CoopLight: Cooperative Vision-Language Agents for Long-Tail Traffic Control

Anonymous authors
Paper under double-blind review

Abstract

Traffic light signals and connected and automated vehicles (CAVs) are increasingly interdependent at urban intersections, yet they are still predominantly studied in isolation. Existing methods mainly rely on low-dimensional traffic states, limiting their ability to handle long-tail scenarios such as incidents, jaywalking pedestrians, and emergency-priority events. Such settings require semantic scene understanding, context-aware decision making, and timely cooperation between infrastructure-side and vehicle-side agents. To address these challenges, we introduce **CoopLight**, a vision-language framework for traffic signal and CAV control. Agent-specific visual question answering (VQA) tasks ground spatial scene understanding and decision making in multi-view observations. Multi-agent cotraining further enhances agents' ability to integrate local observations with received semantic reports. This learned cooperative reasoning enables signal and vehicle agents to use complementary evidence and adapt their decisions under partial observability. Extensive experiments on two real signalized intersections under multiple long-tail traffic scenarios show that CoopLight consistently improves traffic efficiency over rule-based, reinforcement-learning-based, and prior vision-language-model-driven baselines, demonstrating its effectiveness and proactivity for next-generation urban mobility.

1 Introduction

Urban traffic congestion remains a persistent challenge for modern cities, causing substantial economic loss, environmental burden, and reduced travel efficiency (Fattah et al., 2022; Xu et al., 2024a). As urban road networks become denser and more complex, driven by the rise of autonomous vehicles and communication infrastructure, mobility at signalized intersections is increasingly shaped by two tightly coupled forms of control: infrastructure-side traffic signal control (TSC), which allocates right-of-way and regulates junction-level traffic flow, and vehicle-side decision making, where connected and automated vehicles (CAVs) adjust their maneuvers in response to local traffic conditions and signal status. However, effective control at real intersections goes far beyond routine traffic regulation. It must also handle long-tail situations, such as traffic incidents, road blockage, jaywalking pedestrians, and approaching emergency vehicles, where delayed or incorrect signal responses and inefficient vehicle maneuvers can quickly propagate into severe congestion and traffic disruption.

Traditional transportation-based and reinforcement-learning-based TSC methods have shown strong performance under routine traffic, but they typically rely on compact numerical states such as queue length, occupancy, or pressure, which discard the semantic cues needed to distinguish ordinary congestion from anomalies (Eom & Kim, 2020; Varaiya, 2013; Wei et al., 2019). Meanwhile, vehicle-side methods mainly focus on local trajectory planning under normal driving conditions, but lack sufficient awareness of long-tail traffic events at intersections (Xia & Chen, 2024; Liu & Feng, 2024; Jain et al., 2021). For example, high occupancy may indicate heavy inbound traffic under normal conditions, but it may also reflect vehicles physically trapped or obstructed behind a crash or roadside barrier. Without access to richer scene understanding, state-based controllers often respond sub-optimally, leading to ineffective signal decisions and delayed vehicle reactions that further worsen congestion and safety risks. Prior joint signal-vehicle control frameworks are often formulated through deterministic optimization or hierarchical coordination, which can be difficult

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

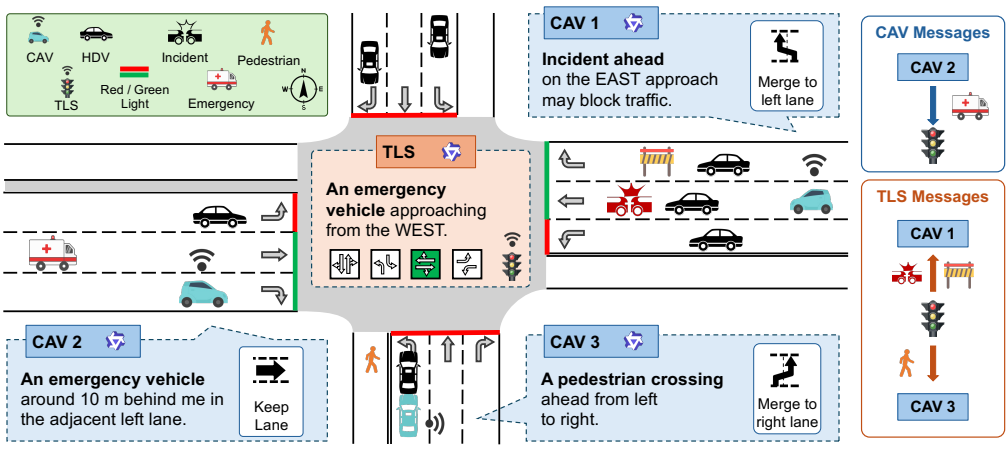


Figure 1: Heterogeneous TLS↔CAV cooperation under long-tail traffic scenarios. The TLS agent observes the intersection from multi-directional cameras while CAV agents observe from ego-centric views. Each agent first performs confidence-aware perception, then exchanges semantic reports so that signal phasing and vehicle maneuvers are jointly grounded on the most reliable evidence.

to scale or adapt under rapidly changing long-tail and complex traffic conditions (Guo et al., 2019; Wu & Qu, 2022; Zheng et al., 2024; Dabiri et al., 2025).

Recent progress in large language models (LLMs) has opened new opportunities for more interpretable traffic control. Compared with traditional TSC methods, LLM-based approaches offer stronger reasoning ability, and prior studies (Lai et al., 2025; Yuan et al., 2026) have shown that language-driven reasoning can support more transparent signal decisions. Vision-language models (VLMs), pre-trained on internet-scale multimodal corpora, are particularly promising in this regard, as their broad world knowledge and commonsense reasoning capabilities make them naturally suited to detecting and classifying long-tail traffic events from roadside and onboard camera observations (Tian et al., 2025; Zhou et al., 2024). Most recently, VLMLight (Wang et al., 2025b) demonstrated the effectiveness of VLM-based reasoning for emergency-aware traffic signal control.

However, these challenges are especially pronounced when infrastructure-side signals and vehicle-side CAVs must reason over the same scene from different viewpoints and levels of abstraction. The infrastructure side has a broad but coarse junction-level view, while each CAV has a detailed but local ego-centric view. How to leverage VLMs for scene understanding while enabling efficient semantic sharing for cooperative decision making remains an important open challenge. To address this, we propose a VLM-grounded heterogeneous agent framework for cooperative traffic signal and CAV control under long-tail traffic scenarios. The infrastructure-side agent reasons over multi-directional images, lane layout, and detector measurements to produce confidence-aware, spatial scene understanding of junction and context-aware signal decisions. In parallel, each CAV agent reasons over ego-centric surrounding-view images to identify lane-level anomalies and make maneuvers that maintain travel efficiency. Both agent types exchange semantic reports containing event types, locations, and confidence levels. Through multi-agent co-training, senders learn to produce accurate reports, while receivers learn to condition their rationales and actions on local and shared evidence, enabling cooperative reasoning and context-aware decision making.

The main contributions are as follows:

- We develop a VLM-grounded cooperative control framework that unifies infrastructure-side traffic signal control and vehicle-side CAV decision making in a joint reasoning and control architecture.
- We construct long-tail traffic scenarios in an image-based simulator and introduce perception and decision-making VQA tasks for both agents at two real intersections. Fine-tuning on these tasks equips agents with confidence-aware spatial understanding and context-aware decision making.
- We propose multi-agent co-training that jointly rewards sender report accuracy and receiver action correctness, supporting proactive decision making and adaptive cooperation. Closed-loop

experiments demonstrate improvements in overall and CAV travel efficiency under long-tail scenarios, measured by average waiting time and average travel time.

2 Related Work

Traffic Signal Control. Traditional traffic signal control methods (Webster, 1958; Varaiya, 2013) rely on handcrafted rules and struggle under dynamic traffic conditions. RL-based controllers (Wei et al., 2018; 2019; Liang et al., 2019; Aslani et al., 2017) improve adaptability, but their reliance on compact numerical states limits their ability to reason about long-tail events in real intersections. Recent LLM-based controllers (Lai et al., 2025; Zou et al., 2025; Yuan et al., 2026) enhance interpretability but require intensive multi-step reasoning on structured textual descriptions. Meanwhile, V2X-enabled frameworks (Mo et al., 2022; Wang et al., 2024; Zhang et al., 2025) extend signal-side observability with vehicle-side telemetry for adaptive signal control. CoTV (Guo et al., 2023) further explores joint signal and CAV coordination to improve overall traffic efficiency. Despite these advances, existing methods remain largely dominated by low-dimensional traffic states, limiting their ability to exploit rich visual cues and perform joint multi-modal reasoning, and thus restricting robustness in complex urban intersection scenarios.

VLM-Based Driving and Signal Control. Recent VLMs have shown strong capabilities in visual understanding and reasoning, enabling more interpretable scene analysis and decision making (Liu et al., 2023; Bai et al., 2025; Zhang et al., 2024). In autonomous driving, prior works leverage VLMs for scene understanding, high-level driving commands, and waypoint planning (Xu et al., 2024b; Tian et al., 2025; Mao et al., 2024). VLMLight (Wang et al., 2025b) introduces VLM-based visual scene understanding for safety-critical signal control with an RL-hybrid pipeline. However, its modular design remains focused on local signal control and is vulnerable to cumulative error propagation. To the best of our knowledge, CoopLight is the first to address cooperative traffic signal and CAV control in long-tail urban intersection scenarios through VLM-grounded reasoning, enabling TLS and CAV agents to exchange semantic information and jointly improve decision making.

3 Problem Formulation

Our goal is to reduce traffic delay at complex urban signalized intersections through joint control of traffic light signals and the CAVs that traverse them, measured by the average travel time and waiting time of vehicles at the intersection.

