# PARETO-CONSISTENT PREFERENCE LEARNING FOR TRAFFIC SIGNAL CONTROL

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Urban traffic signal control is inherently multi-objective, yet traffic quality is not directly available as a well-defined reward vector. Instead, it must be inferred from competing criteria such as efficiency, safety, fairness, and stability, and the preferred ordering of traffic states can reverse as preferences change. Methods based on preference-independent scalar rewards or learned scores reduce these competing criteria to a single implicit compromise. We formulate traffic signal control as multi-objective preference learning under the risk of representation collapse. Our method learns a structured vector-valued traffic-quality representation and a unified preference-conditioned policy family, rather than a preference-independent scalar score and a fixed controller. To supervise this representation, we construct deterministic, schema-constrained traffic abstractions and combine rule-based objective-specific comparisons with preference-conditioned comparisons generated offline by an LLM. We further encourage dominance consistency to preserve Pareto-relevant relational structure. Based on the learned representation, we train a single controller that can realize diverse preference-conditioned trade-offs without retraining for each preference or querying the LLM during deployment. Results on CityFlow benchmarks across multiple urban networks, unseen preferences, and shifted traffic conditions indicate competitive traffic performance and improvements in ordering preservation, preference alignment, and empirical trade-off quality within the evaluated settings.

## 1 INTRODUCTION

Traffic signal control must adapt to operating priorities that change across scenarios. Peak-hour control may emphasize throughput and delay reduction, while school zones, incidents, adverse weather, or persistent congestion may require safer, smoother, fairer, or more stable behavior. These objectives can induce different judgments over the same traffic conditions: a state with lower delay may be less safe or less fair, while a conservative state may sacrifice efficiency for risk reduction or service balance. Thus, traffic quality is not naturally described by a single fixed ordering, but by preference-dependent comparisons among competing objectives.

This creates a fundamental difficulty for scalar reward learning. Consider two traffic states or policies: one is more efficient but less safe or less fair, while the other improves safety or fairness at the cost of higher delay. Under an efficiency-oriented preference, the former should be ranked higher; under a safety- or fairness-oriented preference, the ranking may reverse. A preference-independent scalar score cannot faithfully represent both orderings at the same time. When multi-objective supervision is absorbed into one fixed scalar reward or learned quality score, these preference-dependent relations collapse into a fixed compromise. The key issue is therefore *representation collapse*: scalar scoring can erase the relational structure needed to represent multi-objective traffic quality.

We address this problem by learning a structured traffic-quality representation rather than a single scalar score. Our key object is a vector-valued representation $q_\phi(o)$, whose components correspond to traffic-quality dimensions such as efficiency, safety, fairness, and stability. Objective-specific comparisons ground each head in its intended semantic dimension; preference-conditioned comparisons teach how these dimensions compose under different priorities; and dominance consistency encourages basic Pareto-relevant orderings not to be reversed. The scalar quantity $w^\top q_\phi(o)$ is a query of the learned vector for a given preference $w$, not a preference-independent learned score.



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.

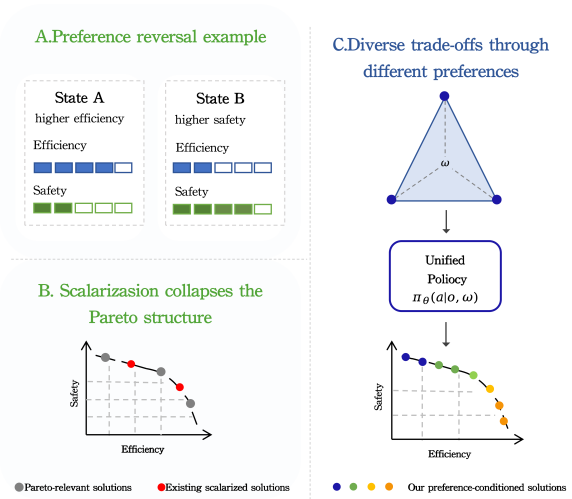


Figure 1: Original conceptual illustration: **A**, preference reversal; **B**, one fixed, preference-independent scalar compromise; **C**, a preference-conditioned policy family. The plotted points are schematic, not measured Pareto fronts.

On top of this representation, we train a preference-conditioned controller that can realize different trade-offs without rebuilding the policy for each preference.

Figure 1 illustrates preference reversal and a shared policy family. Its scalar panel concerns only a fixed, preference-independent score; a conditioned scalar baseline is evaluated separately.

Existing work touches parts of this problem but usually avoids the representation question. RL and MARL methods for traffic signal control typically optimize fixed scalar rewards, even when they improve coordination, scale, or state representation, as exemplified by phase-competition, pressure-based, and graph-coordination approaches (Zheng et al., 2019; Wei et al., 2019a;b; Chen et al., 2020; Zhang et al., 2022). Multi-objective traffic studies often enrich the metric set or compare several hand-designed trade-offs, but may still rely on scalarized objectives or sparse operating points. Preference-conditioned MORL methods are closer in spirit, and recent work studies reusable policies over preferences, Pareto-conditioned networks, Pareto-front recovery, and controllable preference-conditioned optimization (Yang et al., 2019; Reymond et al., 2022; Lu et al., 2023; Liu et al., 2025; Yang et al., 2025). Related ideas also appear in multi-reward alignment and controllable preference adaptation, where one model is expected to realize diverse trade-offs across preferences (Ramé et al., 2023; Yang et al., 2024; Zhong et al., 2024). Many of these methods assume that a reward vector or objective structure is already available. In traffic control, by contrast, the traffic-quality representation itself must be inferred from conflicting and preference-dependent supervision, and the central challenge is to prevent that representation from collapsing into a disguised scalar score.

Recent LLM-based work also provides useful context, but in a different role. Some methods place LLMs directly in the control loop as online traffic controllers or coordinators (Lai et al., 2025; Yuan et al., 2025; Pang et al., 2024), while others use language models for reward or objective design (Kwon et al., 2023). More broadly, comparison-based supervision is well established in preference learning and RLHF-style pipelines (Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022), and recent work has studied the reliability and bias of LLM-based judges (Zheng et al., 2023; Zhu et al., 2025). Our method uses a different division of labor. We do not place the LLM in the control loop, and we do not use it as a generic reward generator. Instead, we use it as an *offline pairwise supervision interface* for preference-conditioned comparisons; metric- or rule-based comparisons provide the objective-specific labels in the main Hybrid protocol.

We formulate traffic signal control as multi-objective preference learning under representation collapse risk. Our framework converts raw simulator observations into deterministic, schema-constrained traffic abstractions and obtains offline pairwise supervision to learn a structured traffic-

quality representation. The learned representation $q_\phi(o)$ is trained with objective fidelity, preference compositionality, and dominance-preservation objectives, and is then used to optimize a preference-conditioned policy family. Our evaluation therefore includes not only conventional traffic metrics, but also representation diagnostics, preference reversals, dominance preservation, Pareto-aware metrics, and preference-response alignment.

**Our contributions are:**

- **Representation-centered formulation.** We identify scalar representation collapse as a key obstacle in multi-objective traffic signal control, where preference-dependent ranking reversals cannot be captured by a single preference-independent score.
- **Structured traffic-quality representation.** We learn a vector-valued representation $q_\phi(o)$ constrained by objective-specific supervision, preference-conditioned supervision, and dominance consistency, with $w^\top q_\phi(o)$ serving as a preference query rather than a collapsed scalar reward.
- **Offline LLM-assisted supervision.** We use deterministic traffic abstractions and offline pairwise comparisons for multi-objective preference learning, using rule-based objective labels and LLM preference-conditioned labels in the main Hybrid protocol.
- **Preference-conditioned control and diagnostics.** We train a unified policy family on top of $q_\phi(o)$ and evaluate it through traffic metrics, empirical Pareto-aware metrics, preference-response alignment, unseen-preference generalization, and representation-level diagnostics.

## 2 Related Work

**RL/MARL-based traffic signal control under fixed scalar rewards.** RL and MARL have been widely applied to traffic signal control, with representative methods such as PressLight (Wei et al., 2019a), CoLight (Wei et al., 2019b), and MPLight (Chen et al., 2020). These methods improve coordination, scalability, and network-level control under fixed scalar objectives such as pressure, delay, queue length, or throughput. As a result, they optimize one predefined compromise rather than explicitly represent trade-offs among efficiency, safety, fairness, and stability.

**Multi-objective traffic control with richer traffic metrics.** A related line enriches traffic control with more traffic metrics or compares a small number of manually chosen trade-off settings, and may further strengthen traffic-state representations (Zhang et al., 2022). Some such settings rely on scalarized objectives or sparse trade-off comparisons. In contrast, our focus is not merely to add more metrics, but to avoid collapsing comparative multi-objective supervision into a preference-independent scalar score.

**Preference-conditioned RL and multi-objective reinforcement learning.** MORL studies Pareto-optimal trade-offs instead of a single optimum. Existing work has explored reusable policies over preferences (Yang et al., 2019), Pareto-conditioned policy families (Reymond et al., 2022), Pareto-front recovery (Lu et al., 2023; Liu et al., 2025), and controllable preference-conditioned optimization (Yang et al., 2025). Similar ideas also appear in controllable alignment and multi-reward modeling (Ramé et al., 2023; Yang et al., 2024; Zhong et al., 2024). Many of these methods assume the reward vector or objective structure is given, whereas our problem is to learn the traffic-quality representation itself from comparative supervision. A preference-conditioned scalar model can express reversals; our question is whether objective-wise and dominance relations are better preserved by an explicitly structured representation.

**LLMs as controller, reward designer, or supervision source.** LLMs have recently been used as online traffic controllers or coordinators (Lai et al., 2025; Yuan et al., 2025; Pang et al., 2024), as reward or objective designers (Kwon et al., 2023), and as pairwise judges in preference-learning pipelines (Christiano et al., 2017; Ouyang et al., 2022). Recent work also studies the reliability of LLM-based judges (Zheng et al., 2023; Zhu et al., 2025). Our setting uses the LLM as an *offline pairwise supervision interface* for preference-conditioned labels; objective-specific labels in the main protocol come from rules, and control is carried out without online LLM involvement.

# 3 PRELIMINARIES

We do not consider the standard MORL setting in which a clean reward vector is given. Instead, traffic quality is only indirectly reflected by heterogeneous signals, and the relative preference between two traffic states may depend on the desired trade-off.

