ICML 2026 · SEOUL, SOUTH KOREA

RAST-MoE-RL

A Regime-Aware Spatio-Temporal Mixture-of-Experts framework for deep reinforcement learning in ride-hailing

Yuhan Tang*[1] · Kangxin Cui*[2] · Jung Ho Park*[3] · Yibo Zhao*[1] · Xuan Jiang[1] · Haoze He[4] · Jiangbo Yu[5] · Haris Koutsopoulos[6] · Jinhua Zhao[1]

[1]Massachusetts Institute of Technology · [2]Tongji University · [3]Uiversity of California, Berkeley · [4]Carnegie Mellon University · [5]McGill University · [6]Northeastern University

*equal contribution

PROBLEM STATEMENT

The Matching Decision

Every request poses one question, tens of thousands of times an hour: match this passenger to a driver now, or wait for a better one to appear?

Match now

Assign to the nearest driver immediately

Match	**Fast**
PICKUP	**Slow — driver may be far**

vs

Wait

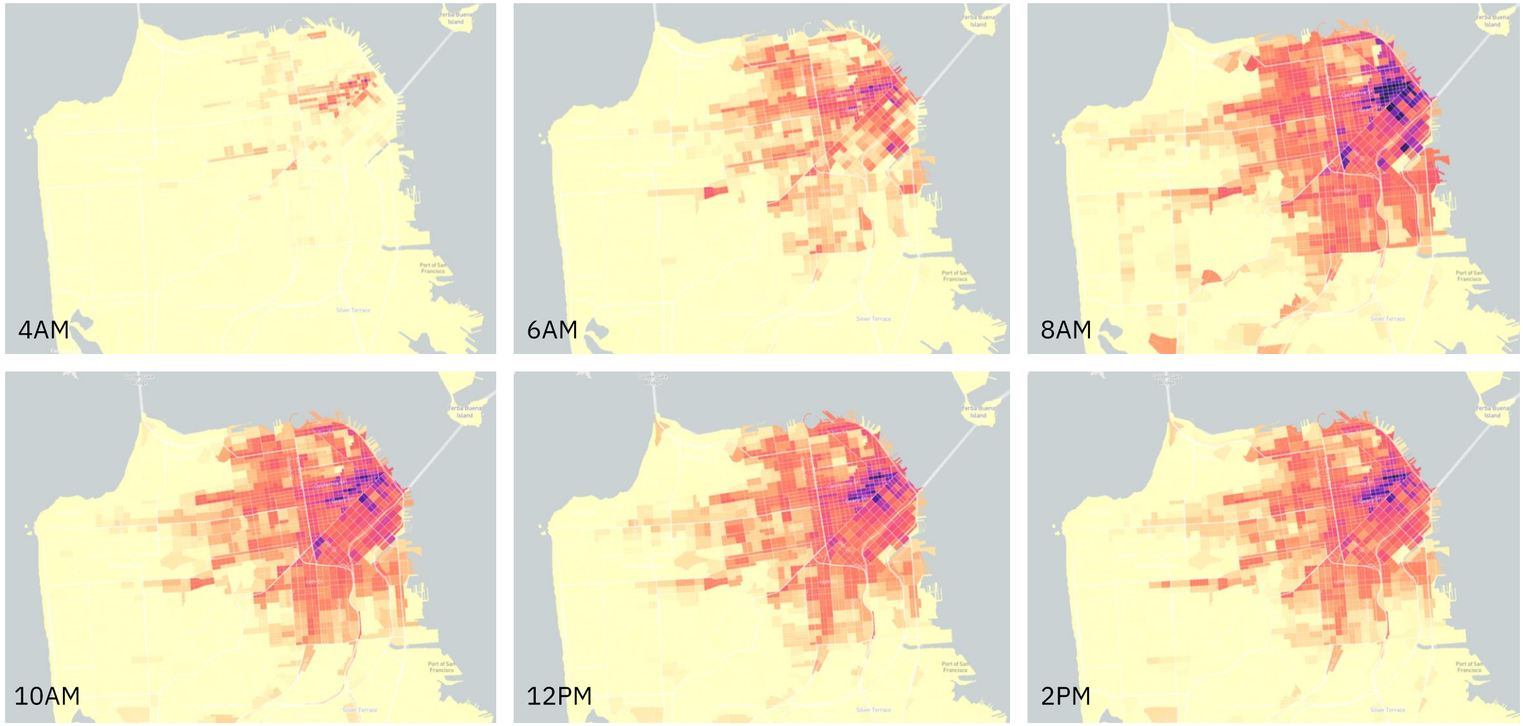
Batch requests, hold for a better match

Match	**Slow — passenger waits**
PICKUP	**Fast — closer driver**