Long-Tail Scenarios. We focus on four types of long-tail scenarios that can severely degrade intersection efficiency and control effectiveness: incidents, blockages, jaywalking pedestrians, and emergency vehicles. (a) *Incident.* A stalled or crashed vehicle blocks one or more lanes, causing abrupt capacity loss and dynamic queue buildup behind the incident. Such cases are challenging because persistent detector occupancy may resemble ordinary congestion even though part of the queue is unable to clear. (b) *Blockage.* Barriers, such as construction cones, lane closures, or blocked road segments, obstruct one or more lanes on selected approaches. These obstacles partially reduce lane availability, requiring both adaptive lane changes and adjusted phase allocation (c) *Jaywalking Pedestrians.* Pedestrians unexpectedly cross the roadway during normal signalized movements, creating sudden conflicts that require timely driving adjustments while maintaining traffic efficiency. (d) *Emergency Vehicles.* Emergency vehicles, such as ambulances, fire trucks, or police cars, enter from random approaches and require timely right-of-way.

Heterogeneous-Agent Framework. We consider a signalized intersection controlled by a traffic signal agent, $\mathcal{A}^{\text{TLS}}$, together with a set of CAV agents, $\mathcal{A}^{\text{CAV}} = \{c_i\}$, where each agent c_i controls a CAV traversing the intersection. These two classes of autonomous agents operate with heterogeneous observations and action spaces. Specifically, $\mathcal{A}^{\text{TLS}}$ acts over a set of K feasible phases, $\mathcal{P} = \{p_1, \ldots, p_K\}$. Each phase p_k activates a subset of non-conflicting traffic movements, where each movement connects an incoming lane to an outgoing lane. At each signal decision step t_s, $\mathcal{A}^{\text{TLS}}$ receives four directional camera views $\mathbf{I}_{t_s}^{\text{TLS}}$, per-movement vehicle counts $\mathbf{n}_{t_s}$, and a system prompt specifying the intersection-specific phase-to-movement mapping derived from the signal program, and selects a phase$a_{t_s}^{\text{TLS}} \in \mathcal{P}$ to minimize traffic delay. Each CAV agent $c_i \in \mathcal{A}^{\text{CAV}}$ is equipped with six surround-view cameras. At each vehicle decision step t_v, c_i observes its surround-view images $\mathbf{I}_{i,t_v}^{\text{CAV}}$, and ego state, $\mathbf{x}_{i,t_v}$, which includes its current speed and lane, and outputs a lateral maneuver

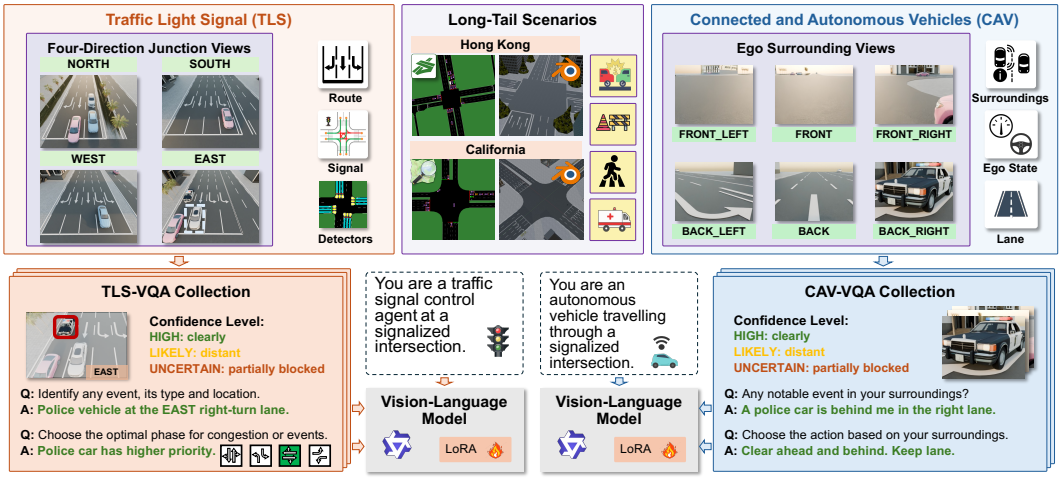


Figure 2: **VQA generation and supervised fine-tuning pipeline.** Four long-tail scenario types are constructed for route generation and visual and state data collection. We curate TLS-VQA and CAV-VQA records for two task classes: confidence-aware perception and perception-grounded decision making, where each record is annotated with a confidence label $\gamma \in \{\text{HIGH}, \text{LIKELY}, \text{UNCERTAIN}\}$. A shared VLM backbone is then fine-tuned with agent-specific objectives for $\mathcal{A}^{\text{TLS}}$ and $\mathcal{A}^{\text{CAV}}$.

$a^{\text{CAV}}_{i,t_v} \in \mathcal{U}^{\text{CAV}}$ to support efficient progress. Both agents share a common VLM backbone, f_θ, parameterized by θ, thereby benefiting from the semantic knowledge acquired during pre-training. They are further optimized under agent-specific objectives for perception and decision making.

4 METHOD

To connect language-based reasoning with control, we formulate two classes of visual question answering (VQA) tasks for semantic scene understanding and perception-grounded decision making, as illustrated in Figure 2. First, *confidence-aware spatial scene understanding* enables both $\mathcal{A}^{\text{TLS}}$ and $\mathcal{A}^{\text{CAV}}$ to produce junction-level or ego-centric scene interpretations, together with varying confidence levels and supporting rationales. Second, *context-aware decision making* grounds each agents action in its local observations and inferred traffic conditions. Third, *cooperative belief reasoning* enables the two agent types to exchange semantic reports and learn through multi-agent co-training, thereby supporting more proactive and adaptive decisions across long-tail scenarios.

4.1 CONFIDENCE-AWARE SPATIAL SCENE UNDERSTANDING

To fully understand dynamic traffic scenarios, we design a perception-focused task for both $\mathcal{A}^{\text{TLS}}$ and $\mathcal{A}^{\text{CAV}}$. Each agent identifies, describes, and localizes the key objects in the current scene. Given multi-view images and detector measurements, $\mathcal{A}^{\text{TLS}}$ jointly reasons over visual semantics and measured traffic states to produce a structured junction-level description $s^{\text{TLS}}_{t_s} = f_\theta(\mathbf{I}^{\text{TLS}}_{t_s}, \mathbf{n}_{t_s})$. It identifies anomaly events, their types, and spatial locations, such as affected lanes or incoming approaches. For blockages caused by crashes or barriers, it also assesses queue severity behind the obstruction; in the absence of safety-critical events, it compares traffic conditions across approaches. Similarly, each CAV agent c_i uses surround-view images and ego state to generate a maneuver-relevant description $s^{\text{CAV}}_{i,t_v} = f_\theta(\mathbf{I}^{\text{CAV}}_{i,t_v}, \mathbf{x}_{i,t_v})$. This description identifies incidents, slowdowns, or other hazards from the ego perspective, including the types of critical surrounding objects and their relative lane offsets and distances. When no critical event is observed, it characterizes the local driving context, such as nearby lane clearance. However, visual evidence varies in reliability across viewpoints: an event recognizable in a nearby unobstructed view may be weakly visible at a distance, while occluded or out-of-view events may still affect traffic. Each agent therefore outputs a confidence level $\gamma \in \{\text{HIGH}, \text{LIKELY}, \text{UNCERTAIN}\}$ alongside its description. For the TLS, we distinguish four cases: (i) clearly visible ($\gamma = \text{HIGH}$), with reliable event localization and characterization; (ii) marginally visible ($\gamma = \text{LIKELY}$), due to distance or partial observability; (iii) suspicious

($\gamma = \text{UNCERTAIN}$), with no directly visible anomaly but anomalously unbalanced flows across lanes within the same movement; and (iv) normal ($\gamma = \text{HIGH}$), supported by consistent visual and detector evidence. For each CAV agent, confidence is HIGH when a hazard is directly visible ahead, or the nearby road region is clearly observable and free of anomalies. A visible but distant event receives LIKELY, whereas a partially observable event receives UNCERTAIN. The resulting local beliefs are $b_{t_s}^{\text{TLS}} = (s_{t_s}^{\text{TLS}}, \gamma_{t_s}^{\text{TLS}})$ and $b_{i,t_v}^{\text{CAV}} = (s_{i,t_v}^{\text{CAV}}, \gamma_{i,t_v}^{\text{CAV}})$, where $\gamma_{t_s}^{\text{TLS}}$ and $\gamma_{i,t_v}^{\text{CAV}}$ denote the confidence levels within each agent, capturing both scene semantics and observational reliability.

4.2 Context-Aware Decision Making

Building on the perception task, we design a decision task in which each agent generates an interpretable action grounded in local observations and inferred event semantics, distinguishing routine demand from anomaly-affected conditions. Target expert actions are supplied by privileged rule-based experts using ground-truth event types and locations, per-lane dischargeable vehicle counts, and inter-vehicle gaps. These privileged states are unavailable to the agents, which learn to imitate expert decisions from their own visual and state inputs.

Using the same observations as in perception, $\mathcal{A}^{\text{TLS}}$ predicts a signal phase and its rationale: $(r_{t_s}^{\text{TLS}}, a_{t_s}^{\text{TLS}}) = f_\theta(\mathbf{I}_{t_s}^{\text{TLS}}, \mathbf{n}_{t_s})$, with $a_{t_s}^{\text{TLS}} \in \mathcal{P}$. The rationale identifies whether a safety-critical event exists, the affected lane or movement when visible, and the demand and priority that justify the selected phase. Under routine traffic, the agent serves the phase with the highest demand. Under crashes or barriers, high occupancy may reflect physically blocked queues that cannot benefit from additional green time. The agent therefore evaluates the effective demand of each candidate phase by distinguishing blocked vehicles from traffic that can still move under green. In emergency scenarios, it prioritizes the corresponding movement to facilitate right-of-way. For marginally visible events, it uses tentative reasoning consistent with the perceptual confidence. Each CAV agent likewise predicts a lateral maneuver and its rationale from its own observations: $(r_{i,t_v}^{\text{CAV}}, a_{i,t_v}^{\text{CAV}}) = f_\theta(\mathbf{I}_{i,t_v}^{\text{CAV}}, \mathbf{x}_{i,t_v})$, with $a_{i,t_v}^{\text{CAV}} \in \mathcal{U}^{\text{CAV}}$. The rationale considers traffic ahead, adjacent-lane clearance, and detected anomalies, explaining whether the current lane remains safe and clear, a neighboring lane offers a better path around a blockage, or yielding is necessary for an approaching emergency vehicle.