Formally, let $o \in \mathcal{O}$ denote a traffic observation, and let

$$\mathbf{z}^\star(o) \in \mathbb{R}^K \tag{1}$$

denote an unknown latent multi-objective traffic-quality representation. To select among trade-offs, we introduce a preference vector

$$w \in \Delta^K, \qquad \Delta^K = \left\{ w \in \mathbb{R}^K_{\geq 0} \;\middle|\; \sum_{k=1}^{K} w_k = 1 \right\}, \tag{2}$$

and the corresponding preference-conditioned utility

$$u^\star(o; w) = w^\top \mathbf{z}^\star(o). \tag{3}$$

A key difficulty is that the ordering between two states may reverse under different preferences. For two observations $o_i, o_j$, there may exist $w_1, w_2 \in \Delta^K$ such that

$$u^\star(o_i; w_1) > u^\star(o_j; w_1), \qquad u^\star(o_i; w_2) < u^\star(o_j; w_2). \tag{4}$$

This preference-dependent ranking reversal is what makes a fixed, preference-independent scalar score insufficient in general; it does not rule out preference-conditioned scalar models.

We also use the standard notion of Pareto dominance: for $\mathbf{u}, \mathbf{v} \in \mathbb{R}^K$, $\mathbf{u} \succ \mathbf{v}$ if $u_k \geq v_k$ for all $k$ and $u_k > v_k$ for at least one $k$. Throughout the paper, *Pareto consistency* refers only to a weak relational notion: preserving objective-wise orderings, preference-conditioned orderings, and basic dominance relations, rather than exact Pareto front recovery.

Finally, our method learns from pairwise comparisons. For a pair $(o_i, o_j)$, we use objective-specific labels $y_{ij}^{(k)} \in \{-1, 0, 1\}$ and preference-conditioned labels $y_{ij}^{(w)} \in \{-1, 0, 1\}$, which are used to learn a structured multi-objective representation and, on top of it, a preference-conditioned policy family $\pi(a \mid o, w)$.

# 4 METHOD

## 4.1 OVERVIEW

Our problem is not standard MORL with known reward vectors. In traffic signal control, traffic quality is only indirectly reflected by heterogeneous signals such as delay, queue imbalance, TTC, and stability proxies, and the relative ranking between two traffic states may reverse under different preferences. If such supervision is absorbed into a single preference-independent scalar score, the underlying multi-objective structure collapses into a pseudo-total score. We refer to this failure mode as *representation collapse*.

To address this, we first learn a *structured traffic-quality representation*

$$\mathbf{q}_\phi(o) = \left[ q_\phi^{(1)}(o),\, q_\phi^{(2)}(o),\, \ldots,\, q_\phi^{(K)}(o) \right] \in \mathbb{R}^K, \tag{5}$$

rather than a scalar reward. Given a preference vector $w \in \Delta^K$, we query this representation through

$$s_\phi(o; w) = w^\top \mathbf{q}_\phi(o), \tag{6}$$

where $s_\phi(o; w)$ is a preference-dependent query of the learned vector, not the primary learning object. On top of $\mathbf{q}_\phi(o)$, we then learn a preference-conditioned policy family

$$\pi_\theta(a \mid o, w), \tag{7}$$

so that one shared controller can realize different traffic behaviors under different preferences.

As shown in Figure 2, the framework has four stages: (1) structured traffic abstraction, (2) pairwise multi-objective supervision, (3) structured representation learning, and (4) preference-conditioned policy learning.

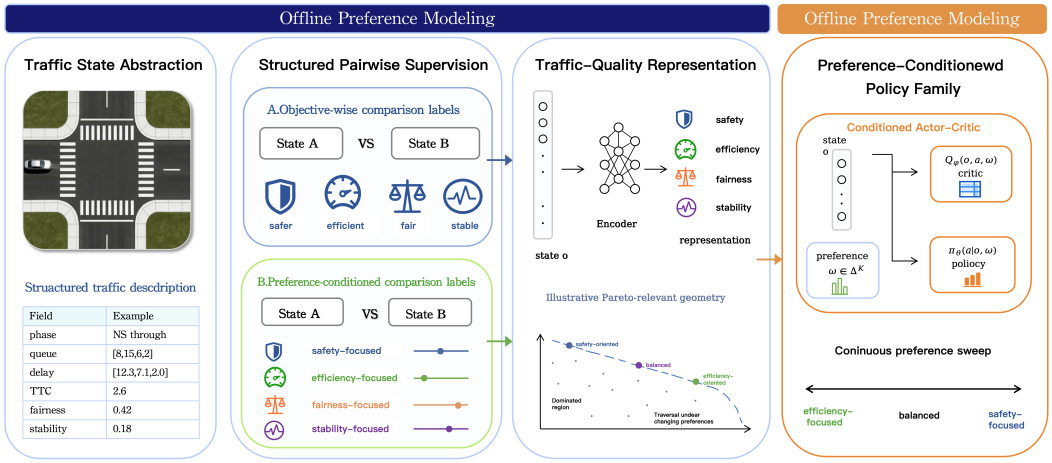


Figure 2: Conceptual overview. In the main Hybrid protocol, objective-specific labels are rule-based and preference-conditioned labels come from an offline LLM. Objective-specific LLM queries occur only in the LLM-only ablation; deployment makes no LLM calls. The trade-off curve is illustrative.

### 4.2 Structured Abstraction and Pairwise Supervision

We first map each raw observation $o$ to a deterministic structured abstraction

$$x = T(o; \Xi), \tag{8}$$

where $T(\cdot; \Xi)$ is defined by a fixed schema $\Xi$. This abstraction makes traffic-quality factors such as efficiency, safety, fairness, and stability explicitly comparable. For supervision, we further serialize it into a schema-constrained text form

$$c = \mathcal{C}(x; \Xi). \tag{9}$$

From a buffer of observations, we construct candidate state pairs

$$\mathcal{P} = \{(o_i, o_j)\},$$

prioritizing pairs with visible contrast in at least one quality dimension. We then collect two complementary forms of pairwise supervision.

For each pair $(o_i, o_j)$ and each objective $k$, we obtain an objective-specific label

$$y_{ij}^{(k)} \in \{-1, 0, 1\}, \tag{10}$$

which indicates which state is better with respect to objective $k$. In the main Hybrid protocol, these labels are obtained from metric- or rule-based comparisons. LLM-generated objective-specific labels are used only in the LLM-only ablation.

For each pair $(o_i, o_j)$ and each sampled preference $w \in \mathcal{W}_{\text{train}} \subset \Delta^K$, we obtain a preference-conditioned label

$$y_{ij}^{(w)} \in \{-1, 0, 1\}, \tag{11}$$

which indicates which state is preferred under trade-off $w$. Here, $w$ is externally specified by the training protocol. In the core protocol, these labels are provided by the LLM as an *offline comparative supervision interface*, which is particularly useful when conflicting objectives induce preference-dependent ranking reversals that are difficult to encode manually.

Objective-specific labels tell the model what each dimension should mean, while preference-conditioned labels tell the model how these dimensions should interact under different trade-offs. Both label types have distinct roles in the proposed objective; their individual contributions are examined by ablation.

### 4.3 Structured Traffic-Quality Representation

The central object of our method is the structured representation $\mathbf{q}_\phi(o)$. Its four outputs alone do not establish distinct semantics. Instead, it is jointly constrained by three components, whose contributions are evaluated separately.

**Objective fidelity.** For a pair $(o_i, o_j)$ and objective $k$, define

$$\Delta_{ij}^{(k)} = q_\phi^{(k)}(o_i) - q_\phi^{(k)}(o_j). \tag{12}$$

We supervise each head with

$$\mathcal{L}_{\text{obj}} = \sum_{k=1}^{K} \mathbb{E}_{(i,j)} \big[\ell(\Delta_{ij}^{(k)}, y_{ij}^{(k)})\big]. \tag{13}$$

**Preference compositionality.** Given preference $w$, we query the representation through

$$s_\phi(o; w) = w^\top \mathbf{q}_\phi(o). \tag{14}$$

For a pair $(o_i, o_j)$, define

$$\Delta_{ij}^{(w)} = s_\phi(o_i; w) - s_\phi(o_j; w). \tag{15}$$

We then match preference-conditioned supervision via

$$\mathcal{L}_{\text{pref}} = \mathbb{E}_{(i,j,w)} \big[\ell(\Delta_{ij}^{(w)}, y_{ij}^{(w)})\big]. \tag{16}$$

This is where the model learns trade-off-sensitive composition and preference-dependent ranking reversal. Importantly, scalarization in this loss is a preference-dependent query over the learned representation, not a preference-independent score.

**Dominance consistency.** To encourage preservation of basic Pareto ordering on sampled comparisons, we construct pseudo-dominance pairs

$$\mathcal{D}_{\text{dom}} = \Big\{(i,j) \;\Big|\; y_{ij}^{(k)} \geq 0,\ \forall k,\ \text{and } \exists k:\ y_{ij}^{(k)} = 1\Big\}, \tag{17}$$

and regularize them with

$$\mathcal{L}_{\text{dom}} = \mathbb{E}_{(i,j)\in\mathcal{D}_{\text{dom}}} \left[\max\Big(0,\ m - \min_k \big(q_\phi^{(k)}(o_i) - q_\phi^{(k)}(o_j)\big)\Big)\right]. \tag{18}$$

For $m > 0$, this surrogate encourages a margin in every coordinate, including coordinates labeled as ties; it is stronger than the weak-dominance definition on those pairs. It does not imply exact Pareto-front recovery or Pareto-optimal policies.

The full representation objective is

$$\mathcal{L}_{\text{repr}} = \lambda_{\text{obj}}\mathcal{L}_{\text{obj}} + \lambda_{\text{pref}}\mathcal{L}_{\text{pref}} + \lambda_{\text{dom}}\mathcal{L}_{\text{dom}}. \tag{19}$$

### 4.4 Preference-Conditioned Policy Family

On top of the learned representation, we train a preference-conditioned policy family

$$\pi_\theta(a \mid o, w), \tag{20}$$

optionally paired with a conditioned critic

$$Q_\psi(o, a, w). \tag{21}$$

Given a preference $w$, we query the representation through $s_\phi(o; w) = w^\top \mathbf{q}_\phi(o)$ and define

$$r_{\text{mix}}(o_t, a_t, o_{t+1}; w) = r_{\text{ext}}(o_t, a_t, o_{t+1}) + \lambda\, s_\phi(o_{t+1}; w). \tag{22}$$

This is only a policy-learning interface: the learned object remains $\mathbf{q}_\phi(o)$, while (22) specifies how a chosen preference reads that structure for control. The same preference query also appears in the

comparison loss (16); neither use replaces the learned vector with a preference-independent scalar representation.

