.41	64.8	22.4
Caption-only	0.73	0.49	63.1	22.1
Fusion	0.78	0.56	61.9	21.6
Shuffled / No-units	0.60	0.30	66.1	23.9
C^2T (full)	**0.78**	**0.56**	**59.3**	**20.8**

6. Conclusion

We introduced C^2T, a framework that leverages structured traffic captions and LLM-based preference learning to build a human-aligned intrinsic reward for multi-intersection traffic signal control. By rendering simulator states into deterministic, unit-aware captions and training a lightweight scorer from offline LLM preferences, C^2T decouples reward design from online interaction and avoids any runtime dependence on LLM calls. The learned intrinsic signal is integrated into a standard PPO pipeline via asymmetric reward mixing, a simple safety mask, and per-stream normalization, making it drop-in compatible with existing RL controllers.

On CityFlow-based benchmarks built from real-world Jinan, Hangzhou, and New York networks, C^2T consistently improves traffic efficiency, safety proxies, and an energy-related metric over strong RL baselines. Ablation studies highlight the importance of the intrinsic reward, the safety mask, and the normalization schedule for stable gains, and show that the structured caption design itself already provides measurable benefits.

Looking forward, we plan to extend C^2T to explicit TLC–CAV coordination, explore richer prompt designs and multi-objective heads for finer-grained trade-offs between efficiency, safety, and energy, and investigate deployment in real-world sensing and control loops where human feedback and LLM judgment can be iteratively refined.

References

[1] 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*, 2020. 2, 6

[2] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, 2017. 1, 2

[3] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. In *Proceedings of the International Conference on Learning Representations*, 2018. 2

[4] Jiaying Guo, Long Cheng, and Shen Wang. CoTV: Cooperative control for traffic light signals and connected autonomous vehicles using deep reinforcement learning. *arXiv preprint arXiv:2201.13143*, 2023. 1, 2

[5] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, 2016. 2

[6] Yuichi Inoue, Yuki Yada, Kotaro Tanahashi, and Yu Yamaguchi. NuScenes-MQA: Integrated evaluation of captions and QA for autonomous driving datasets using markup annotations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 586–595, 2024. 3

[7] Siqi Lai, Zheyuan Xu, Wenjun Zhang, Haoran Liu, and Hui Xiong. LLMLight: Large language models as traffic signal control agents. *arXiv preprint arXiv:2312.16044*, 2023. 3, 6, 7

[8] Mehdi Oroojlooy, Lawrence V. Snyder, Reza Samadi, and Bo Zeng. Attendlight: Universal attention-based reinforcement learning for traffic signal control. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2020. 2, 6

[9] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. 1, 3

[10] John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations (ICLR)*, 2016. 6

[11] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 6

[12] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beisswenger, Ping Luo, Andreas Geiger, and Hongyang Li. DriveLM: Driving with graph visual question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 3

[13] Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M. Alvarez. OmniDrive: A holistic vision–language dataset for autonomous driving with counterfactual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[14] Yiming Wang, Kaiyan Zhao, Yan Li, and Leong Hou U. BILE: An effective behavior-based latent exploration scheme for deep reinforcement learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 6497–6505, 2025. 2

[15] Yiming Wang, Kaiyan Zhao, Xu Li, Yan Li, Jiayu Chen, Steven Morad, and Leong Hou U. Explore to learn: Latent exploration through disentangled synergy patterns for reinforcement learning in overactuated control. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 26670–26678, 2026. 2

[16] Yiming Wang, Kaiyan Zhao, Ming Yang, Yan Li, Furui Liu, Jiayu Chen, and Leong Hou U. DSAP: Enhancing generalization in goal-conditioned reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 26679–26687, 2026. 2

[17] Yiming Wang, Kaiyan Zhao, Borong Zhang, Yan Li, and Leong Hou U. Latent state-predictive exploration for deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 26661–26669, 2026. 2

[18] Hua Wei, Guanjie Zheng, Huaxiu Yao, and Zhenhui Li. IntelliLight: A reinforcement learning approach for intelligent traffic light control. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018. 1, 2

[19] Hua Wei, Chacha Chen, Guanjie Zheng, Kan Wu, Vikash Gayah, Kai Xu, and Zhenhui 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 and Data Mining*, 2019. 1, 2, 6

[20] 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. 2, 3, 6

[21] Kan Wu, Hua Wei, Chacha Chen, Guanjie Zheng, and Zhenhui Li. Efficient-CoLight: Learning efficient pressure representation for network-level traffic signal control. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pages 2060–2069, 2021. 2, 6

[22] Zhenghao Xu, Qin Lu, Qingru Zhang, Liang Qiu, Ilgee Hong, Changlong Yu, Wenlin Yao, Yao Liu, Haoming Jiang, Lihong Li, Hyokun Yun, and Tuo Zhao. Ask a strong LLM judge when your reward model is uncertain. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 3

[23] 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*, 2020. 2

[24] Huichu Zhang, Siyuan Feng, Chang Liu, Yaoyao Ding, Yichen Zhu, Zihan Zhou, Weinan Zhang, Yong Yu, Haiming Jin, and Zhenhui Li. Cityflow: A multi-agent reinforcement learning environment for large scale city traffic scenario. In *Proceedings of the World Wide Web Conference*, pages 3620–3624. ACM, 2019. 2, 6

[25] Liang Zhang, Qiang Wu, Jun Shen, Linyuan Lü, Bo Du, and Jianqing Wu. Expression might be enough: Representing pressure and demand for reinforcement learning-based traffic signal control. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pages 26645–26654. PMLR, 2022. 2, 6

[26] Kaiyan Zhao, Yiming Wang, Yuyang Chen, Yan Li, Leong Hou U, and Xiaoguang Niu. Efficient diversity-based experience replay for deep reinforcement learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 7083–7091, 2025. 2

[27] 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 with FRAP. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019. 2, 3

Acknowledgments

This work was supported by the Science and Technology Development Fund of Macau [0007/2025/RIC, 0122/2024/RIB2, 0215/2024/AGJ, 0074/2025/AMJ, 001/2024/SKL], the Research Services and Knowledge Transfer Office, University of Macau [SRG2023-00037-IOTSC, MYRG-GRG2024-00284-IOTSC], the Shenzhen-Hong Kong-Macau Science and Technology Program Category C [SGDX20230821095159012], the Science and Technology Planning Project of Guangdong [2025A0505010016], the National Natural Science Foundation of China [52572354], the State Key Lab of Intelligent Transportation System [2024-B001], and the Jiangsu Provincial Science and Technology Program [BZ2024055].