PROBLEM STATEMENT

Challenges

The city runs in distinct regimes that shift continuously — and the right strategy in one actively fails in another.



Heatmap of spatiotemporal demand patterns in San Francisco (SFTCA)

HOLD TOO LONG Penalize pickup too hard and the agent **holds requests indefinitely** — matching waits explode past 20 minutes.

MATCH TOO FAST Penalize batching too hard and it **matches everything instantly** — pickup waits past 25 minutes.

GAP

Limitations of Prior Work

Methods using centralized coordinator do not account for non-stationary regimes

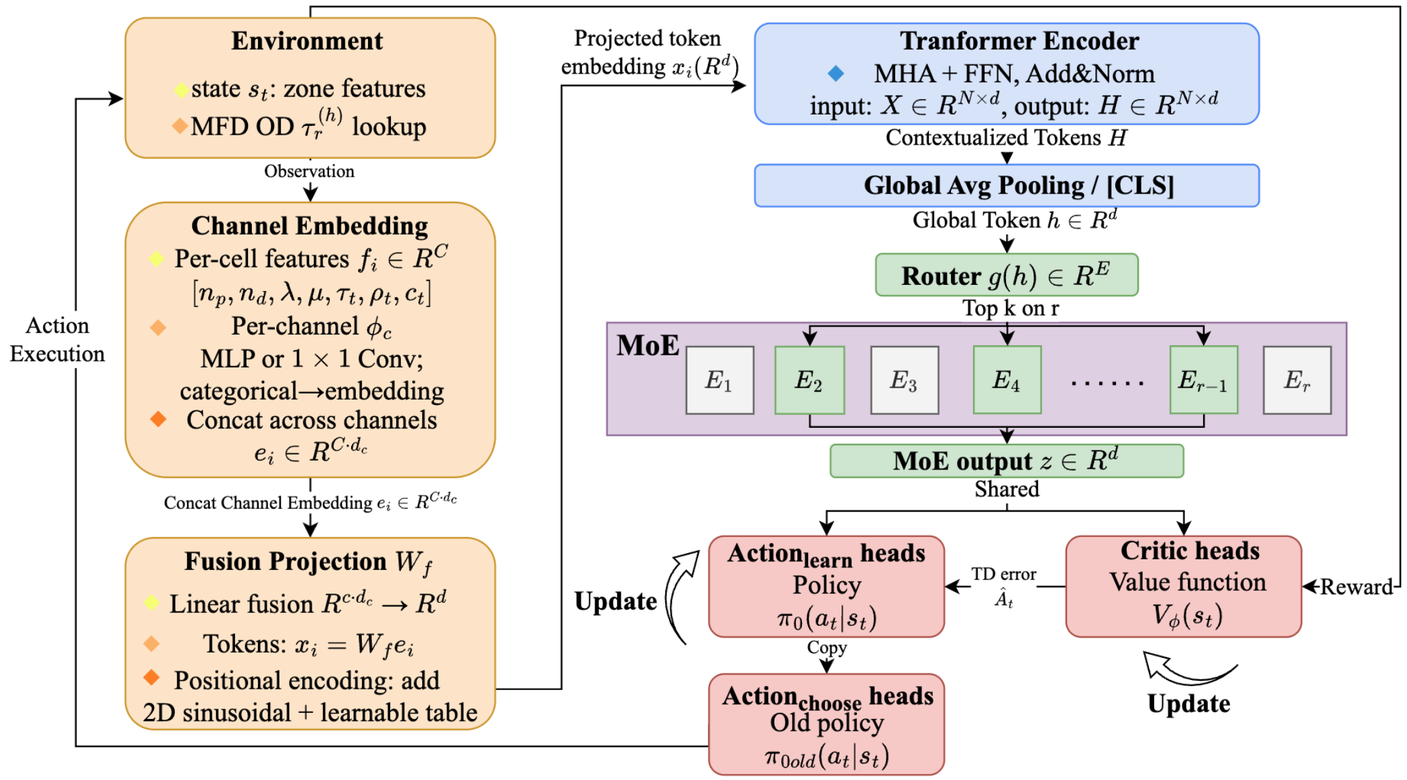
APPROACH	HOW IT WORKS	LIMITATION
Heuristics Fixed windows · time-of-day	Simple, interpretable rules tuned for average conditions.	Does not adapt in real time.
Monolithic RL Single policy network	One set of weights compresses every regime into one model.	Does not generalize well
RAST-MoE-RL Our work	**Experts that each handle a distinct urban regime.**	