We further equip both agents with a lightweight self-reflection step on the fly. For $\mathcal{A}^{\text{TLS}}$, the agent examines whether the previously selected phase actually improved traffic flow in the intended direction. If the observed outcome does not match the expected effect, the agent revises its prior belief when making the next phase decision. Similarly, each CAV agent checks whether its previous maneuver choice led to satisfactory progress. If lane keeping results in low speed or poor progress, the agent reconsiders the earlier decision and adjusts its next maneuver accordingly. Finally, the VLM adapter θ is fine-tuned on both tasks:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\,\mathbb{E}_{(x,\,y)\sim\mathcal{D}_{\text{SFT}}}[\log f_\theta(y \mid x)]\,, \tag{1}$$

where x comprises the task prompt and the agent's local observation $\mathbf{o}_t^\alpha$, which includes multi-view images and agent-specific state measurements. and y denotes the target output. Specifically, for the perception task, y is the local belief $b_t^\alpha = (s_t^\alpha, \gamma_t^\alpha)$, while for the decision task, y is the decision output (r_t^α, a_t^α), $\alpha \in \{\mathcal{A}^{\text{TLS}}, \mathcal{A}^{\text{CAV}}\}$ denotes either agent type.

4.3 Cooperative Belief Reasoning

Since $\mathcal{A}^{\text{TLS}}$ and $\mathcal{A}^{\text{CAV}}$ observe the traffic scene from different viewpoints, each agent may possess beliefs that are beneficial to the other's decision. We therefore introduce *cooperative belief reasoning*, where agents exchange semantic reports and learn to integrate them with local evidence for cooperative decision making. Cooperation enables $\mathcal{A}^{\text{TLS}}$ and $\mathcal{A}^{\text{CAV}}$ to enhance each other's situational awareness. In particular, the TLS can extend a CAV's early awareness by informing it of a blocked lane or a jaywalking pedestrian ahead, especially when the CAV's forward view is occluded. It can also indicate which lane is affected and whether an early lane change is advisable, thereby supporting more proactive maneuvering. Conversely, CAVs provide close-range or upstream evidence that may not yet be visible or clear to the TLS, such as an approaching emergency vehicle. These reports help the TLS make better-informed phase decisions at the next control step.

A key challenge is to align the semantic outputs of $\mathcal{A}^{\text{TLS}}$ and $\mathcal{A}^{\text{CAV}}$ into a shared context that supports agent-specific reasoning and decision adaptation. To this end, each agent communicates the event

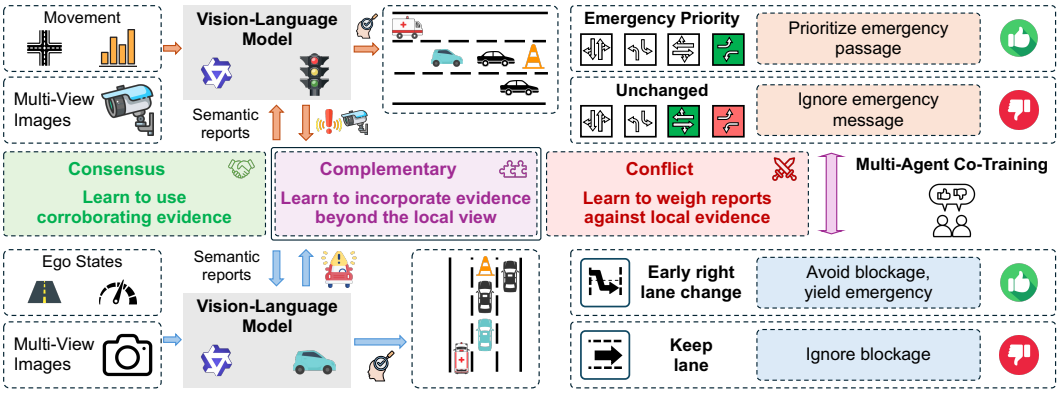


Figure 3: **Cooperative Belief Reasoning.** TLS and CAV agents exchange semantic reports and fuse them with local observations to generate rationales and actions. Multi-agent co-training jointly refines both agents using a shared reward, enabling them to adaptively incorporate shared evidence into their reasoning and decisions across CONSENSUS, COMPLEMENTARY, and CONFLICT cases.

fields of its local belief $b = (s, \gamma)$ as a semantic report $m = (e, \ell, \gamma)$, where e is the event type and ℓ is its affected location. The receiving agent α' then conditions on its current observation and the report received from agent α, $x = (\mathbf{o}_t^{\alpha'}, m_t^{\alpha})$, to jointly generate a rationale and an action, $(\tilde{r}_t^{\alpha'}, \tilde{a}_t^{\alpha'}) \sim f_\theta(\cdot \mid x)$. To learn this cooperative reasoning, we supervise the model with a rationale r_E derived from global event annotations and an agent-specific expert action a_E computed from privileged simulator states, by minimizing

$$\mathcal{L}_{\text{coop}}(\theta) = -\mathbb{E}_{(x, r_E, a_E) \sim \mathcal{D}_{\text{coop}}} \log f_\theta(r_E, a_E \mid x), \tag{2}$$

where $\mathcal{D}_{\text{coop}}$ pairs cooperative inputs containing ground-truth semantic reports with target rationales and expert actions. This objective teaches the agent to integrate local and shared evidence and to adapt its reasoning and decisions to the available context.

Multi-Agent Co-Training. In closed-loop execution, shared beliefs are generated by the agents' own perception and may be noisy. We therefore optimize the TLS and CAV agents jointly under a shared cooperative objective. At each decision step, we sample G joint rollouts from the same state. In each rollout, the agents generate and exchange reports and then produce their rationales and actions. Each rollout g receives a shared reward:

$$R_g = \sum_{\alpha \in \mathcal{A}} \Big(\lambda_r \, \phi\big(\tilde{b}_g^{\alpha}, b_E^{\alpha}\big) + \mathbb{I}\big[a_g^{\alpha} = a_E^{\alpha}\big] \Big), \tag{3}$$

where $\tilde{b}_g^{\alpha}$ is the belief reported by agent α, b_E^{α} is its reference belief derived from the ground-truth events observable by α, ϕ evaluates their consistency in event type and location, and $\lambda_r > 0$ balances the two terms. Expert actions a_E^{α} are computed from privileged simulator states: the TLS expert serves dischargeable demand to improve traffic efficiency while accounting for emergency priority, whereas the CAV expert selects maneuvers to improve travel efficiency and reduce waiting while considering emergency yielding. We optimize all agents with GRPO (Shao et al., 2024), applying the group-relative advantage of R_g to all tokens generated in rollout g. In consensus cases, shared reports corroborate the receiver's local observations, reinforcing the evidence for action selection. In complementary cases, reports extend the receivers awareness to events outside its view, supporting better-informed decisions. In conflict cases, discrepancies between shared reports and local observations require agents to reconcile inconsistent evidence and determine which information should guide their actions. By jointly optimizing report accuracy and action correctness, multi-agent co-training encourages agents to adaptively integrate shared and local evidence across these cases.

5 EXPERIMENTS

Experiments are conducted in the TranSimHub (Wang et al., 2025a), which integrates SUMO(Alvarez Lopez et al., 2018) for traffic simulation and Blender for rendering for multi-directional intersection and onboard surrounding-view camera images. We evaluate CoopLight on

two real-world signalized intersections reconstructed from OpenStreetMap: the Yau Ma Tei (YMT) intersection in Hong Kong, China, and an arterial intersection in California (CAL), U.S. These two intersections cover clearly different urban settings, traffic rules, and signal designs. YMT lies in a dense urban core with narrow roads, and uses a split-phasing design with 8 movements over 4 approaches, while CAL uses a wider multi-lane layout with 12 movements over 4 approaches and follows right-hand traffic with dedicated protected left-turn phases.

5.1 Experimental Setup

For each intersection, we construct 40 routes covering four long-tail scenario categories, with 10 routes per category under varying traffic-flow density levels, yielding a total of 120 minutes of traffic data. Each intersection is equipped with multi-directional cameras, and several CAVs travel through each route, each equipped with multi-view cameras. For each category, seven routes are used for training, and three are used for evaluation. We further include six routes where multiple scenario events co-occur. The VQA dataset contains 7828 pairs, including balanced coverage across scenarios. Both TLS and CAV agents are built on Qwen2.5-VL-3B-Instruct (Bai et al., 2025).

Baselines and Ablations. We compare against representative baselines from three categories. Rule-based methods include Fixed-Time and Max-Pressure (Varaiya, 2013), which use handcrafted timing rules or pressure-based heuristics. RL-based methods include PressLight (Wei et al., 2019), which learns signal control policies from vectorized traffic states, and CoTV (Guo et al., 2023), which jointly optimizes traffic signals and CAV control. VLM-based methods include VLMLight (Wang et al., 2025b), a hybrid of the VLM and RL method for traffic signal control under emergency-event scenarios. *CoopLight w/o CAV* keeps only the TLS agent. *CoopLight w/o Comm* uses both TLS and CAV agents, but without cross-agent communication. *CoopLight* further enables semantic sharing for cooperative decision making.

Evaluation Metrics. We evaluate CoopLight under closed-loop settings. In the closed-loop setting, following prior work, we report travel efficiency using Average Travel Time (ATT), and Average Waiting Time (AWT). ATT measures the mean time required for vehicles to reach their destinations. AWT measures the mean duration for which a vehicle remains nearly stationary, with speed below $0.1\,\mathrm{m/s}$. In emergency scenarios, we report the same metrics for emergency vehicles to assess responsiveness to high-priority traffic. We further report travel efficiency metrics for CAVs to assess how cooperation affects CAV performance. All results are averaged over 5 runs.