Let $p(w)$ denote the training-time preference distribution induced by five anchor templates and local sampling (Appendix A.8). We train the policy over this distribution:

$$\max_{\theta} \ \mathbb{E}_{w\sim p(w)} \left[ J(\pi_\theta(\cdot \mid w), w) \right], \tag{23}$$

where

$$J(\pi_\theta(\cdot \mid w), w) = \mathbb{E}_{\pi_\theta(\cdot|w)} \left[ \sum_{t=0}^{\infty} \gamma^t \, r_{\text{mix}}(o_t, a_t, o_{t+1}; w) \right]. \tag{24}$$

At deployment time, the controller receives the current observation $o$ and a desired preference $w$, then outputs an action directly through $\pi_\theta(a \mid o, w)$. By sharing parameters across many preferences, the model learns one reusable policy family rather than isolated point solutions. Optionally, continuity over nearby preferences can be further encouraged via a policy-level smoothness regularizer, but this is not required by the core method.

# 5 Experiments

## 5.1 Setup

We evaluate the proposed method on CityFlow using three standard traffic benchmarks: `Jinan`, `Hangzhou`, and `New York`, which represent small-, medium-, and large-scale networks. We consider four traffic-quality dimensions throughout: *efficiency*, *safety*, *fairness*, and *stability*. Efficiency is measured by ATT, AWT, and AQL; safety by TTC percentiles and harsh braking proxies; fairness by waiting-time imbalance; and stability by oscillation-related phase switching. These simulation-based safety proxies do not measure real-world collisions. ATT and AWT are in seconds and AQL counts queued vehicles; all three are minimized. TTC-P10 (seconds) and normalized Fair are maximized, whereas harsh-braking rate, phase-switch rate, and Oscillation are minimized (Appendix A.8).

For preference-conditioned learning, we define a training preference set $\mathcal{W}_{\text{train}} \subset \Delta^K$ using five templates: efficiency-focused, safety-focused, fairness-focused, stability-focused, and balanced. Each objective-focused anchor assigns $0.55$ to its dominant objective and $0.15$ to each other objective; the balanced anchor assigns $0.25$ to each. The training distribution samples each coordinate within $\pm 0.05$ of an anchor, clips to nonnegative values, and renormalizes. At evaluation time, we distinguish seen anchors, interpolated preferences, and held-out preferences used only for evaluation. Unless otherwise stated, the default supervision pipeline is *Hybrid*: objective-specific labels are rule-based, while preference-conditioned labels are obtained from the LLM offline.

We compare against three classes of baselines: strong traffic-control baselines (MaxPressure, Advanced-CoLight), scalar reward-learning baselines (C2T-scalar), and preference-aware baselines (Cond-Scalar-RL, Weighted-RL). Here, Cond-Scalar-RL takes the preference vector $w$ as input but directly optimizes a scalarized reward, while Weighted-RL trains one controller per preference template and treats the resulting discrete set as a policy-family approximation. Each trained policy receives 1.0 million environment steps, so the five-controller Weighted-RL set uses 5.0 million in total, versus 1.0 million for our shared controller. Advanced-CoLight denotes Advanced-CoLight (Zhang et al., 2022); C2T-scalar adapts the learned scalar score of C$^2$T (Chen et al., 2026).

In the main paper, we focus on three types of evidence: conventional traffic performance, representation diagnostics, and trade-off controllability. Full conventional tables, label-source study, unseen-preference generalization, full ablations, additional Pareto visualizations, metric definitions, and LLM annotation details are reported in the appendix. The main tables give reported point estimates; five-seed paired analyses are reported for selected principal comparisons in Appendix A.12.

## 5.2 Conventional Traffic Performance

We first verify that the proposed method remains competitive as a practical traffic controller. The full conventional traffic table is reported in the appendix.

Table 1: Representation diagnostics on the reported held-out comparison pairs. Obj.Acc tests head semantics, Pref.Acc tests preference-conditioned comparison, and DPR measures dominance preservation. Strict paired-reversal accuracy requires predictions under both opposing preferences for each pair and is not reported without pair-level records (Appendix A.8).

| Method | Obj.Acc↑ | Pref.Acc↑ | DPR↑ |
|---|---|---|---|
| scalar head | 0.8124 | 0.7413 | 0.7364 |
| Cond-Scalar-RL | 0.8347 | 0.7688 | 0.7749 |
| w/o $\mathcal{L}_{\mathrm{obj}}$ | 0.8615 | 0.8564 | 0.8786 |
| w/o $\mathcal{L}_{\mathrm{pref}}$ | 0.9072 | 0.8126 | 0.8914 |
| w/o $\mathcal{L}_{\mathrm{dom}}$ | 0.9041 | 0.8728 | 0.8365 |
| **Ours** | 0.9238 | 0.8897 | 0.9148 |

**Results.** The reported conventional point estimates and paired-seed comparisons show that our method remains competitive on standard traffic metrics while learning a richer representation (Appendix A.12). Strong fixed-objective baselines improve efficiency, but are less balanced on the reported safety, fairness, and stability proxies. C2T-SCALAR is already a strong single-point baseline, confirming the value of structured supervision, but its gains remain concentrated around one implicit compromise. Compared with the stronger preference-aware baselines COND-SCALAR-RL and WEIGHTED-RL, our method does not rely on large gains in any one scalar metric; instead, the reported point estimates are competitive on efficiency and favorable on safety, fairness, and oscillation. The paired comparisons do not distinguish the ATT differences on the three networks from zero. This pattern is consistent with the intended multi-objective representation, although these controller metrics alone cannot establish its mechanism. In other words, the conventional results mainly establish that the proposed method remains a viable traffic controller, while the more distinctive evidence for our claim comes from the representation- and trade-off-level experiments below.

### 5.3 REPRESENTATION VALIDATION

Our main claim is not just that the controller performs well, but that it avoids collapsing multi-objective supervision into a disguised scalar score. We therefore directly test whether the learned representation preserves objective semantics, preference-dependent rankings, and dominance structure.

**Results.** Table 1 is the core mechanism-verification result. A fixed scalar head may still achieve reasonable traffic performance, but cannot represent opposite rankings of the same pair under two preferences. The preference-conditioned COND-SCALAR-RL baseline can in principle represent such reversals; its reported point estimates on Pref.Acc and DPR are lower than those of the full model. These diagnostics test whether the learned representation preserves relations that a fixed scalar score loses.

By contrast, the full model improves all three reported diagnostics simultaneously. The ablated variants behave consistently with their intended roles: removing $\mathcal{L}_{\mathrm{obj}}$ weakens objective semantics, removing $\mathcal{L}_{\mathrm{pref}}$ most strongly hurts preference-conditioned comparison, and removing $\mathcal{L}_{\mathrm{dom}}$ mainly reduces dominance preservation. Together, the ablation point estimates support the roles of semantic, compositional, and relational supervision. The architecture-matched objective-only multi-head control in Appendix A.12 helps separate these constraints from multi-head capacity alone. Additional ablations and leakage analysis are reported in the appendix.

### 5.4 TRADE-OFF QUALITY AND PREFERENCE RESPONSE

We next test whether the learned representation supports a better empirical trade-off set and whether the resulting controller responds to preference changes at evaluated queries.

**Results.** Table 2 shows that the ranking differs from conventional single-point traffic performance. Although WEIGHTED-RL is already strong as a discrete policy-set approximation, our method has the best reported HV and Coverage point estimates, the lowest dominance violation estimate, and fa-

Table 2: Empirical trade-off quality under a common finite evaluation preference set. HV and Coverage summarize the evaluated outcomes, Dom.-Viol. measures observed structural inconsistency, and Uti/Align evaluate preference-conditioned usefulness. These metrics do not certify recovery of the true Pareto front.

| Method | HV↑ | Coverage↑ | Dom.-Viol.↓ | Uti↑ | Align↓ |
|---|---|---|---|---|---|
| Advanced-Co | 0.4853 | 0.3916 | 0.1384 | 0.6038 | 0.2476 |
| C2T-scalar | 0.5781 | 0.5124 | 0.1139 | 0.6521 | 0.2094 |
| Cond-Scalar-RL | 0.6417 | 0.6632 | 0.0924 | 0.6845 | 0.1718 |
| Weighted-RL | 0.6816 | 0.7368 | 0.0704 | 0.7057 | 0.1486 |
| **Ours** | 0.7198 | 0.8015 | 0.0468 | 0.7288 | 0.1127 |

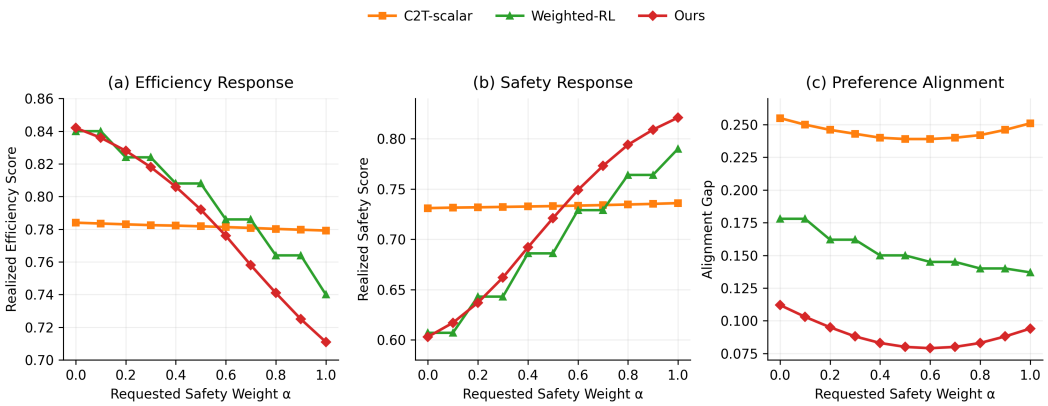


Figure 3: Preference–response alignment under an efficiency–safety preference sweep. **(a)** Realized efficiency decreases smoothly as the requested safety weight increases. **(b)** Realized safety increases accordingly. **(c)** Our method maintains the lowest alignment gap across the sweep. C2T-SCALAR shows weak response to preference changes, whereas WEIGHTED-RL exhibits the expected stepwise behavior of discrete policy switching.

vorable Uti/Align estimates. The contrast with C2T-SCALAR and COND-SCALAR-RL is especially informative: even when these baselines remain competitive on standard traffic metrics, they have lower reported coverage over the evaluated preferences and higher dominance violation estimates. The conditioned scalar baseline is not logically subject to the fixed-scalar reversal impossibility.

Fig. 3 shows the sampled response: ours trades efficiency for safety, while C2T-SCALAR barely responds and WEIGHTED-RL switches stepwise. Paired-seed summaries, template sensitivity, and the 100-pair supervision setting are given in Appendices A.12, A.9, and A.10.