Heuristic vs. monolithic RL

APPROACH

Integration of MoE Architecture

Instead of one network handling every situation, a learned gate routes each state to a few **specialist experts**. Different experts are activated at different operational regimes such as rush hour and late at night. This specialization emerges from training.



12M

parameters in the encoder

4/16

experts active per decision

APPROACH

Reward Design

Our reward design prevents reward hacking and promotes temporal balancing of match and pickup waits.

- **Reward balancing match time and pickup time**

$$r_t = -\left[c_m\,\Delta W_t^{\text{match}} + c_p(t)\,\Delta W_t^{\text{pickup}}\right] - \lambda_t\,(g(s_t,a_t) - \alpha).$$

- **Service quality measure soft constraint**

$$g(s_t,a_t) \leq \alpha$$

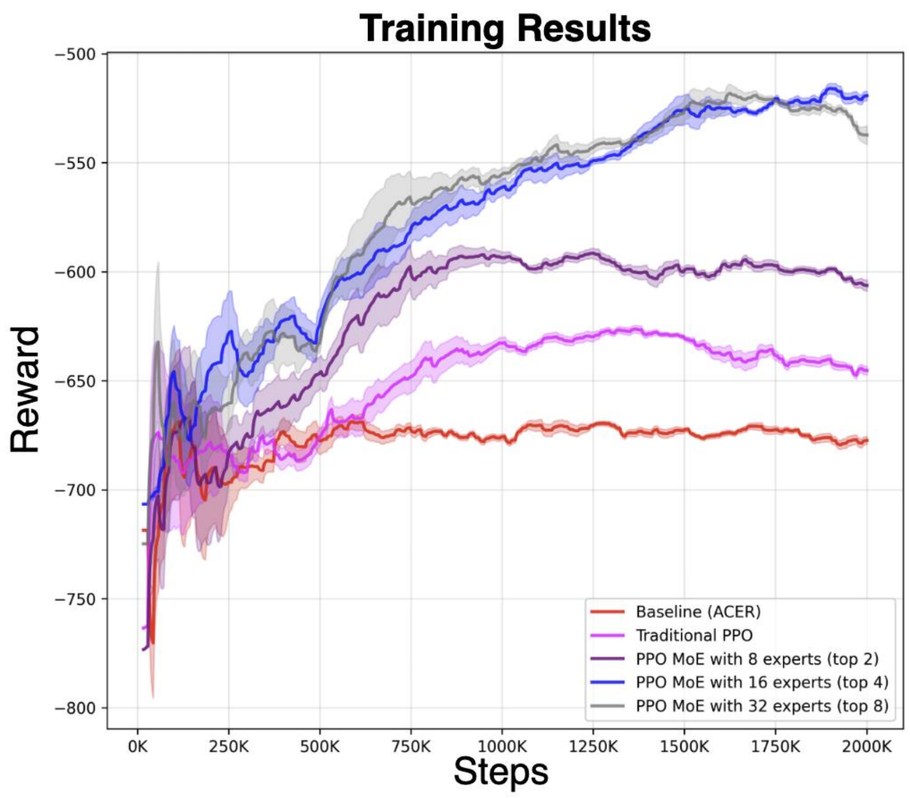
- **Adaptive multiplier update**

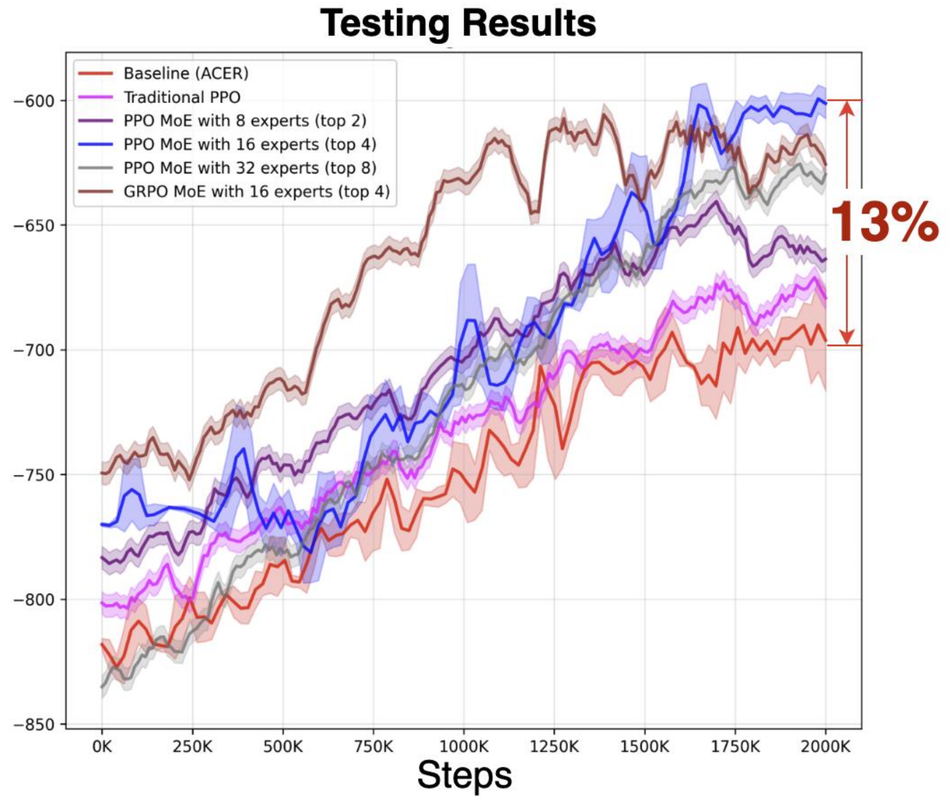
$$\lambda_{t+1} \leftarrow \left[\lambda_t + \xi\,(g(s_t,a_t) - \alpha)\right]_+$$

RESULTS

Training and Testing Report

Comparison of training and testing among baseline algorithms show the advantage of PPO MoE with 16 experts with top=4 expert activation.





RESULTS

Performance Comparison

On real San Francisco TNC data, RAST-MoE-RL beats every baseline — heuristic and RL alike — reducing matching and pickup waits simultaneously.