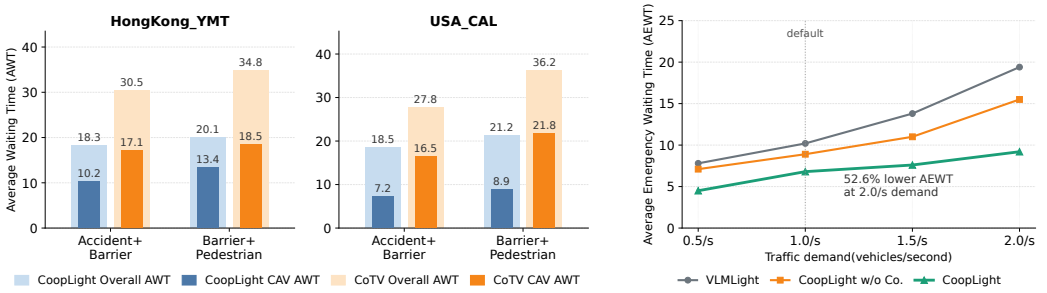
5.2 Main Results

Performance under Long-Tail Traffic Scenarios. Several trends emerge from Table 1. Methods based on fixed heuristics or compact vectorized traffic states are generally less robust under long-tail events, especially in the incident and barrier scenarios. Among the baselines, CoTV is the strongest overall, particularly on CAV travel efficiency, indicating the benefit of explicit TLS and CAV coordination. However, its performance still varies across scenes and event types, suggesting that coordination alone is insufficient without reasoning about anomaly semantics. By contrast, all CoopLight variants reason directly over scene semantics, enabling better understanding of true traffic demand. Moreover, the gains of CoopLight over CoopLight w/o Comm show that semantic sharing helps both agents form complementary beliefs and make more proactive and adaptive decisions.

Jaywalking. CoopLight performs best in both scenes, with the largest margin over CoTV on CAV metrics. A jaywalking pedestrian creates a brief conflict at an unpredictable location, which aggregate traffic states cannot distinguish from a routine fluctuation. The TLS detects the pedestrian in its junction-level camera views and reports it, allowing the CAV to adjust its maneuver before the pedestrian enters its ego view. This report reduces CAV AWT at YMT from $19.8\,\mathrm{s}$ to $10.2\,\mathrm{s}$, a larger gain than at CAL because narrower roads and denser traffic at YMT more often occlude the CAV's forward view. ***Incident/Barrier.*** Unlike this transient conflict, incidents and barriers block lanes and test whether the controller can distinguish blockage-induced queues from ordinary congestion. Under incidents, CoopLight achieves the best overall and CAV performance in both scenes, with clear gains over CoTV and much larger gains over VLMLight. The TLS agent localizes the affected lane and stops allocating green time to a movement whose queue cannot discharge, reducing Overall AWT at CAL from $27.1\,\mathrm{s}$ to $18.1\,\mathrm{s}$; each CAV recognizes the incident ahead from its ego view and changes lane, lowering CAV AWT at CAL from $18.2\,\mathrm{s}$ to $5.1\,\mathrm{s}$, and the TLS report further allows it

Table 1: Closed-loop evaluation across two intersection scenes and three long-tail scenario types.

Scenario	Method	USA_CAL				Hongkong_YMT			
		Overall		CAV		Overall		CAV	
		AWT↓	ATT↓	AWT↓	ATT↓	AWT↓	ATT↓	AWT↓	ATT↓
Jaywalking	Fixed-Time	**28.2** ±1.2	**56.8** ±1.7	18.5 ±0.9	47.3 ±1.5	**36.5** ±1.5	**58.6** ±2.0	27.3 ±1.1	49.5 ±1.6
	Max-Pressure	34.1 ±1.8	61.6 ±2.4	**14.6** ±0.8	**42.3** ±1.4	37.8 ±2.0	62.5 ±2.4	**14.5** ±0.8	**42.1** ±1.4
	PressLight	**29.5** ±2.6	**57.9** ±3.4	**19.2** ±1.8	**46.9** ±2.7	36.4 ±3.2	53.7 ±3.0	22.4 ±2.0	46.5 ±2.7
	CoTV	30.8 ±2.2	58.6 ±3.0	24.3 ±1.8	51.3 ±2.6	**33.3** ±2.4	**49.8** ±2.4	**18.7** ±1.4	**38.4** ±2.0
	VLMLight	26.8 ±1.5	55.3 ±2.4	27.5 ±1.8	54.9 ±2.6	28.3 ±1.6	48.4 ±2.2	28.7 ±1.6	55.2 ±2.5
	CoopLight w/o CAV	24.6 ±1.3	52.9 ±2.0	16.8 ±0.9	44.5 ±1.7	25.8 ±1.4	47.4 ±1.8	20.7 ±1.1	46.3 ±1.7
	CoopLight w/o Comm	22.1 ±1.0	50.4 ±1.6	11.6 ±0.7	39.5 ±1.4	23.5 ±1.1	46.9 ±1.7	19.8 ±0.9	41.3 ±1.5
	CoopLight (ours)	**18.9** ±0.8	**46.8** ±1.3	**9.4** ±0.5	**37.0** ±1.1	**19.5** ±0.8	**45.3** ±1.3	**10.2** ±0.6	**36.8** ±1.1
Incident	Fixed-Time	28.8 ±1.1	59.4 ±1.8	**19.5** ±0.9	**50.8** ±1.6	32.2 ±1.3	**55.5** ±1.9	28.8 ±1.1	**51.0** ±1.6
	Max-Pressure	**28.7** ±1.5	**57.7** ±2.2	23.1 ±1.2	55.3 ±2.0	**30.8** ±1.7	54.8 ±2.1	**22.3** ±1.2	47.5 ±1.7
	PressLight	21.3 ±2.0	51.7 ±2.8	17.3 ±1.6	47.5 ±2.5	26.8 ±2.4	46.9 ±2.4	18.8 ±1.7	48.1 ±2.6
	CoTV	**18.8** ±1.4	**49.6** ±2.2	**10.8** ±1.0	**42.4** ±1.9	**26.0** ±1.8	**44.9** ±1.9	**15.2** ±1.1	**34.2** ±1.5
	VLMLight	27.1 ±1.5	54.3 ±2.3	29.7 ±1.7	53.1 ±2.4	26.4 ±1.5	53.5 ±2.3	27.9 ±1.6	52.7 ±2.4
	CoopLight w/o CAV	18.1 ±1.0	47.9 ±1.7	18.2 ±1.0	48.8 ±1.8	19.8 ±1.1	45.6 ±1.6	20.2 ±1.1	47.8 ±1.8
	CoopLight w/o Comm	16.2 ±0.7	45.4 ±1.4	5.1 ±0.4	33.3 ±1.0	15.9 ±0.7	41.3 ±1.3	15.3 ±0.7	36.1 ±1.1
	CoopLight (ours)	**13.3** ±0.6	**43.7** ±1.2	**4.2** ±0.3	**31.9** ±0.9	**13.2** ±0.6	**37.9** ±1.1	**14.1** ±0.6	**33.6** ±0.9
Barrier	Fixed-Time	**29.3** ±1.2	**58.8** ±2.0	**21.5** ±1.0	**51.8** ±1.7	35.8 ±1.5	**58.8** ±2.0	39.6 ±1.7	62.5 ±2.1
	Max-Pressure	35.5 ±1.8	63.2 ±2.5	24.3 ±1.4	55.6 ±1.9	**34.3** ±1.8	57.2 ±2.1	**24.1** ±1.4	**48.8** ±1.7
	PressLight	**32.2** ±2.8	**61.1** ±3.6	24.5 ±2.0	55.2 ±3.2	46.3 ±3.6	63.5 ±3.4	25.6 ±2.0	49.6 ±2.6
	CoTV	33.1 ±2.4	61.4 ±3.2	**13.2** ±1.0	**43.4** ±2.2	**27.9** ±2.0	**47.4** ±2.2	**8.8** ±0.7	**29.5** ±1.6
	VLMLight	28.9 ±1.6	59.1 ±2.6	37.3 ±2.0	62.3 ±2.8	30.8 ±1.7	57.8 ±2.5	36.5 ±2.0	61.4 ±2.7
	CoopLight w/o CAV	27.3 ±1.4	56.5 ±2.0	22.8 ±1.2	53.2 ±1.9	29.5 ±1.5	51.3 ±1.9	23.5 ±1.2	47.2 ±1.7
	CoopLight w/o Comm	24.5 ±1.2	53.5 ±1.7	3.7 ±0.3	32.1 ±1.2	27.2 ±1.3	46.2 ±1.6	13.5 ±0.7	33.5 ±1.2
	CoopLight (ours)	**19.6** ±0.8	**49.3** ±1.4	**3.6** ±0.3	**31.5** ±1.0	**23.2** ±1.0	**42.7** ±1.3	**12.3** ±0.6	**32.5** ±1.0



(a) (b)

Figure 4: (a) CoopLight consistently achieves a lower AWT than CoTV under co-occurring long-tail scenarios across the YMT and CAL intersections. (b) Under the emergency scenario, CoopLight incurs a smaller increase in average emergency waiting time than VLMLight (Wang et al., 2025b) as traffic demand rises.

to react before the incident enters its view. Barriers, in contrast, persist throughout the episode, so phase switching alone offers little relief: the TLS agent by itself barely improves over VLMLight (28.9 s versus 27.3 s), and the gain instead comes from CAVs rerouting early on the basis of TLS reports, which reduces Overall AWT at CAL from 24.5 s to 19.6 s. The exception is CAV AWT at YMT under barriers, where CoTV remains lowest (8.8 s versus 12.3 s) while CoopLight attains the lower overall AWT, indicating that it improves intersection-level rather than individual-vehicle efficiency. ***Emergency-Priority.*** Emergency vehicles instead require timely right-of-way. Figure 4b compares their average waiting time under increasing traffic demand, where denser traffic makes the approaching vehicle harder to detect and localize. Although VLMLight supports emergency-aware signal switching, its more general scene understanding leads to less effective phase selection as demand increases. CoopLight w/o Comm degrades more slowly owing to its confidence-aware spatial scene understanding, while CoopLight remains the most robust because semantic cooperation provides clearer evidence of emergency presence under heavy traffic.