## 6 CONCLUSION

We studied traffic signal control as multi-objective preference learning under representation collapse risk. Rather than a fixed scalar reward and controller, we learn a structured quality vector and one preference-conditioned policy family. Rule-based objective comparisons, offline LLM preference comparisons, and dominance regularization encourage Pareto-relevant relations. CityFlow results indicate competitive traffic performance and improved ordering preservation, alignment, and empirical trade-off quality in the evaluated settings. These findings support a reusable trade-off-aware controller over one fixed scalar objective.

**Limitations.** The fixed four-objective schema and five training anchors leave omitted objectives and untested preferences uncovered. Offline LLM labels add prompt/version sensitivity and annotation cost despite LLM-free deployment. Dominance regularization does not guarantee Pareto-optimal policies. CityFlow TTC and braking are safety proxies, not collision evidence; deployment needs validation with imperfect sensors, other road users, traffic rules, failures, and rare hazards.

### AI USE STATEMENT

Generative AI was used to assist with manuscript writing and editing. The proposed method also uses an LLM offline to generate preference-conditioned comparison labels, as described in Section 4.2 and Appendix A.7. The authors are responsible for the scientific content and final manuscript.

## REFERENCES

Yuntao Bai et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

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

Yuyang Chen, Kaiyan Zhao, Yiming Wang, Ming Yang, Bin Rao, and Zhenning Li. $C^2T$: Captioning-structure and llm-aligned common-sense reward learning for traffic–vehicle coordination, 2026.

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.

Minae Kwon, Sang Michael Xie, Kalesha Bullard, and Dorsa Sadigh. Reward design with language models. In *The Eleventh International Conference on Learning Representations*, 2023.

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, Volume 1*, pp. 2335–2346, 2025.

Ruohong Liu, Yuxin Pan, Linjie Xu, Lei Song, Pengcheng You, Yize Chen, and Jiang Bian. Efficient discovery of pareto front for multi-objective reinforcement learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Haoye Lu, Daniel Herman, and Yaoliang Yu. Multi-objective reinforcement learning: Convexity, stationarity and pareto optimality. In *The Eleventh International Conference on Learning Representations*, 2023.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022.

Aoyu Pang, Maonan Wang, Man-On Pun, Chung Shue Chen, and Xi Xiong. illm-tsc: Integration reinforcement learning and large language model for traffic signal control policy improvement. *arXiv preprint arXiv:2407.06025*, 2024.

Alexandre Ramé, Guillaume Couairon, Mustafa Shukor, Corentin Dancette, Jean-Baptiste Gaya, Laure Soulier, and Matthieu Cord. Rewarded soups: Towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. In *Advances in Neural Information Processing Systems*, 2023.

Mathieu Reymond, Eugenio Bargiacchi, and Ann Nowé. Pareto conditioned networks. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, 2022.

Hua Wei, Chacha Chen, Guanjie Zheng, Kan Wu, Vikash V. 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*, pp. 1290–1298, 2019a.

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*, pp. 1913–1922, 2019b.

Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 56276–56297, 2024.

Runzhe Yang, Xingyuan Sun, and Karthik Narasimhan. A generalized algorithm for multi-objective reinforcement learning and policy adaptation. In *Advances in Neural Information Processing Systems*, pp. 14610–14621, 2019.

Yucheng Yang, Tianyi Zhou, Mykola Pechenizkiy, and Meng Fang. Preference controllable reinforcement learning with advanced multi-objective optimization. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 71612–71647, 2025.

Zirui Yuan, Siqi Lai, and Hao Liu. Collmlight: Cooperative large language model agents for network-wide traffic signal control. *arXiv preprint arXiv:2503.11739*, 2025.

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*, pp. 26645–26654, 2022.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023.

Yifan Zhong, Chengdong Ma, Xiaoyuan Zhang, Ziran Yang, Haojun Chen, Qingfu Zhang, Siyuan Qi, and Yaodong Yang. Panacea: Pareto alignment via preference adaptation for llms. In *Advances in Neural Information Processing Systems*, 2024.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges. In *The Thirteenth International Conference on Learning Representations*, 2025.

# A Extended Experimental Analysis and Protocol

This appendix provides a detailed breakdown of the experimental analyses summarized in the main paper. Our goal is to examine conventional traffic performance alongside the central mechanism question: whether the learned traffic-quality representation preserves preference-dependent ordering, objective semantics, and empirical trade-off structure.

We organize the appendix into twelve parts: (A.1) full conventional traffic-control results, (A.2) label-source study and human/expert audit, (A.3) extended representation diagnostics and leakage analysis, (A.4) additional Pareto visualization and controllability analysis, (A.5) unseen-preference and traffic-shift generalization, (A.6) full ablations, (A.7) detailed LLM annotation protocol, (A.8) metric definitions and implementation details, (A.9) preference-template sensitivity and the human label audit, (A.10) offline supervision budget and schema sensitivity, (A.11) annotation-model choice and training cost, and (A.12) paired statistics and ranking-reversal analyses.

Unless otherwise stated, all experiments follow the same simulator setup, PPO training recipe, random seeds, and evaluation pipeline used in the main paper. The original point estimates are retained in the full conventional table. The paired-seed summaries and additional analyses reported in the author response are collected in Appendices A.9–A.12.

## A.1 Full Conventional Traffic-Control Results

The main paper intentionally compresses conventional traffic performance to avoid letting standard controller metrics dominate the narrative. Nevertheless, it is still necessary to verify that the proposed method remains practically useful as a traffic controller. Table 3 reports the full conventional performance table across the three benchmark networks.

**Interpretation.** The role of Table 3 is confirmatory rather than central. It shows that the proposed method does not sacrifice ordinary traffic performance in order to obtain a richer representation. At the same time, it should not be over-interpreted: the core contribution of the paper is not that our method wins every conventional metric by a large margin, but that it learns a better structured multi-objective representation and therefore supports a more faithful and controllable trade-off family.

Table 3: **Full conventional traffic-control results.** Lower is better for ATT, AWT, AQL, Harsh, and Oscillation. Higher is better for TTC-P10 and Fairness. Values are originally reported point estimates; uncertainty is not shown.

**Jinan**

| Method | ATT↓ | AWT↓ | AQL↓ | TTC-P10↑ | Harsh↓ | Fair↑ | Osc.↓ |
|---|---|---|---|---|---|---|---|
| Random | 81.2437 | 33.8142 | 20.4518 | 1.4216 | 0.4127 | 0.6714 | 1.2385 |
| FixedTime | 69.3385 | 28.4067 | 17.3829 | 1.6084 | 0.3516 | 0.7028 | 1.0741 |
| MaxPressure | 58.6124 | 20.9871 | 13.9046 | 1.9365 | 0.3018 | 0.7332 | 1.1164 |
| PressLight | 57.4831 | 20.1148 | 13.2117 | 1.9014 | 0.2871 | 0.7219 | 1.0517 |
| MPLight | 56.2746 | 19.4723 | 12.8821 | 1.9486 | 0.2748 | 0.7362 | 0.9964 |
| CoLight | 55.7832 | 19.1086 | 12.7035 | 1.9841 | 0.2634 | 0.7488 | 0.9637 |
| Advanced-Co | 54.8927 | 18.7715 | 12.5819 | 2.0063 | 0.2516 | 0.7543 | 0.9448 |
| C2T-scalar | 54.5500 | 18.6500 | 12.4500 | 2.2000 | 0.2050 | 0.7794 | 0.8873 |
| Cond-Scalar-RL | 53.1842 | 18.0745 | 11.8721 | 2.1684 | 0.1987 | 0.7968 | 0.8542 |
| Weighted-RL | 52.9476 | 17.9241 | 11.6284 | 2.1573 | 0.1968 | 0.8014 | 0.8461 |
| **Ours** | 52.7142 | 17.8126 | 11.5173 | 2.1964 | 0.1927 | 0.8158 | 0.8219 |

**Hangzhou**

| Method | ATT↓ | AWT↓ | AQL↓ | TTC-P10↑ | Harsh↓ | Fair↑ | Osc.↓ |
|---|---|---|---|---|---|---|---|
| Random | 93.4261 | 37.5082 | 22.1844 | 1.3962 | 0.4591 | 0.6582 | 1.3026 |
| FixedTime | 79.2648 | 31.8125 | 18.9217 | 1.5773 | 0.3847 | 0.6921 | 1.1443 |
| MaxPressure | 67.4419 | 23.2786 | 15.6324 | 1.9128 | 0.3325 | 0.7147 | 1.1683 |
| PressLight | 66.5823 | 22.8954 | 15.3028 | 1.9345 | 0.3191 | 0.7036 | 1.1218 |
| MPLight | 65.9137 | 22.2213 | 14.9826 | 1.9654 | 0.3027 | 0.7194 | 1.0661 |
| CoLight | 64.8015 | 21.8441 | 14.6733 | 1.9982 | 0.2895 | 0.7318 | 1.0195 |
| Advanced-Co | 63.9642 | 21.4636 | 14.5081 | 2.0227 | 0.2764 | 0.7412 | 0.9824 |
| C2T-scalar | 63.7500 | 21.5500 | 14.4000 | 2.1100 | 0.2350 | 0.7681 | 0.9248 |
| Cond-Scalar-RL | 62.0416 | 20.9913 | 13.8215 | 2.1129 | 0.2274 | 0.7861 | 0.8756 |
| Weighted-RL | 61.8943 | 20.9417 | 13.7068 | 2.1085 | 0.2258 | 0.7913 | 0.8694 |
| **Ours** | 61.6235 | 20.8031 | 13.5481 | 2.1267 | 0.2197 | 0.8064 | 0.8426 |

**New York**

| Method | ATT↓ | AWT↓ | AQL↓ | TTC-P10↑ | Harsh↓ | Fair↑ | Osc.↓ |
|---|---|---|---|---|---|---|---|
| Random | 129.7534 | 52.6149 | 30.8227 | 1.2163 | 0.5418 | 0.6112 | 1.4264 |
| FixedTime | 114.3221 | 45.7085 | 26.9918 | 1.4386 | 0.4724 | 0.6489 | 1.2815 |
| MaxPressure | 97.6125 | 38.4418 | 22.9043 | 1.7821 | 0.3562 | 0.6828 | 1.1974 |
| PressLight | 96.4083 | 37.9127 | 22.3016 | 1.8043 | 0.3385 | 0.6715 | 1.1537 |
| MPLight | 94.6719 | 36.8824 | 21.9048 | 1.8427 | 0.3216 | 0.6897 | 1.1049 |
| CoLight | 92.1143 | 35.7772 | 21.2086 | 1.8724 | 0.3097 | 0.7031 | 1.0728 |
| Advanced-Co | 90.3368 | 34.9817 | 20.8445 | 1.9038 | 0.2974 | 0.7189 | 1.0316 |
| C2T-scalar | 87.9000 | 34.5000 | 20.3000 | 1.9800 | 0.2700 | 0.7426 | 0.9481 |
| Cond-Scalar-RL | 86.5347 | 33.9624 | 19.9316 | 2.0017 | 0.2613 | 0.7708 | 0.9247 |
| Weighted-RL | 86.2715 | 33.7482 | 19.8427 | 1.9984 | 0.2595 | 0.7736 | 0.9168 |
| **Ours** | 85.9348 | 33.5124 | 19.6743 | 2.0128 | 0.2528 | 0.7894 | 0.8921 |

### A.2 Label-source Study and Human/Expert Audit

A natural concern for our formulation is whether the LLM is actually needed, and if so, for which part of the supervision pipeline. Because the structured traffic state already exposes quantitative variables such as queue length, delay, and TTC, a reviewer may reasonably ask whether a purely metric-driven labeling pipeline is sufficient.