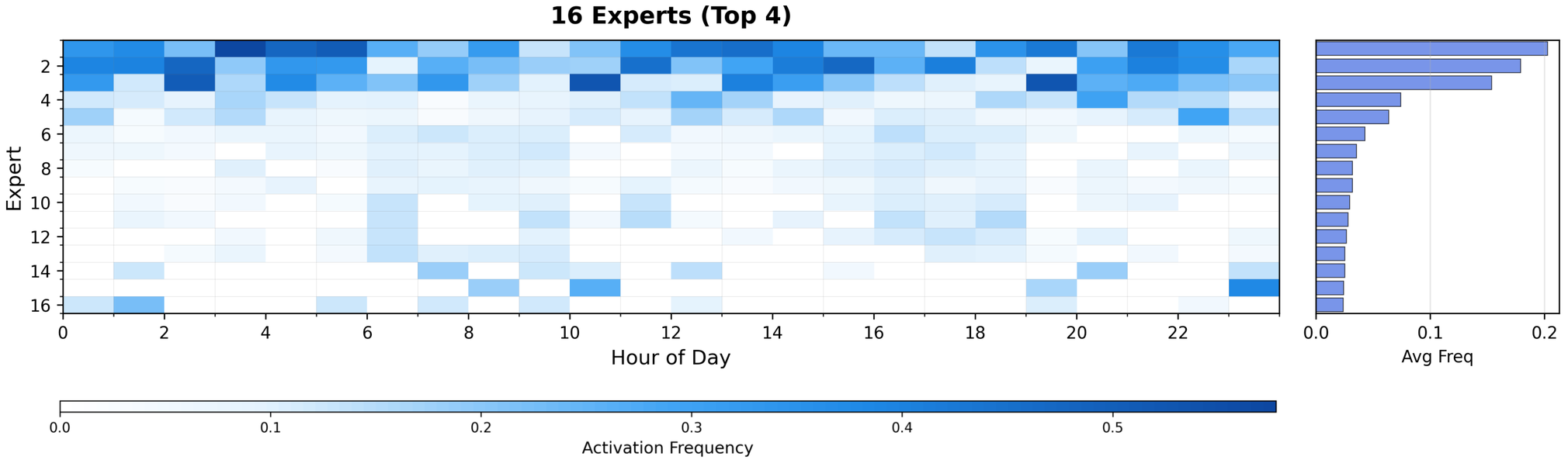
METHOD	MATCHING WAIT SECONDS · LOWER BETTER	PICKUP WAIT SECONDS · LOWER BETTER
RAST-MoE-RL OURS MoE 16/4 · 12M · PPO	**182**	**523**
Dense Transformer 12M · PPO	**201**	**569**
Plain MLP 12M · PPO	**208**	**577**
Best heuristic Conditioned on zone-level congestion	**203**	**784**

−10% matching · −15% pickup · +13% reward. Pickup wait drops **~33%** below the strongest heuristic.

RESULTS

Evidence of Specialization by Regime

The 16 experts do not share the load equally. Each fires in distinct patterns across the 24-hour cycle, matching the city's operational regimes — rush hour, off-peak, late night.