Co-Occurrence. As illustrated in Figure 4a, CoopLight shows stronger adaptability than CoTV when multiple long-tail events occur simultaneously at the same intersection, as evidenced by consistently lower overall and CAV AWT across both evaluated intersections and co-occurrence settings. This advantage is consistent with the complementary roles of the two agent types: the TLS agent allocates green time to movements with dischargeable demand, while CAVs use shared semantic reports to anticipate blockages, change lanes proactively, and yield to crossing pedestrians.

Cross-Intersection Generalization. To assess transfer across intersection layouts, we train on one intersection and evaluate on the other without target fine-tuning. Despite differences in driving side, lane geometry, and signal design, CoopLight achieves lower Overall AWT than VLMLight in both ID and OOD settings (Table 2). Transfer increases AWT for both methods, but CoopLight incurs a smaller mean relative degradation (14.9% versus 25.5%). Moreover, its OOD performance remains better than VLMLight's ID performance at both intersections, showing that its efficiency advantage persists even without training on the target layout. These results support transfer across the two evaluated intersections. Per-scenario results are provided in Appendix D.1.

Table 2: Overall AWT (s), averaged over three scenarios. ID/OOD: trained on the same/other intersection, with no target fine-tuning for OOD. Δ: mean per-scenario relative increase across both transfer directions.

Method	YMT		CAL		$\Delta \downarrow$
	ID	OOD	ID	OOD	
VLMLight	28.5	35.2	27.6	35.1	25.5%
CoopLight	18.6	**21.4**	17.3	**19.7**	**14.9%**

Robustness to Visual Degradation. Table 3 evaluates the same traffic settings under different weather and illumination conditions. All methods incur higher AWT under adverse visibility, but CoopLight remains the most efficient in every condition. Relative to the no-communication variant, it reduces AWT by 16.7–17.8% across the four conditions, indicating that the benefit of semantic exchange persists under visual degradation. From clear daytime to fog, its AWT increases by 28.3%, compared with 30.1% without communication and 36.7% for VLMLight. Thus, cooperation preserves an efficiency advantage while only modestly reducing relative degradation. Full AWT and ATT results are provided in Appendix D.2.

Table 3: AWT (s) $\downarrow$ under visual condition variations, averaged over both intersections. Traffic demand, events, and seeds are fixed; no retraining.

Method	Clear	Dusk	Rain	Fog
VLMLight	28.1	31.3	34.6	38.4
CoopLight w/o Comm	21.6	23.6	25.3	28.1
CoopLight	**18.0**	**19.5**	**20.8**	**23.1**

5.3 Effect of Joint Decision Making and Cooperation

Across the three scenarios and two intersections, CoopLight reduces Overall AWT by 26% and CAV AWT by 56% relative to *CoopLight w/o CAV*. The largest gain appears in the CAL barrier scenario, where CAV AWT drops from 22.8 s to 3.6 s. This shows that joint decision making and cooperation provide substantial benefits beyond task-specific VLM grounding alone. In particular, moving from *CoopLight w/o CAV* to *CoopLight w/o Comm* already yields large gains, especially in the incident and barrier scenarios, indicating that long-tail traffic events affect both signal control and CAV travel efficiency. Therefore, vehicle-side perception-grounded decision making is already important for handling local lane-level disruptions. *CoopLight* further improves over *CoopLight w/o Comm*, suggesting that semantic sharing enriches the shared context and improves how both vehicle-side and infrastructure-side agents use it under long-tail scenarios.

6 Conclusion

This paper presents CoopLight, a vision-language framework for cooperative traffic signal and CAV control under long-tail traffic scenarios. By combining task-specific VQA supervision, confidence-aware spatial scene understanding, context-aware decision making, and semantic cross-agent cooperation, CoopLight enables both infrastructure-side and vehicle-side agents to reason over complex traffic conditions and act more proactively. Experiments on simulated replicas of two real signalized intersections across diverse long-tail scenarios show that CoopLight consistently improves traffic efficiency over rule-based, reinforcement-learning-based, and prior vision-language-driven baselines, highlighting the value of semantic reasoning and cooperation for next-generation urban mobility.

AI USE STATEMENT

In this work, we did not use generative AI tools for any task with required disclosure, including developing the conceptual framework, formulating hypotheses, designing the methodology and experiments, implementing methods, generating or processing data, and interpreting results; translation, qualitative data analysis, and proof writing are not applicable to this work. Additionally, we used generative AI tools to edit the manuscript for grammar and readability. All AI-edited text was reviewed and revised by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

CoopLight has the potential to reduce waiting and travel times at urban signalized intersections and improve the reliability of emergency vehicle passage. Potential negative impacts arise from the use of roadside and onboard cameras, which introduce privacy considerations because real-world deployments may capture personally identifiable information from pedestrians and drivers.

We therefore recommend that any real-world deployment apply on-device anonymization (e.g., face and license-plate blurring) and comply with local data-protection regulations.

REPRODUCIBILITY STATEMENT

We describe the problem formulation in Section 3, and the full method in Section 4. The experimental setup, including the simulation platform, the two real-world intersections, long-tail scenario construction, traffic signal and CAV settings, baselines, and compute resources, is given in Section 5.1 and Appendix B. Representative VQA prompts and responses used to train and evaluate both agents are provided in Appendix E. Code, VQA data generation scripts, and trained checkpoints will be released upon publication.

REFERENCES

Pablo Alvarez Lopez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wießner. Microscopic traffic simulation using SUMO. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 2575–2582. IEEE, 2018.

Mohammad Aslani, Mohammad Saadi Mesgari, and Marco Wiering. Adaptive traffic signal control with actor-critic methods in a real-world traffic network with different traffic disruption events. *Transportation Research Part C: Emerging Technologies*, 85:732–752, 2017.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Azita Dabiri, Giray Önür, Sebastien Gros, and Bart De Schutter. Optimization-based coordination of traffic lights and automated vehicles at intersections. *arXiv preprint arXiv:2502.01315*, 2025.

Myungeun Eom and Byung-In Kim. The traffic signal control problem for intersections: A review. *European Transport Research Review*, 12(1):50, 2020.

Md. Abdul Fattah, Syed Riad Morshed, and Abdulla-Al Kafy. Insights into the socio-economic impacts of traffic congestion in the port and industrial areas of chittagong city, bangladesh. *Transportation Engineering*, 9:100122, 2022.

Jiaying Guo, Long Cheng, and Shen Wang. CoTV: Cooperative control for traffic light signals and connected autonomous vehicles using deep reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems*, 24(10):10501–10512, 2023.

Qiangqiang Guo, Li Li, and Xuegang (Jeff) Ban. Urban traffic signal control with connected and automated vehicles: A survey. *Transportation Research Part C: Emerging Technologies*, 101: 313–334, 2019.

Ashesh Jain, Luca Del Pero, Hugo Grimmett, and Peter Ondruska. Autonomy 2.0: Why is self-driving always 5 years away? *arXiv preprint arXiv:2107.08142*, 2021.

Siqi Lai, Zhao Xu, Weijia Zhang, Hao Liu, and Hui Xiong. LLMLight: Large language models as traffic signal control agents. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 2335–2346, 2025.

Xiaoyuan Liang, Xunsheng Du, Guiling Wang, and Zhu Han. A deep reinforcement learning network for traffic light cycle control. *IEEE Transactions on Vehicular Technology*, 68(2):1243–1253, 2019.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023.

Henry X. Liu and Shuo Feng. Curse of rarity for autonomous vehicles. *Nature Communications*, 15 (1):4808, 2024.

Jiageng Mao, Junjie Ye, Yuxi Qian, Marco Pavone, and Yue Wang. A language agent for autonomous driving. In *Conference on Language Modeling*, 2024.

Zhaobin Mo, Wangzhi Li, Yongjie Fu, Kangrui Ruan, and Xuan Di. CVLight: Decentralized learning for adaptive traffic signal control with connected vehicles. *Transportation Research Part C: Emerging Technologies*, 141:103728, 2022.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. DriveVLM: The convergence of autonomous driving and large vision-language models. In *Proceedings of the 8th Conference on Robot Learning (CoRL)*, volume 270 of *Proceedings of Machine Learning Research*, pp. 4698–4726. PMLR, 2025.

Pravin Varaiya. Max pressure control of a network of signalized intersections. *Transportation Research Part C: Emerging Technologies*, 36:177–195, 2013.

Maonan Wang, Xi Xiong, Yuheng Kan, Chengcheng Xu, and Man-On Pun. UniTSA: A universal reinforcement learning framework for V2X traffic signal control. *IEEE Transactions on Vehicular Technology*, 73(10):14354–14369, 2024.

Maonan Wang, Yirong Chen, Yuxin Cai, Aoyu Pang, Yuejiao Xie, Zian Ma, Chengcheng Xu, Kemou Jiang, Ding Wang, Laurent Roullet, Chung Shue Chen, Zhiyong Cui, Yuheng Kan, Michael Lepech, and Man-On Pun. TranSimHub: A unified air-ground simulation platform for multi-modal perception and decision-making. *arXiv preprint arXiv:2510.15365*, 2025a.

Maonan Wang, Yirong Chen, Aoyu Pang, Yuxin Cai, Chung Shue Chen, Yuheng Kan, and Man On Pun. VLMLight: Safety-critical traffic signal control via vision-language meta-control and dual-branch reasoning architecture. In *Advances in Neural Information Processing Systems*, volume 38, pp. 39590–39621, 2025b.