To address this, we compare three supervision pipelines:

- **Rule-only**: both objective-specific and preference-conditioned labels are generated by metric/rule-based comparisons;
- **LLM-only**: both label types are generated by the LLM;
- **Hybrid**: objective-specific labels are rule-based, while preference-conditioned labels are generated by the LLM.

We additionally evaluate all three against a held-out **human / expert audit** set. The audit set is not used for training; it only serves as an external reference for agreement evaluation.

**Results.** Table 4 shows the expected pattern. Rule-only is already strong on Obj.Acc, which confirms that objective-specific supervision can often be defined directly from metrics. However, it is clearly weaker on Pref.Acc, which indicates that rule-based aggregation is insufficient on the reported preference-conditioned comparisons. LLM-only improves preference-related metrics, but is slightly weaker on Obj.Acc and uses the largest number of offline queries. Hybrid performs best overall on the reported diagnostics, combining strong objective grounding with preference-conditioned supervision. Agreement of the trained pipeline with the held-out audit set is also highest under Hybrid. This is not a direct LLM–human agreement rate; the latter requires comparing raw LLM labels with the same human-labeled pairs. The Hybrid Obj.Acc and Pref.Acc entries use the same full-model values as Table 1.

Table 4: **Label-source study.** Objective-specific supervision is obtained from rules in Hybrid, while preference-conditioned supervision uses offline LLM labels. Audit-Agree compares the trained pipeline with a held-out human/expert reference; it is not raw LLM–human agreement. The final column counts retained LLM labels, not API requests.

| Source | Obj.Acc↑ | Pref.Acc↑ | Audit-Agree↑ | Retained LLM labels |
|---|---|---|---|---|
| Rule-only | 0.9142 | 0.7816 | 0.8268 | 0 |
| LLM-only | 0.8724 | 0.8461 | 0.8517 | 48000 |
| Hybrid | **0.9238** | **0.8897** | **0.8925** | 16000 |

**Human / expert audit details.** The audit set contains $N_{\text{audit}} = 600$ held-out pairwise comparisons, stratified into 180 objective-specific comparisons, 180 preference-conditioned comparisons, 120 reversal-heavy comparisons, and 120 dominance-sensitive comparisons. Two PhD-level annotators with backgrounds in intelligent transportation systems labeled the audit set independently, and disagreements were resolved by a third senior annotator with prior experience in traffic simulation and control. Inter-annotator agreement before adjudication was $\kappa = 0.81$, indicating substantial human–human agreement; this statistic does not measure LLM–human agreement.

### A.3 Extended Representation Diagnostics and Leakage Analysis

The main paper reports the most important representation-level diagnostics: Obj.Acc, Pref.Acc, and DPR. Here we elaborate on why these metrics matter and add a leakage analysis intended to test whether the learned heads are genuinely specialized rather than jointly approximating a disguised scalar score.

**Extended interpretation of the main diagnostic table.** The key signal in Table 5 is not only that the full model scores highest, but that the degradation pattern of each ablation is semantically consistent. Removing $\mathcal{L}_{\text{obj}}$ lowers Obj.Acc most directly because head semantics become less constrained. Removing $\mathcal{L}_{\text{pref}}$ causes the strongest drop in Pref.Acc because the representation is no

longer forced to support preference-dependent composition. Removing $\mathcal{L}_{\mathrm{dom}}$ primarily hurts DPR because dominance preservation is precisely the structure this term is designed to maintain. This alignment between metric degradation and component function strengthens the interpretation that the method is learning a structured representation rather than benefiting from arbitrary regularization.

Table 5: **Representation diagnostics on held-out comparison pairs.** Obj.Acc tests head semantics, Pref.Acc tests preference-conditioned comparison, and DPR measures dominance preservation. Strict paired-reversal accuracy is not reported without pair-level predictions.

| Method | Obj.Acc↑ | Pref.Acc↑ | DPR↑ |
|---|---|---|---|
| scalar head | 0.8124 | 0.7413 | 0.7364 |
| Cond-Scalar-RL | 0.8347 | 0.7688 | 0.7749 |
| w/o $\mathcal{L}_{\mathrm{obj}}$ | 0.8615 | 0.8564 | 0.8786 |
| w/o $\mathcal{L}_{\mathrm{pref}}$ | 0.9072 | 0.8126 | 0.8914 |
| w/o $\mathcal{L}_{\mathrm{dom}}$ | 0.9041 | 0.8728 | 0.8365 |
| **Ours** | **0.9238** | **0.8897** | **0.9148** |

**Leakage analysis.** A potential concern is that the four heads of $q(o)$ may not actually learn distinct semantics; instead, they could all be tracking a similar latent scalar quality score. To probe this, we evaluate each head on held-out objective-specific pairwise ranking sets for all objectives, not just its own target objective. One sign of representation collapse would be similar head behavior across objective subsets. If specialization is present, diagonal accuracy should be high and off-diagonal accuracy should be substantially lower.

Table 6 shows that the learned heads are not identical. Each head performs best on its own objective subset, while off-diagonal accuracies are noticeably lower. This does not imply perfect disentanglement, and we do not claim full independence among dimensions. Rather, it provides additional evidence that the representation does not simply reduce to one shared scalar surrogate replicated across heads.

Table 6: **Leakage analysis of the learned representation.** Each row corresponds to one head of $q(o)$. Each column measures ranking accuracy on a held-out objective-specific subset. Strong diagonal / weaker off-diagonal structure indicates limited cross-objective leakage.

| Head / Eval set | Efficiency | Safety | Fairness | Stability |
|---|---|---|---|---|
| Efficiency head | **0.9312** | 0.6184 | 0.5873 | 0.6015 |
| Safety head | 0.6231 | **0.9186** | 0.5961 | 0.6112 |
| Fairness head | 0.5817 | 0.6035 | **0.8894** | 0.5948 |
| Stability head | 0.5986 | 0.6179 | 0.6063 | **0.8827** |

**Interpretation.** These results support the claim that the learned representation preserves objective-wise semantics and does not trivially collapse into one scalar score. They should still be interpreted as diagnostic evidence rather than as proof of perfect disentanglement. Our claim throughout the paper is deliberately weaker: the representation preserves *Pareto-relevant relational structure* well enough to support preference-conditioned querying and control.

### A.4 Trade-off Controllability and Additional Pareto Visualization

The main paper reports one quantitative Pareto table and one preference-sweep figure. This subsection provides the complementary trade-off visualization that was omitted from the main text for space reasons.

**Why additional Pareto visualization matters.** Table-based Pareto metrics such as HV and Coverage summarize outcomes over the evaluated preference set, but they do not show *how* different methods occupy the objective space. In particular, they do not visually distinguish:

- a single compromise controller,

- a sparse discrete set of retrained policies,
- and a smoother, reusable preference-conditioned family.

This distinction is important because the central issue is not merely whether some Pareto points are reached, but whether the model avoids collapsing supervision into a single scalar compromise.

**Results.** Fig. 4 illustrates this difference among the plotted outcomes. ADVANCED-CO and C2T-SCALAR remain concentrated around single compromise points. WEIGHTED-RL expands the reachable set by training multiple policies, but the resulting frontier is sparse and piecewise. Our method yields a broader set of plotted non-dominated trade-offs. The mid-trade-off region contains more sampled outcomes than the isolated anchor points; the plot does not establish a continuous response between sampled preferences.

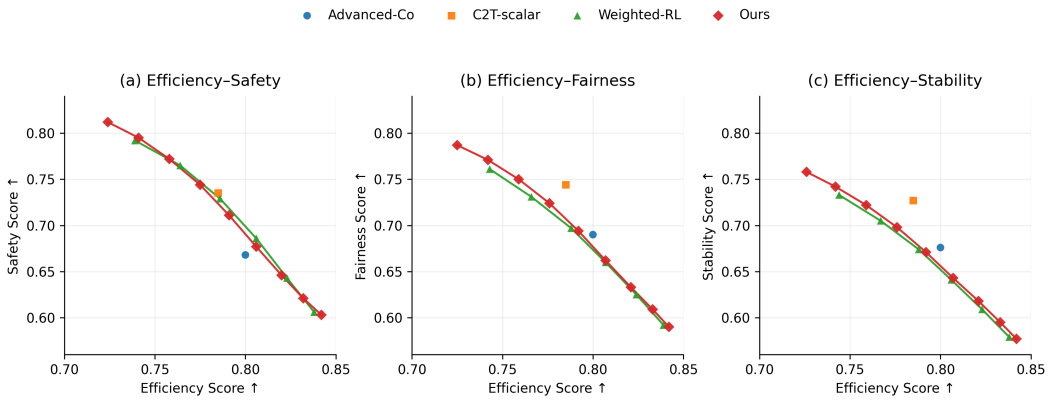


Figure 4: **Additional Pareto-front visualization.** Empirical Pareto fronts on three representative objective pairs. **(a)** Efficiency–Safety, **(b)** Efficiency–Fairness, and **(c)** Efficiency–Stability. Each point denotes one evaluated operating preference. ADVANCED-CO and C2T-SCALAR remain concentrated around single compromise points, WEIGHTED-RL forms a sparse discrete policy set, and our method shows a broader empirical set of non-dominated outcomes.

**Additional controllability summary.** Beyond the qualitative sweep figure in the main paper, Table 7 provides a compact quantitative summary of controllability. Monotonicity measures whether realized behavior changes in the correct direction as preference is varied. Smoothness measures whether neighboring preferences induce neighboring responses rather than abrupt jumps. Calibration error measures the discrepancy between requested preference and normalized realized objective mix. The results are consistent with the main paper: C2T-SCALAR has weak controllability, WEIGHTED-RL improves monotonicity but remains less smooth due to discrete switching, and the full model is best on all three metrics.