F. V. Webster. Traffic signal settings. Road Research Technical Paper 39, Road Research Laboratory, London, UK, 1958.

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 & Data Mining (KDD)*, pp. 2496–2505, 2018.

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 & Data Mining (KDD)*, pp. 1290–1298, 2019.

Jiaming Wu and Xiaobo Qu. Intersection control with connected and automated vehicles: A review. *Journal of Intelligent and Connected Vehicles*, 5(3):260–269, 2022.

Taokai Xia and Hui Chen. A survey of autonomous vehicle behaviors: Trajectory planning algorithms, sensed collision risks, and user expectations. *Sensors*, 24(15):4808, 2024.

Shuhua Xu, Chuanwang Sun, and Nian Liu. Road congestion and air pollution – analysis of spatial and temporal congestion effects. *Science of The Total Environment*, 945:173896, 2024a.

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K. Wong, Zhenguo Li, and Hengshuang Zhao. DriveGPT4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 9(10):8186–8193, 2024b.

Zirui Yuan, Siqi Lai, and Hao Liu. CoLLMLight: Cooperative large language model agents for network-wide traffic signal control. In *International Conference on Learning Representations*, volume 2026, pp. 66687–66715, 2026.

Yifeng Zhang, Ping Gong, Weiyi He, Yilin Liu, Mingfeng Fan, and Guillaume Sartoretti. V2XFormer: A multi-stage transformer for multi-agent reinforcement learning in v2x-enabled traffic signal control. In *2025 IEEE 28th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 2686–2693, 2025.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024.

Shuai Zheng, Yugang Liu, Kui Fu, Rongrong Li, You Zhang, and Hongtai Yang. Optimization of isolated intersection signal timing and trajectory planning under mixed traffic environment: The flexible catalysis of connected and automated vehicles. *Physica A: Statistical Mechanics and its Applications*, 640:129668, 2024.

Xingcheng Zhou, Mingyu Liu, Ekim Yurtsever, Bare Luka Zagar, Walter Zimmer, Hu Cao, and Alois C. Knoll. Vision language models in autonomous driving: A survey and outlook. *IEEE Transactions on Intelligent Vehicles*, pp. 1–20, 2024. Early Access.

Xingchen Zou, Yuhao Yang, Zheng Chen, Xixuan Hao, Yiqi Chen, Chao Huang, and Yuxuan Liang. Traffic-R1: Reinforced LLMs bring human-like reasoning to traffic signal control systems. *arXiv preprint arXiv:2508.02344*, 2025.

A Appendix Overview

In this appendix, we provide supplementary material to further elaborate on the experimental design, implementation details, and additional analyses of CoopLight:

- Experimental setup, including the simulation platform, compute resources, inference and communication latency, and map construction (Section B).
- Perception-level evaluation (Section C).
- Cross-intersection generalization and robustness to weather and illumination variation (Section D).
- Representative VQA examples illustrating TLS-side and CAV-side perception and decision making under barrier, incident, emergency-vehicle, and jaywalking-pedestrian scenarios in Section E.

B Experimental Setup

In this section, we describe the experimental setup used to evaluate CoopLight, our cooperative vision-language framework for traffic signal and CAV control. The experiments involve a realistic traffic simulator, multi-view visual rendering and map construction.

Simulation Platform. All experiments are conducted in the TranSimHub traffic simulator, which integrates SUMO for microscopic traffic simulation and Blender-based rendering for visual observation generation. SUMO (Version 1.26) simulates vehicle dynamics, lane-level traffic flow, signal phases, and interactions between human-driven vehicles and CAVs, while the Blender-based renderer provides multi-view roadside and onboard camera images. This setup allows CoopLight to evaluate traffic control decisions using both structured traffic measurements and visual information.

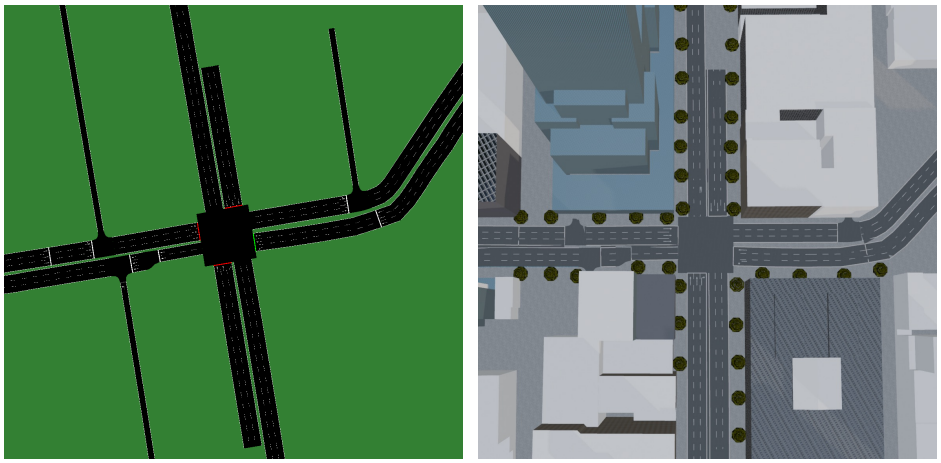
Compute Resources. All training and inference experiments are conducted on a local workstation equipped with an Intel Core i9-14900KF CPU, 64 GB of RAM, and an NVIDIA GeForce RTX 4090 GPU, running Ubuntu 22.04 LTS. All VLM agents are deployed locally rather than through external APIs, ensuring reproducible inference behavior and avoiding network-induced latency variation during closed-loop simulation.

Inference and Communication Latency. The TLS agent makes a decision every $5\,\mathrm{s}$ when the current phase is retained, or every $8\,\mathrm{s}$ when a phase switch includes a $3\,\mathrm{s}$ yellow transition; each CAV agent acts at every $1\,\mathrm{s}$ control step. A semantic report contains the event type, affected location, confidence level, and a brief description, with a serialized size of at most $0.5\,\mathrm{KB}$. Table 4 reports the average inference latency and the communication latency measured on the workstation above.

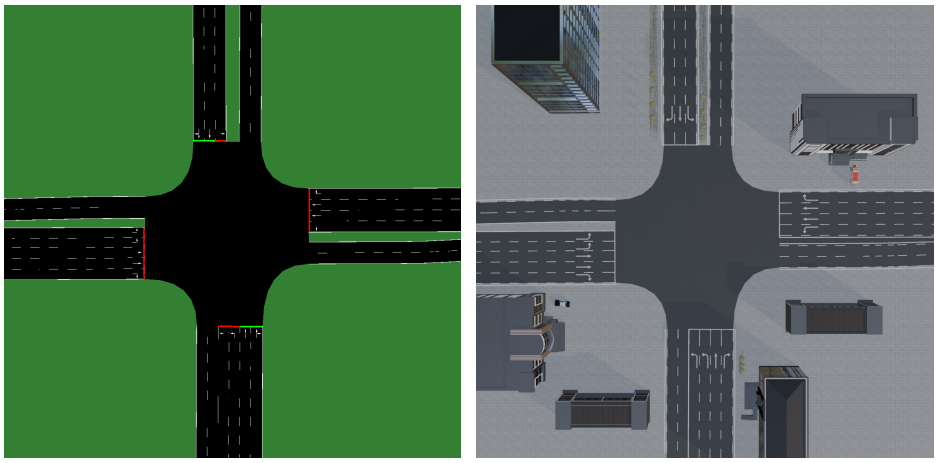
Table 4: Decision intervals and measured latencies (in seconds) of the TLS and CAV agents.

Agent	Decision interval	Inference latency	Communication latency
TLS	5.000 (retain) / 8.000 (switch)	0.727	0.0053 (TLS→CAV)
CAV	1.000	0.814	0.0056 (CAV→TLS)

Map Construction. We use two realistic signalized intersections in experiments. The Hong Kong Yau Ma Tei (YMT) map is adopted from the TranSimHub simulation platform, including its SUMO road network and Blender-based visual scene. For the Southern California (CAL) intersection, we construct the SUMO road network from OpenStreetMap and build the corresponding Blender scene for high-fidelity image rendering. The CAL map is manually checked to ensure valid lane connectivity, turning movements, route definitions, and traffic signal phases for closed-loop simulation.



(a) Hong Kong Yau Ma Tei (YMT).



(b) Southern California (CAL).

Figure 5: Two real-world intersections: (a) Yau Ma Tei, Hong Kong, and (b) Southern California, United States. Each row shows the SUMO simulation view on the left and the Blender rendering on the right.

C Perception-Level Evaluation

Semantic perception is evaluated on the VQA test set, stratified by intersection and agent type. Table 5 reports, for each agent's belief without and with communication (w/o Comm and w/ Comm), (i) detection accuracy for event type, approach, and lane; (ii) confidence-label accuracy; and (iii) event macro-F1 and the normal-scene false-positive rate (FPR). Without communication, perception is weaker for CAVs than for the TLS, because ego-centric views cover only part of the scene, and weaker at YMT than at CAL because denser traffic and narrower roads at YMT cause more occlusion. Communication improves every metric, most notably CAV-side lane-level detection, where the intersection-level view of the TLS complements the partial view of a CAV. Reports from the other agent recover events missed locally, reinforce detections on which both agents agree, and correct uncertain local beliefs. Averaged over the four agent–intersection settings, communication raises detection accuracy by 4.6, 4.1, and 6.6 points at the event, approach, and lane levels, confidence-label accuracy by 6.5, and macro-F1 by 5.5, while reducing the FPR from 7.8% to 4.1%.

Table 5: Perception accuracy (%) of beliefs without and with communication (w/o Comm and w/ Comm), by agent and intersection. FPR: normal-scene false-positive rate.

		Detection Acc. ↑											
		Event		Approach		Lane		Conf. Acc. ↑		Macro-F1 ↑		FPR ↓	
Agent	Intersection	w/o Comm	w/ Comm	w/o Comm	w/ Comm	w/o Comm	w/ Comm	w/o Comm	w/ Comm	w/o Comm	w/ Comm	w/o Comm	w/ Comm
TLS	CAL	92.4	**95.1**	89.5	**93.2**	88.2	**91.4**	73.6	**80.8**	90.2	**93.4**	5.6	**3.0**
	YMT	87.6	**91.3**	89.4	**92.1**	81.5	**86.7**	78.4	**82.6**	85.6	**90.6**	7.4	**4.0**
CAV	CAL	85.8	**91.6**	88.1	**92.7**	78.6	**86.9**	76.7	**83.5**	84.1	**90.3**	7.9	**4.2**
	YMT	81.2	**87.4**	84.6	**89.8**	72.3	**82.1**	72.5	**80.4**	79.7	**87.1**	10.3	**5.2**
Average		86.8	**91.4**	87.9	**92.0**	80.2	**86.8**	75.3	**81.8**	84.9	**90.4**	7.8	**4.1**

Table 6 reports perception accuracy at each confidence level, to test whether higher confidence indicates a more reliable perception. For both agents, accuracy drops sharply as confidence decreases: without communication, it falls from 93.5% at HIGH to 78.2% at LIKELY and 55.4% at UNCERTAIN for the TLS, and from 90.1% to 72.8% and 49.6% for CAVs. HIGH-confidence beliefs are therefore reliable, whereas about half of UNCERTAIN CAV beliefs are wrong. Communication further raises accuracy at every confidence level while preserving this ordering, with the largest gains at UNCERTAIN (6.7 points for the TLS and 8.7 for CAVs), where local evidence is weakest.

Table 6: Perception accuracy (%) at each confidence level, by agent and averaged over intersections. A belief is correct if both the event type and the lane are correct.

	TLS		CAV	
Confidence level	w/o Comm	w/ Comm	w/o Comm	w/ Comm
HIGH	93.5	**95.8**	90.1	**94.2**
LIKELY	78.2	**83.6**	72.8	**80.5**
UNCERTAIN	55.4	**62.1**	49.6	**58.3**

D Generalization and Robustness

D.1 Cross-Intersection Generalization

To assess transfer to an unseen layout, each model is trained on one intersection and evaluated on the other without fine-tuning. The two intersections differ in driving side, geometry, and signal design: CAL has right-hand traffic, wide multi-lane approaches, and protected left-turn phases, whereas YMT has left-hand traffic, narrow roads, and split phasing. Table 7 reports Overall AWT. Relative to in-domain training, cross-trained CoopLight incurs an average AWT increase of 14.9%, versus 25.5% for VLMLight. The smaller degradation is consistent with the perception task design: events

are localized to named approaches and movements defined per intersection, so the learned reasoning transfers across layouts.

Table 7: Leave-one-intersection-out evaluation (Overall AWT, seconds). Cross-trained models are trained only on the other intersection.

Test intersection	Setting	Jaywalking ↓	Incident ↓	Barrier ↓
YMT	VLMLight (trained on CAL)	39.8	32.3	33.5
	VLMLight (in-domain)	28.3	26.4	30.8
	CoopLight (trained on CAL)	22.4	15.8	26.0
	CoopLight (in-domain)	**19.5**	**13.2**	**23.2**
CAL	VLMLight (trained on YMT)	36.6	29.2	39.6
	VLMLight (in-domain)	26.8	27.1	28.9
	CoopLight (trained on YMT)	21.6	15.1	22.5
	CoopLight (in-domain)	**18.9**	**13.3**	**19.6**

D.2 Robustness to Visual Degradation

Weather and Illumination. The same long-tail events are evaluated under the four weather and illumination conditions supported by TranSimHub: clear daytime, cloudy dusk, rainfall, and fog. All conditions share the same traffic demand, event types and locations, and random seeds, and differ only in the renderer's weather and illumination settings; all methods are evaluated without retraining. Table 8 reports closed-loop results. From clear daytime to fog, CoopLight's AWT and ATT increase by 28.3% and 13.3%, compared with 30.1% and 16.5% for CoopLight w/o Comm and 36.7% and 22.1% for VLMLight. Table 9 reports the perception quality of CoopLight's beliefs without and with communication under the same conditions. Communication limits the macro-F1 decrease from clear daytime to fog to 9.5 points, versus 12.8 points without communication, with similar reductions in lane-level detection loss and false positives. The complementary viewpoints of the TLS and CAVs compensate for each other's degraded local evidence, and this effect becomes more pronounced as visibility decreases.

Table 8: Closed-loop performance (seconds) under weather and illumination variation, averaged over both intersections.

	VLMLight		CoopLight w/o Comm		CoopLight	
Condition	AWT ↓	ATT ↓	AWT ↓	ATT ↓	AWT ↓	ATT ↓
Clear daytime	28.1	54.7	21.6	47.3	**18.0**	**44.3**
Cloudy dusk	31.3	58.4	23.6	49.8	**19.5**	**46.1**
Rainfall	34.6	62.3	25.3	51.9	**20.8**	**47.5**
Fog	38.4	66.8	28.1	55.1	**23.1**	**50.2**

Table 9: Perception accuracy (%) of beliefs without and with communication (w/o Comm and w/ Comm) under weather and illumination variation, averaged over agents and intersections. The clear-daytime row corresponds to the average row of Table 5.

	Detection Acc. ↑											
	Event		Approach		Lane		Conf. Acc. ↑		Macro-F1 ↑		FPR ↓	
Condition	w/o Comm	w/ Comm	w/o Comm	w/ Comm	w/o Comm	w/ Comm	w/o Comm	w/ Comm	w/o Comm	w/ Comm	w/o Comm	w/ Comm
Clear daytime	86.8	**91.4**	87.9	**92.0**	80.2	**86.8**	75.3	**81.8**	84.9	**90.4**	7.8	**4.1**
Cloudy dusk	84.5	**89.0**	86.1	**90.6**	76.4	**84.0**	73.1	**79.9**	82.7	**87.8**	9.8	**5.6**
Rainfall	80.3	**85.9**	83.0	**88.3**	71.2	**80.1**	69.8	**77.2**	78.2	**84.6**	12.6	**7.9**
Fog	74.6	**82.3**	78.4	**85.5**	64.5	**75.6**	64.9	**73.8**	72.1	**80.9**	17.4	**10.8**

E VQA EXAMPLES

This section provides representative VQA examples for the two agent types in CoopLight: the infrastructure-side TLS agent and the vehicle-side CAV agent. For each agent type, we show examples under four long-tail scenarios: barrier, incident, emergency vehicle, and jaywalking pedestrian.

E.1 INFRASTRUCTURE-SIDE VQA EXAMPLES

E.1.1 BARRIER



[EVENT DETECTION] Examine all junction camera images carefully. Is there any special event at this intersection? Answer Yes or No first, then describe if yes.

Detector readings: Phase 0: 10 vehicles, Phase 1: 5 vehicles, Phase 2: 4 vehicles, Phase 3: 7 vehicles

Yes. Road barrier on the SOUTH right-turn lane (Phase 0). 2 vehicles blocked. **Confidence: HIGH.**

[PHASE DECISION] Based on the junction camera images and detector readings, which traffic phase should receive green next?

Detector readings: Phase 0: 10 vehicles, Phase 1: 5 vehicles, Phase 2: 4 vehicles, Phase 3: 7 vehicles

Options: A. Phase 0 B. Phase 1 C. Phase 2 **D. Phase 3**

<think>Road barrier on SOUTH right-turn lane (Phase 0, North+South Straight+Right).
Detector reads 10 on Phase 0, but only 5 vehicles in Phase 0 can clear on green.
Phase 3, East+West Left has the highest effective demand. Serve Phase 3.
Confidence: HIGH.</think> **D**

Figure 6: TLS-side barrier example.

E.1.2 INCIDENT



[EVENT DETECTION] Examine all junction camera images carefully. Is there any special event at this intersection? Answer Yes or No first, then describe if yes.

Detector readings: Phase 0: 7 vehicles, Phase 1: 10 vehicles, Phase 2: 9 vehicles, Phase 3: 8 vehicles

Yes. Crash on the WEST straight lane (Phase 2). 2 vehicle blocked. **Confidence: HIGH.**

[PHASE DECISION] Based on the junction camera images and detector readings, which traffic phase should receive green next?

Detector readings: Phase 0: 7 vehicles, Phase 1: 10 vehicles, Phase 2: 9 vehicles, Phase 3: 8 vehicles

Options: A. Phase 0 **B. Phase 1** C. Phase 2 D. Phase 3

Crash on WEST straight lane (Phase 2), ~2 vehicles blocked. Phase 1 has highest demand. Serve **Phase 1**. Confidence: UNCERTAIN.

Figure 7: TLS-side incident example.



[EVENT DETECTION] Examine all junction camera images carefully. Is there any special event at this intersection? Answer Yes or No first, then describe if yes. Detector readings: Phase 0: 2 vehicles, Phase 1: 4 vehicles, Phase 2: 6 vehicles, Phase 3: 4 vehicles

Yes. Crash on the SOUTH right-turn lane (Phase 0).
Confidence: UNCERTAIN.

[PHASE DECISION] Based on the junction camera images and detector readings, which traffic phase should receive green next?

Detector readings: Phase 0: 2 vehicles, Phase 1: 4 vehicles, Phase 2: 6 vehicles, Phase 3: 4 vehicles

Options: A. Phase 0 B. Phase 1 **C. Phase 2** D. Phase 3

Possible crash on SOUTH right-turn lane (Phase 0, North+South Straight+Right), details unclear at this range. If present, Phase 0, North+South Straight+Right still has 1 vehicles that can move. Phase 2, East+West Straight+Right has highest demand. **Serve Phase 2**. Confidence: UNCERTAIN.

Figure 8: TLS-side incident example.