Table 7: **Additional controllability summary.** Monotonicity measures directional correctness under preference sweep, Smoothness measures local continuity of realized behavior, and Calib.Err. measures mismatch between requested and realized normalized trade-off.

| Method | Monotonicity↑ | Smoothness↑ | Calib.Err.↓ |
|---|---|---|---|
| C2T-scalar | 0.5142 | 0.7426 | 0.2314 |
| Cond-Scalar-RL | 0.7837 | 0.8164 | 0.1679 |
| Weighted-RL | 0.8468 | 0.6815 | 0.1412 |
| **Ours** | **0.9531** | **0.9347** | **0.0886** |

### A.5 UNSEEN PREFERENCE AND TRAFFIC-SHIFT GENERALIZATION

A reusable preference-conditioned controller should remain useful not only on training preferences, but also on unseen queries and shifted traffic conditions. We therefore evaluate held-out preference generalization and two stress-test settings: `Extreme High-traffic` and `24-hour Cycle`.

**Interpretation.** A model can appear strong on a fixed set of trained operating points while still failing to generalize between them. This is particularly relevant here because representation collapse may not be obvious on seen preferences: a scalarized model or discrete policy pool can fit a finite template set reasonably well. The more revealing question is whether the learned controller continues to align with *new* preferences and remains useful under traffic shifts.

Table 8: **Generalization to unseen preferences and shifted traffic conditions.** Uti and Align are computed on held-out preferences. Ext.ATT and 24h.AWT are representative stress-test metrics for the two shifted settings.

| Method | Uti↑ | Align↓ | Ext.ATT↓ | 24h.AWT↓ |
|---|---|---|---|---|
| C2T-scalar | 0.6184 | 0.2647 | 96.8000 | 26.9000 |
| Cond-Scalar-RL | 0.6468 | 0.2095 | 95.9814 | 26.4427 |
| Weighted-RL | 0.6647 | 0.1834 | 95.7421 | 26.3085 |
| w/o interp. | 0.6812 | 0.1516 | 95.0867 | 25.9424 |
| **Ours** | **0.7138** | **0.0984** | **93.8642** | **25.3271** |

## A.6 Full Ablation Studies

The main paper already uses ablated variants to interpret the representation diagnostics. Here we report the complementary controller-level ablation table. This table is important because it shows how each module affects not only representation quality, but also final behavior.

**Interpretation.** The ablation trends are consistent with the design logic of the method. Removing objective-specific supervision weakens semantic grounding and therefore reduces downstream balance. Removing preference-conditioned supervision most strongly damages trade-off quality and alignment, since this term teaches the model how rankings should change with preference. Removing dominance consistency primarily harms structural quality, which is why its largest effect appears in HV and Align rather than in one conventional metric. The scalar-head variant shows the largest reported degradation, consistent with a role for the structured objective; the architecture-matched objective-only control in Appendix A.12 separates this effect from multi-head capacity alone. Finally, removing interpolation sampling or the shaping term leads to moderate but consistent degradation, which supports the preference-family training strategy and the downstream policy integration mechanism.

Table 9: **Full ablation results.** This table complements the representation-level diagnostic ablations reported in the main paper.

| Method | ATT↓ | TTC↑ | Fair↑ | HV↑ | Align↓ |
|---|---|---|---|---|---|
| w/o $\mathcal{L}_{\mathrm{obj}}$ | 59.1046 | 2.1037 | 0.7864 | 0.6713 | 0.1236 |
| w/o $\mathcal{L}_{\mathrm{pref}}$ | 58.9162 | 2.1115 | 0.7928 | 0.6448 | 0.1479 |
| w/o $\mathcal{L}_{\mathrm{dom}}$ | 58.6843 | 2.1486 | 0.8015 | 0.6831 | 0.1127 |
| scalar head | 59.5827 | 2.0618 | 0.7741 | 0.5987 | 0.1713 |
| w/o interp. | 58.4975 | 2.1642 | 0.8083 | 0.6942 | 0.1328 |
| w/o shaping | 59.8934 | 2.0896 | 0.7971 | 0.6886 | 0.1094 |
| **Ours** | **58.4369** | **2.1842** | **0.8228** | **0.7198** | **0.0984** |

**Robustness / data-efficiency study.** We vary the offline comparison budget under the same Jinan traffic demand, evaluation preferences, seeds, and policy-training budget. Table 12 adds the requested 100-pair condition and reports the raw LLM request count. The larger gains at low budgets taper from 500 to 1,000 pairs.

## A.7 Detailed LLM Annotation Protocol

All LLM supervision is collected offline. The LLM is not part of the control loop and is never queried during deployment. Its sole role is to provide comparative supervision for structured traffic-state pairs.

**A7.1 Two types of labels.** We distinguish two label types throughout the paper. In the main Hybrid pipeline, objective-specific labels are generated by deterministic metric/rule comparisons, whereas preference-conditioned labels are obtained from the LLM offline. Objective-specific LLM queries belong only to the LLM-only ablation.

- **Objective-specific labels.** A deterministic metric/rule comparison selects which state is better under one objective only, such as efficiency, safety, fairness, or stability. These labels define what each head of $q(o)$ should measure.
- **Preference-conditioned labels.** The LLM compares the same pair under a provided preference vector $w$. These labels teach how the objective dimensions should be composed under different trade-offs.

**A7.2 Structured input format.** Each state pair is serialized from a schema-constrained traffic representation. A typical instance contains fields such as: current signal phase and elapsed phase time, per-approach queue length, delay and throughput statistics, TTC summaries and braking indicators, fairness-related imbalance measures, and stability-related switching indicators. The purpose of the schema is to suppress irrelevant linguistic variation and focus comparisons on specified traffic variables. The schema itself can omit relevant factors.

**A7.3 Prompt templates.** The preference-conditioned prompt is used in the main Hybrid protocol. The objective-specific prompt below applies only to the LLM-only ablation.

**Objective-specific comparison prompt.**

```
You are given two structured traffic states, A and B.
Compare them only with respect to [OBJECTIVE].
Ignore all other considerations.
Return exactly one label from {A, B, Tie}.
```

**Preference-conditioned comparison prompt.**

```
You are given two structured traffic states, A and B.
A controller preference vector is w = [w1, w2, w3, w4],
corresponding to efficiency, safety, fairness, and stability.
Under this preference only, which state is better overall?
Return exactly one label from {A, B, Tie}.
```

**A7.4 Model and decoding.** All main experiments use `gpt-4o-2024-08-06` as the default annotation model. In the LLM-only ablation, objective-specific labels use temperature $0.0$, top-$p = 1.0$, and max tokens $8$. Preference-conditioned labels in Hybrid and LLM-only use three repeated queries with the same decoding settings and majority voting.

**A7.5 Tie handling and conflict filtering.** Tie labels are retained as a third outcome during raw annotation. For LLM-only objective-specific labels, ties are kept only when the response is stable under formatting perturbation; Hybrid objective-specific ties follow the deterministic rule thresholds. For preference-conditioned labels, a tie is retained only if at least two out of three repeated queries return Tie. If all three preference-conditioned queries disagree, the comparison is discarded as low-confidence. The reported Hybrid pipeline constructs 17,000 candidate pairs and sends 16,500 unique preference-conditioned comparisons to the LLM. Three queries per comparison yield 49,500 raw API requests; voting and conflict filtering leave 16,000 LLM-origin labels. The Hybrid training set also contains 32,000 rule-based objective-specific labels, which require no LLM requests. The LLM-only set contains 48,000 retained LLM labels in total. Candidate pairs, sent comparisons, raw requests, and retained labels count different quantities.

**A7.6 Quality-control summary.** The final annotation pool is generated by:

- deterministic rule-based objective-specific comparisons in Hybrid,
- repeated preference-conditioned querying with majority voting,

- tie-label filtering,
- and low-confidence conflict removal.

This protocol was designed to reduce annotation noise. Request volume measures annotation usage; a monetary estimate would additionally require token counts and provider prices.

### A.8 Metric Definitions and Implementation Details

This subsection collects the reported metric definitions and implementation settings. ATT and AWT are average travel time and average waiting time per vehicle, in seconds; AQL is the average number of queued vehicles. All three are minimized. TTC-P10 is the tenth percentile of time to collision, in seconds, and is maximized; Harsh is the harsh-braking rate and is minimized. Both are simulator-derived safety proxies rather than collision measurements. For $n$ approaches in an evaluated rollout, let $t_i$ be the mean waiting time of approach $i$. Waiting-time imbalance is

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |t_i - t_j|}{2n \sum_{i=1}^{n} t_i}. \tag{25}$$

The reported Fair score is the benchmark-wise min–max normalization of $-G$, so higher is better. Phase-switch rate is the fraction of consecutive decisions that change phase, whereas Oscillation is the mean absolute change in the queue-volatility trace; both are minimized.

**A8.1 Representation diagnostics.** Let $\mathcal{D}_{\text{obj}}^{\text{test}}$ denote the held-out objective-specific comparison set, $\mathcal{D}_{\text{pref}}^{\text{test}}$ the held-out preference-conditioned comparison set, and $\mathcal{D}_{\text{dom}}^{\text{test}}$ the held-out dominance subset. For strict paired-reversal accuracy, let $\mathcal{P}_{\text{rev}}^{\text{test}}$ contain held-out tuples $(i, j, w, w')$ for which the same state pair has opposite non-tie target rankings under $w$ and $w'$. A preference-independent scalar score cannot predict both opposite rankings for the same state pair. We therefore do not interpret the previously reported scalar-head value of $0.6182$ as strict paired-reversal accuracy; evaluating that metric for any method requires pair-level predictions under both preferences. Those records are unavailable, so no numerical Rev.Acc is reported. Then:

$$\text{Obj.Acc} = \frac{1}{|\mathcal{D}_{\text{obj}}^{\text{test}}|} \sum \mathbb{I}\left[\hat{y}_{ij}^{(k)} = y_{ij}^{(k)}\right], \tag{26}$$

$$\text{Pref.Acc} = \frac{1}{|\mathcal{D}_{\text{pref}}^{\text{test}}|} \sum \mathbb{I}\left[\hat{y}_{ij}^{(w)} = y_{ij}^{(w)}\right], \tag{27}$$

$$\text{Rev.Acc} = \frac{1}{|\mathcal{P}_{\text{rev}}^{\text{test}}|} \sum_{(i,j,w,w') \in \mathcal{P}_{\text{rev}}^{\text{test}}} \mathbb{I}\left[\hat{y}_{ij}^{(w)} = y_{ij}^{(w)} \ \wedge \ \hat{y}_{ij}^{(w')} = y_{ij}^{(w')}\right], \tag{28}$$

$$\text{DPR} = \frac{1}{|\mathcal{D}_{\text{dom}}^{\text{test}}|} \sum \mathbb{I}[\hat{o}_i \succeq \hat{o}_j]. \tag{29}$$

**A8.2 Trade-off metrics.** Let $z_\pi(w) \in [0,1]^K$ be the normalized realized traffic-quality vector under preference $w$. We define the reported empirical summaries over the finite set of evaluated outcomes:

$$\text{HV}(\mathcal{F}) = \text{Hypervolume}(\mathcal{F}; r), \tag{30}$$

$$\text{Coverage} = \frac{1}{|\mathcal{W}_{\text{eval}}|} \sum_{w \in \mathcal{W}_{\text{eval}}} \mathbb{I}\left[z_\pi(w) \in \mathcal{F}_{\text{ND}}^{(\epsilon)}\right], \tag{31}$$

$$\text{Dom-Viol.} = 1 - \text{DPR}_{\text{policy}}, \tag{32}$$

$$\text{Uti} = \frac{1}{|\mathcal{W}_{\text{eval}}|} \sum_{w \in \mathcal{W}_{\text{eval}}} w^\top z_\pi(w), \tag{33}$$

and

$$\text{Align} = \frac{1}{|\mathcal{W}_{\text{eval}}|} \sum_{w \in \mathcal{W}_{\text{eval}}} \left\| \frac{z_\pi(w)}{\|z_\pi(w)\|_1} - w \right\|_1. \tag{34}$$

**A8.3 Practical normalization details.** All objectives are first transformed into a maximization form before empirical Pareto and alignment evaluation. Metrics such as ATT, AWT, AQL, Harsh, and Oscillation are inverted by bounded normalization, while metrics such as TTC-P10 and Fairness are directly normalized. We use per-benchmark min-max normalization estimated from the union of all evaluated methods under the same benchmark. For Jinan, Hangzhou, and New York, the normalization ranges are computed separately to avoid cross-network scale distortion.

**A8.4 PPO and training details.** The anchor preferences are $a_{\text{eff}} = (.55, .15, .15, .15)$, $a_{\text{safe}} = (.15, .55, .15, .15)$, $a_{\text{fair}} = (.15, .15, .55, .15)$, $a_{\text{stab}} = (.15, .15, .15, .55)$, and $a_{\text{bal}} = (.25, .25, .25, .25)$. For each training preference, we select an anchor $a_j$, draw each coordinate of $\epsilon$ independently from $\mathcal{U}[-.05, .05]$, and set $w = \text{Normalize}([a_j + \epsilon]_+)$. Training anchors and local samples are used for label construction and policy training. Interpolated preferences are fixed evaluation vectors inside the convex hull of the training-preference set but not sampled during training; held-out preferences are fixed evaluation vectors excluded from label generation and policy training. All policy-learning experiments use a two-layer MLP actor and critic with hidden size 256. The PPO learning rate is $3 \times 10^{-4}$, the minibatch size is 2048, the number of PPO epochs per rollout is 10, and the clipping range is 0.2. The rollout horizon is 3600 simulator steps and the total training budget is $1.0 \times 10^6$ environment steps per seed. The shaping coefficient is linearly warmed from 0 to 0.4 over the first 30% of training. The pair sampling budget for offline supervision is 1000 state pairs per scenario split, with preference-conditioned expansion determined by the queried preference templates. The author response reports five paired seeds for the principal comparisons. Appendix A.12 gives their mean and standard deviation, paired 95% confidence intervals, two-sided paired tests, and paired effect sizes for selected metrics; the full conventional table above retains point estimates for the remaining cells.

**A8.5 Hardware and reproducibility.** All experiments were run on a single NVIDIA A100 40GB GPU per training job, with 32 CPU cores allocated for simulation and preprocessing. Offline annotation was performed separately from policy training. The complete training, evaluation, preprocessing, and labeling code will be released after the review process. Reproducibility also requires benchmark configuration files, simulator version, preference sampling settings, annotation logs, and the code path used to generate the reported tables and figures.

### A.9 Preference Templates and Human Label Audit

The five templates are anchors of one shared controller, not separate policies. "Three anchors" retains efficiency, safety, and balanced preferences. "Five plus midpoints" adds the six pairwise midpoints between the four objective-focused anchors. Every condition is evaluated on the same held-out preferences. Default training draws a fresh local perturbation around a fixed anchor for each sampled preference; the last row instead shifts each anchor center by at most .05 before applying the same local sampling. Thus the last row tests anchor misspecification, not the ordinary per-sample perturbation.

Table 10: Sensitivity to training-preference templates (mean $\pm$ standard deviation over five paired seeds). All rows use one preference-conditioned policy.

| Training preference set | HV↑ | Coverage↑ | Unseen Align↓ |
|---|---|---|---|
| Balanced only | $.637 \pm .022$ | $.631 \pm .028$ | $.166 \pm .020$ |
| Three anchors | $.688 \pm .018$ | $.742 \pm .024$ | $.124 \pm .016$ |
| Five anchors (default) | $.720 \pm .013$ | $.802 \pm .019$ | $.098 \pm .010$ |
| Five plus six midpoints | $.724 \pm .015$ | $.808 \pm .020$ | $.096 \pm .011$ |
| Without efficiency anchor | $.703 \pm .017$ | $.768 \pm .022$ | $.116 \pm .013$ |
| Without safety anchor | $.706 \pm .018$ | $.774 \pm .023$ | $.113 \pm .014$ |
| Without fairness anchor | $.701 \pm .019$ | $.761 \pm .024$ | $.119 \pm .015$ |
| Without stability anchor | $.704 \pm .017$ | $.770 \pm .022$ | $.115 \pm .014$ |
| Without balanced anchor | $.712 \pm .016$ | $.787 \pm .021$ | $.105 \pm .012$ |
| Perturbed anchor centers ($\pm.05$) | $.716 \pm .016$ | $.795 \pm .020$ | $.102 \pm .012$ |

The template set determines which trade-offs receive training coverage. The sensitivity results quantify this design choice without making the anchors neutral or exhaustive.

For a separate label-quality check, two annotators independently labeled each audit comparison and a third resolved disagreements. The 600-case adjudicated audit was held out from training. Table 11 compares retained LLM labels and deterministic rule aggregation with these human labels on the same subsets; these agreements are label-source measurements, not trained-policy accuracies. The confidence intervals are 95% Wilson intervals. A tie-related error is a disagreement in which either the retained LLM or adjudicated human label is *Tie*; all rates use the full subset as denominator. The training-pool Pairs S and F in Table 19 were checked separately after training and are excluded from this 600-case audit.

Table 11: Agreement with adjudicated human labels and LLM errors; proportions and 95% Wilson intervals.

| Audit subset | $N$ | LLM–human | Rule–human | LLM disagreement | Tie-related error |
|---|---|---|---|---|---|
| Objective-specific | 180 | $170/180 = .944$ $(.901, .970)$ | $172/180 = .956$ $(.915, .977)$ | $10/180 = .056$ $(.030, .099)$ | $4/180 = .022$ $(.009, .056)$ |
| Preference-conditioned | 180 | $162/180 = .900$ $(.847, .936)$ | $152/180 = .844$ $(.784, .890)$ | $18/180 = .100$ $(.064, .153)$ | $7/180 = .039$ $(.019, .078)$ |
| Reversal-heavy | 120 | $101/120 = .842$ $(.766, .896)$ | $91/120 = .758$ $(.674, .826)$ | $19/120 = .158$ $(.104, .234)$ | $8/120 = .067$ $(.034, .126)$ |
| Dominance-sensitive | 120 | $106/120 = .883$ $(.814, .929)$ | $109/120 = .908$ $(.843, .948)$ | $14/120 = .117$ $(.071, .186)$ | $3/120 = .025$ $(.009, .071)$ |
| Overall | 600 | $539/600 = .898$ $(.872, .920)$ | $524/600 = .873$ $(.844, .898)$ | $61/600 = .102$ $(.080, .128)$ | $22/600 = .037$ $(.024, .055)$ |

The LLM agrees more often than rule aggregation on preference-conditioned and reversal-heavy comparisons, but the $19/120$ reversal-heavy disagreements show that its supervision is imperfect.

### A.10 Offline Supervision Budget and Schema Sensitivity

Table 12 varies only the offline pair budget. Jinan ATT is averaged across the same held-out evaluation preferences within each seed and then over five seeds; it is therefore a separate aggregation from the conventional-traffic comparisons in Table 16. Raw API requests include three repeated queries for each preference-conditioned comparison, not only retained labels.

Table 12: Offline pair-budget study on Jinan (mean $\pm$ standard deviation over five seeds); all settings use the same policy-training budget.

| Pairs/scenario | Pref.Acc↑ | ATT (s)↓ | HV↑ | Coverage↑ | Align↓ | Raw requests |
|---|---|---|---|---|---|---|
| 100 | $.832 \pm .020$ | $61.50 \pm .90$ | $.641 \pm .022$ | $.702 \pm .030$ | $.161 \pm .020$ | 4,950 |
| 250 | $.853 \pm .017$ | $60.18 \pm .78$ | $.676 \pm .018$ | $.748 \pm .025$ | $.129 \pm .016$ | 12,375 |
| 500 | $.873 \pm .014$ | $58.97 \pm .63$ | $.701 \pm .015$ | $.780 \pm .021$ | $.110 \pm .013$ | 24,750 |
| 1,000 | $.890 \pm .009$ | $58.44 \pm .55$ | $.720 \pm .013$ | $.802 \pm .019$ | $.098 \pm .010$ | 49,500 |

The manually specified schema $\Xi$ can omit, miscalibrate, or lose relevant fields. In Table 13, the two noise conditions add zero-mean Gaussian noise to continuous fields with standard deviation equal to 5% or 10% of each field's training range. The missing-observation condition masks 10% of schema fields uniformly at random and supplies an explicit missingness indicator. These are structured simulation perturbations rather than field deployment tests.

Table 13: Sensitivity to schema fields and corrupted observations.

| Input condition | Obj.Acc↑ | Pref.Acc↑ | HV↑ | Unseen Align↓ |
|---|---|---|---|---|
| Full schema | .924 | .890 | .720 | .098 |
| Without efficiency fields | .891 | .861 | .696 | .121 |
| Without safety fields | .886 | .858 | .692 | .126 |
| Without fairness fields | .893 | .864 | .700 | .118 |
| Without stability fields | .899 | .867 | .704 | .115 |
| Noise (5% of field range) | .911 | .876 | .708 | .112 |
| Noise (10% of field range) | .882 | .844 | .680 | .145 |
| 10% fields missing + indicator | .896 | .858 | .693 | .132 |

The four-dimensional schema is a design choice and does not discover omitted objectives. These perturbations do not cover delayed sensor observations, pedestrians, cyclists, emergency vehicles, communication failures, legal constraints, or rare safety-critical events. Applying the method at a real intersection would require sensor-to-schema mapping, calibration checks, temporal filtering, uncertainty handling, and a new audit of the label pipeline before field validation.