E.1.3 EMERGENCY VEHICLE



[EVENT DETECTION] Examine all junction camera images carefully. Is there any special event at this intersection? Answer Yes or No first, then describe if yes. Detector readings: Phase 0: 3 vehicles, Phase 1: 1 vehicles, Phase 2: 5 vehicles, Phase 3: 3 vehicles

Yes. Fire engine on the WEST straight lane (Phase 2), heading toward the intersection. **Confidence: HIGH.**

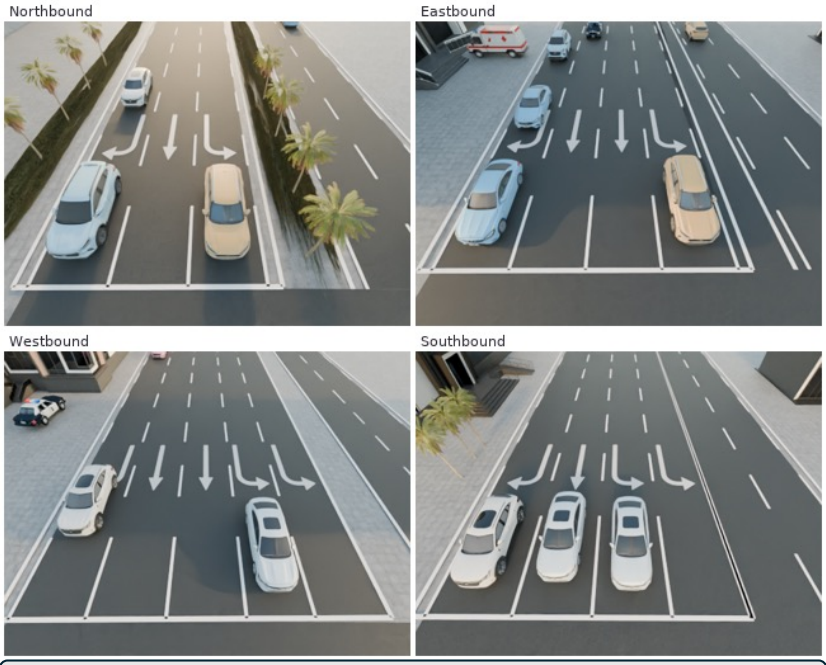
[PHASE DECISION] Based on the junction camera images and detector readings, which traffic phase should receive green next?

Detector readings: Phase 0: 3 vehicles, Phase 1: 1 vehicles, Phase 2: 5 vehicles, Phase 3: 3 vehicles

Options: A. Phase 0 B. Phase 1 **C. Phase 2** D. Phase 3

Fire engine approaching. Phase 2 has the highest effective demand. Serve Phase 2. Confidence: HIGH. **C**

Figure 9: TLS-side emergency-vehicle example.



[EVENT DETECTION] Examine all junction camera images carefully. Is there any special event at this intersection? Answer Yes or No first, then describe if yes.

Detector readings: Phase 0: 4 vehicles, Phase 1: 2 vehicles, Phase 2: 6 vehicles, Phase 3: 2 vehicles

Yes. Police vehicle on the EAST straight lane (Phase 2), heading toward the intersection. Confidence: LIKELY.

[PHASE DECISION] Based on the junction camera images and detector readings, which traffic phase should receive green next?

Detector readings: Phase 0: 4 vehicles, Phase 1: 2 vehicles, Phase 2: 6 vehicles, Phase 3: 2 vehicles

Options: A. Phase 0 B. Phase 1 **C. Phase 2** D. Phase 3

Police vehicle possibly approaching on EAST straight lane (Phase 2) — priority green. Serve Phase 2. Confidence: LIKELY. **C**

Figure 10: TLS-side emergency-vehicle example.

E.1.4 JAYWALKING PEDESTRIAN



[EVENT DETECTION] Examine all junction camera images carefully. Is there any special event at this intersection? Answer Yes or No first, then describe if yes.

Detector readings: Phase 0: 2 vehicles, Phase 1: 1 vehicles, Phase 2: 0 vehicles, Phase 3: 0 vehicles

Yes. Crossing pedestrian on the NORTH right-turn lane (Phase 0). **Confidence: HIGH**.

[PHASE DECISION] Based on the junction camera images and detector readings, which traffic phase should receive green next?

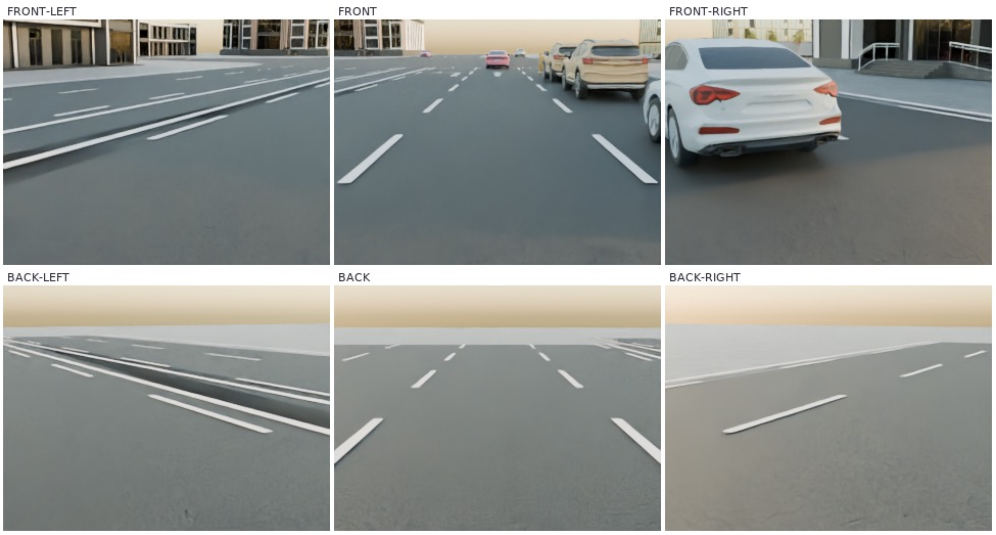
Detector readings: Phase 0: 2 vehicles, Phase 1: 1 vehicles, Phase 2: 0 vehicles, Phase 3: 0 vehicles

Options: **A. Phase 0** B. Phase 1 C. Phase 2 D. Phase 3

<think>Pedestrian obstruction on NORTH right-turn lane (Phase 0, North+South Straight+Right). Detector reads 2 on Phase 0, but only 1 vehicles in Phase 0 can clear on green. Phase 0, North+South Straight+Right has the highest effective demand. Serve Phase 0. Confidence: HIGH.</think> **A**

Figure 11: TLS-side jaywalking-pedestrian example.

E.2 VEHICLE-SIDE VQA EXAMPLES



CAV ego_1: Examine your multi-view camera images. Is there any anomalous event, object, or special vehicle visible? If yes, identify the event type, its relative position, and how many vehicles appear blocked behind it. If no anomaly is visible, briefly describe the traffic conditions ahead that may be relevant for driving decisions. Reason step-by-step through the cameras before concluding; end with "Confidence: HIGH, LIKELY, or UNCERTAIN".

No. Stopped vehicle ahead at 35m while other traffic flows; no anomaly visible. **Confidence: LIKELY.**

CAV ego_1 (Speed: 16.6 m/s, current lane: 3, ~40m to stop line). Available actions: Keep lane, Change lane left, Change lane right. Based on your multi-view camera images, what is the optimal lane decision? Note: last tick you choose "change lane left"; speed decreased 0.9 m/s, lane unchanged.

[ANOMALY DETECTION] Local: No. Stopped vehicle ahead at 35m while other traffic flows; no anomaly visible. Confidence: LIKELY. Peer (ITS): A crashed vehicle on the Eastbound lane 2, you are at Eastbound lane 3, your lane is not affected. Confidence: HIGH.

Options: A. Keep lane B. Change lane left C. Change lane right

B. Change lane left — stopped vehicle ahead at 35m, sides also clear. Confidence: HIGH.

Figure 12: CAV-side barrier example.

E.2.1 EMERGENCY VEHICLE



CAV ego_2: Examine your multi-view camera images. Is there any anomalous event, object, or special vehicle visible? If yes, identify the event type, its relative position, and how many vehicles appear blocked behind it. If no anomaly is visible, briefly describe the traffic conditions ahead that may be relevant for driving decisions. Reason step-by-step through the cameras before concluding; end with "Confidence: HIGH, LIKELY, or UNCERTAIN".

Yes. Emergency vehicle behind, same lane, approximately 14m away, visible in back camera. **Confidence: HIGH.**

CAV ego_2 (Speed: 12.5 m/s, current lane: 1, ~29m to stop line). Available actions: Keep lane, Change lane left, Change lane right. Based on your multi-view camera images, what is the optimal lane decision? [ANOMALY DETECTION] Yes. Emergency vehicle behind, same lane, approximately 14m away, visible in back camera. Confidence: HIGH.

Options: A. Keep lane B. Change lane left C. Change lane right

A. Keep lane

Figure 13: CAV-side emergency-vehicle example.

E.2.2 JAYWALKING PEDESTRIAN



CAV ego_0: Examine your multi-view camera images. Is there any anomalous event, object, or special vehicle visible? If yes, identify the event type, its relative position, and how many vehicles appear blocked behind it. If no anomaly is visible, briefly describe the traffic conditions ahead that may be relevant for driving decisions. Reason step-by-step through the cameras before concluding; end with "Confidence: HIGH, LIKELY, or UNCERTAIN".

Yes. Crossing pedestrian ahead, on the road, approximately 24m away, visible in front camera — blocks my path. **Confidence: HIGH.**

CAV ego_0 (Speed: 13.8 m/s, current lane: 1, ~34m to stop line). Available actions: Keep lane, Change lane left. Based on your multi-view camera images, what is the optimal lane decision?

Options: A. Keep lane **B. Change lane left** C. Change lane right

<think>Peer (ITS) reports lane 2 least congested on my approach, I am in lane 1. Changing lane left. Confidence: HIGH.</think> **B**

Figure 14: CAV-side jaywalking-pedestrian example.