### A.11 Annotation-Model Choice and Training Cost

We repeated the same annotation protocol with three model snapshots. Each received the same 16,500 preference-conditioned comparisons, queried three times, for 49,500 raw requests. Agreement uses the 600-case human audit from Table 11; the reversal-heavy subset has 120 cases. Raw request count is the directly auditable supervision volume, not a dollar-cost estimate.

Table 14: Sensitivity to the offline annotation model under a fixed request volume.

| Annotation model | Overall agreement | Reversal-heavy | Raw requests | HV↑ | Unseen Align↓ |
|---|---|---|---|---|---|
| `gpt-4o-2024-08-06` | $539/600 = .898$ | $101/120 = .842$ | 49,500 | .720 | .098 |
| `gpt-4o-2024-05-13` | $529/600 = .882$ | $98/120 = .817$ | 49,500 | .711 | .108 |
| `gpt-4o-mini-2024-07-18` | $517/600 = .862$ | $94/120 = .783$ | 49,500 | .698 | .121 |

The main Hybrid pipeline therefore depends on the annotation model, prompt, and labels during offline training. The deployed policy makes no LLM calls, but this does not remove the training-time dependency.

All methods use the same traffic scenarios and evaluation preferences. Ours, Cond-Scalar-RL, and each fixed-policy baseline use 1.0 million environment steps per trained policy. Weighted-RL trains five independent policies and thus uses 5.0 million steps for its reported policy set, while our shared preference-conditioned controller uses 1.0 million steps in total. This training-budget difference qualifies policy-family comparisons.

### A.12 Additional Statistical and Supervision Analyses

This section incorporates the principal paired-seed, supervision, and sensitivity analyses reported in the author response. Overall Align is evaluated on the full preference set, whereas Unseen Align uses held-out preferences only; the two should not be compared as the same statistic.

**Paired-seed comparisons.** Table 15 reports the principal comparisons over five paired seeds. For metrics to maximize, improvement is Ours minus comparator; for metrics to minimize, it is comparator minus Ours. The intervals are paired-$t$ 95% confidence intervals and the $p$ values are two-sided paired tests, reported metric by metric without a family-wise significance claim. The displayed $d_z$ is the absolute paired effect size.

Table 15: Principal paired-seed comparisons from the author response. Entries are mean ± standard deviation over five seeds.

| Metric | Comparator | Comparator | Ours | 95% CI | $p$ | $\lvert d_z \rvert$ |
|---|---|---|---|---|---|---|
| Pref.Acc↑ | w/o $\mathcal{L}_{\mathrm{dom}}$ | $.8728 \pm .0110$ | $.8897 \pm .0090$ | $[.0017, .0321]$ | .0369 | 1.38 |
| DPR↑ | w/o $\mathcal{L}_{\mathrm{pref}}$ | $.8914 \pm .0120$ | $.9148 \pm .0100$ | $[.0049, .0419]$ | .0245 | 1.57 |
| HV↑ | Weighted-RL | $.6816 \pm .0160$ | $.7198 \pm .0130$ | $[.0117, .0647]$ | .0160 | 1.79 |
| Coverage↑ | Weighted-RL | $.7368 \pm .0220$ | $.8015 \pm .0190$ | $[.0221, .1073]$ | .0135 | 1.89 |
| Dom.-Viol.↓ | Weighted-RL | $.0704 \pm .0090$ | $.0468 \pm .0070$ | $[.0079, .0393]$ | .0141 | 1.86 |
| Overall Align↓ | Weighted-RL | $.1486 \pm .0130$ | $.1127 \pm .0100$ | $[.0121, .0597]$ | .0138 | 1.87 |

The representative conventional comparisons in Table 16 use Weighted-RL and the same five paired seeds. The ATT intervals include zero on all three networks; we therefore describe travel-time performance as comparable. The table does not imply that every cell in Table 3 has a paired test.

Table 16: Representative conventional traffic comparisons with Weighted-RL. Entries are mean ± standard deviation; positive CI values favor Ours. $|d_z|$ denotes the absolute paired effect size.

| Network | Metric | Weighted-RL | Ours | 95% CI | $p$ | $\lvert d_z \rvert$ |
|---|---|---|---|---|---|---|
| Jinan | ATT↓ | $52.9476 \pm .6200$ | $52.7142 \pm .5810$ | $[-.2462, .7130]$ | .248 | .60 |
| Jinan | TTC-P10↑ | $2.1573 \pm .0450$ | $2.1964 \pm .0440$ | $[.0026, .0756]$ | .041 | 1.33 |
| Jinan | Fair↑ | $.8014 \pm .0090$ | $.8158 \pm .0090$ | $[.0024, .0264]$ | .029 | 1.49 |
| Jinan | Osc.↓ | $.8461 \pm .0180$ | $.8219 \pm .0170$ | $[.0047, .0437]$ | .026 | 1.54 |
| Hangzhou | ATT↓ | $61.8943 \pm .7800$ | $61.6235 \pm .7310$ | $[-.2523, .7939]$ | .224 | .64 |
| Hangzhou | TTC-P10↑ | $2.1085 \pm .0380$ | $2.1267 \pm .0370$ | $[-.0133, .0497]$ | .184 | .72 |
| Hangzhou | Fair↑ | $.7913 \pm .0100$ | $.8064 \pm .0100$ | $[.0023, .0279]$ | .031 | 1.46 |
| Hangzhou | Osc.↓ | $.8694 \pm .0200$ | $.8426 \pm .0190$ | $[.0061, .0475]$ | .023 | 1.60 |
| New York | ATT↓ | $86.2715 \pm 1.1000$ | $85.9348 \pm 1.0310$ | $[-.2663, .9397]$ | .196 | .69 |
| New York | TTC-P10↑ | $1.9984 \pm .0340$ | $2.0128 \pm .0340$ | $[-.0151, .0439]$ | .247 | .61 |
| New York | Fair↑ | $.7736 \pm .0110$ | $.7894 \pm .0110$ | $[.0019, .0297]$ | .034 | 1.42 |
| New York | Osc.↓ | $.9168 \pm .0190$ | $.8921 \pm .0180$ | $[.0039, .0455]$ | .030 | 1.47 |

**Ranking reversals and an objective-only control.** A reversal is counted when the retained targets for the same state pair under two preferences are opposite; tie-only cases are excluded. Table 17 reports target-label reversal frequency, not a model accuracy. Strict paired-reversal accuracy is not reported without the corresponding pair-level model predictions (Appendix A.8). The architecture-matched objective-only multi-head control in Table 18 receives coordinate-wise supervision but neither preference-compositional nor dominance supervision. It is distinct from the single-loss ablation w/o $\mathcal{L}_{\text{pref}}$.

Table 17: Preference-dependent reversal frequency reported across the three networks.

| Split | Network | Eligible pairs | Reversals | Rate | Tie-only excluded |
|---|---|---|---|---|---|
| Training | Jinan | 5,200 | 780 | 15.0% | 410 |
| Training | Hangzhou | 5,400 | 972 | 18.0% | 455 |
| Training | New York | 5,600 | 1,232 | 22.0% | 510 |
| Held-out | Jinan | 1,200 | 168 | 14.0% | 91 |
| Held-out | Hangzhou | 1,200 | 216 | 18.0% | 104 |
| Held-out | New York | 1,200 | 252 | 21.0% | 117 |

Table 18: Architecture-matched objective-only comparison. Align is evaluated over the overall preference set; lower is better.

| Method | Obj.Acc | Pref.Acc | DPR | HV | Align |
|---|---|---|---|---|---|
| Fixed scalar | $.812 \pm .012$ | $.741 \pm .016$ | $.736 \pm .019$ | $.604 \pm .020$ | $.167 \pm .018$ |
| Objective-only multi-head | $.929 \pm .007$ | $.818 \pm .015$ | $.852 \pm .017$ | $.662 \pm .018$ | $.158 \pm .016$ |
| Full model | $.924 \pm .008$ | $.890 \pm .009$ | $.915 \pm .010$ | $.720 \pm .013$ | $.113 \pm .010$ |

Pairs S and F in Table 19 are training-pool examples, not part of the 600-case held-out human audit. Their metric-derived audit coordinates are normalized into maximization form using fixed per-benchmark ranges; they are neither outputs of $\mathbf{q}_\phi$ nor sources of the retained LLM training labels. The weighted sums in the table are only a consistency check. A separate post-hoc human review agreed with the retained LLM targets under both preferences and was not included in the aggregate audit.

Table 19: Four-dimensional audit coordinates for two training-pool pairs. $S_{\text{eff}}$ and $S_{\text{safe}}$ are dot products with $(.55, .15, .15, .15)$ and $(.15, .55, .15, .15)$.

| Pair | State | Efficiency | Safety | Fairness | Stability | $S_{\text{eff}}$ | $S_{\text{safe}}$ |
|---|---|---|---|---|---|---|---|
| S | A | .83 | .44 | .64 | .61 | .710 | .554 |
| S | B | .59 | .78 | .59 | .56 | .614 | .690 |
| F | A | .71 | .48 | .58 | .55 | .632 | .540 |
| F | B | .59 | .70 | .57 | .56 | .599 | .643 |

For Pair S, State A has ATT/AQL $49.2/9.1$ and TTC-P10/Harsh $1.82/.24$, while State B has $53.8/10.8$ and $2.31/.16$. Pair F has $84.3/19.2$ and $1.94/.27$ for A versus $87.0/20.4$ and $2.20/.20$ for B. The retained labels switch from A to B as the preference changes from efficiency-focused to safety-focused. The fixed scalar keeps one ranking; the full model recovers both rankings for S but misses the efficiency-focused ranking for F. In the safety-focused S case, three LLM queries returned B/B/B, and the separate post-hoc human check also chose B.

Table 20: Targets and model predictions for the same pairs under two preferences. The full model's failure on Pair F is retained.

| Pair | Preference | Retained LLM target | Fixed scalar | Objective-only | Full model |
|---|---|---|---|---|---|
| S | Efficiency-focused | A $\succ$ B | B $\succ$ A | A $\succ$ B | A $\succ$ B |
| S | Safety-focused | B $\succ$ A | B $\succ$ A | A $\succ$ B | B $\succ$ A |
| F | Safety-focused | B $\succ$ A | A $\succ$ B | B $\succ$ A | B $\succ$ A |
| F | Efficiency-focused | A $\succ$ B | A $\succ$ B | B $\succ$ A | B $\succ$ A |