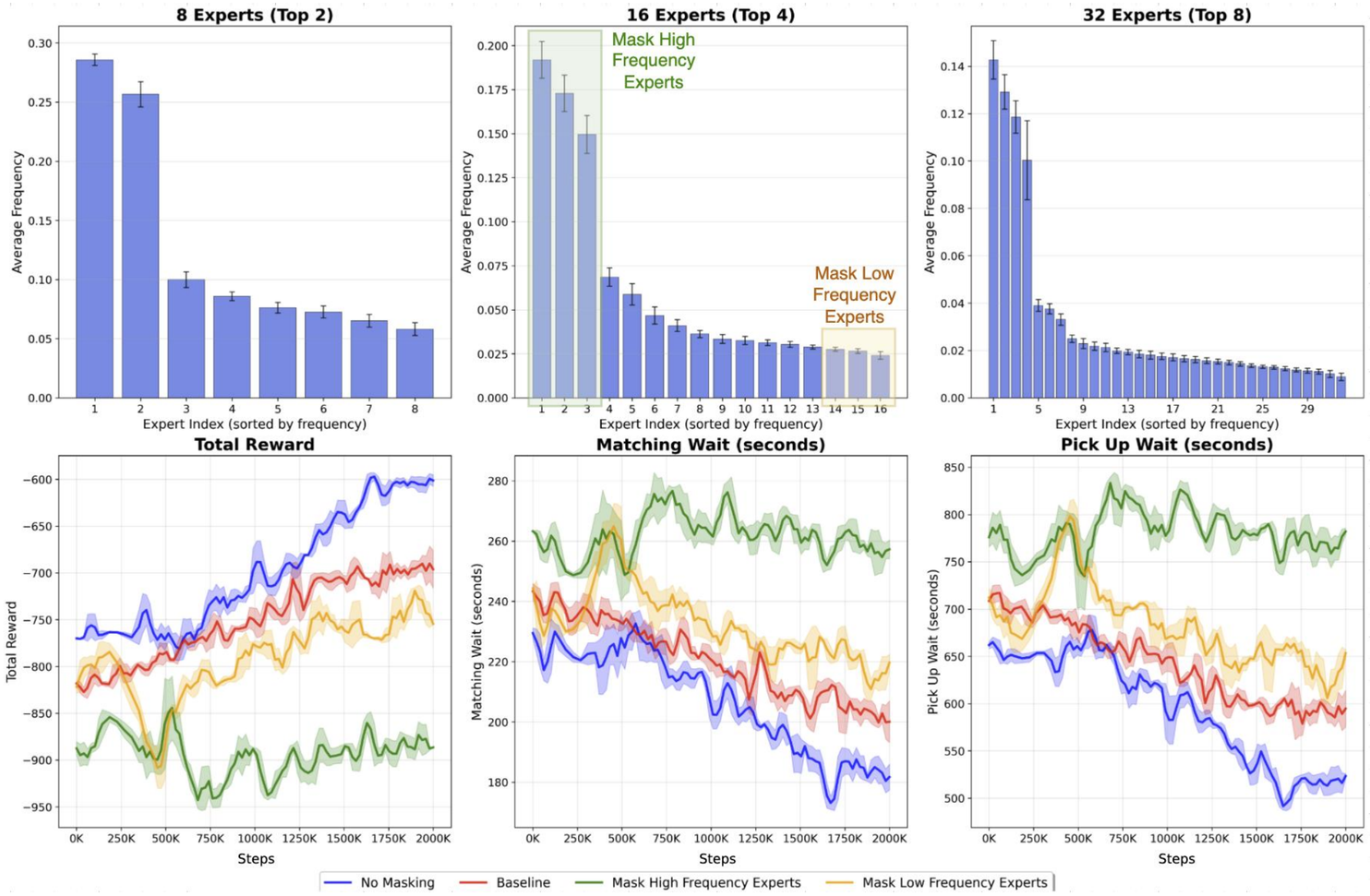


Expert activation heatmap

RESULTS

Importance of Individual Experts

Are the experts genuinely specializing? Every expert turns out to be irreplaceable.



Expert masking

RESULTS

Cross-City Generalization

Application of policy trained in SF city to San Diego and Alameda County reveals good generalization of the policy.

Alameda County

within 3%

	SF → ALA ZERO-SHOT	ALA → ALA NATIVE
Matching wait (s)	196	191
Pickup wait (s)	595	578

San Diego

within 1.5%

	SF → SD ZERO-SHOT	SD → SD NATIVE
Matching wait (s)	205	208
Pickup wait (s)	626	617

The SF-trained policy lands within **1.5–3%** of a model trained natively in each city. The regime-aware representation captures something general about urban ride-hailing.

Cross-city transfer

CONCLUSION

Three Contributions

01 **A regime-aware environment.**

Adaptive delayed matching as a spatio-temporal MDP, with a physics-based travel-time surrogate fast enough for millions of rollouts.

02 **An adaptive reward for model training.**

An online Lagrangian multiplier tightens service constraints in real time, stabilizing long-horizon training.

03 **A compact 12M-parameter MoE encoder.**

Experts specialize across regimes, beat every baseline on both waits, and generalize zero-shot across cities.

Regime awareness — in how you model the world and represent its state likely extends to any large-scale spatio-temporal resource allocation problem.

CORRESPONDENCE

xuanj@mit.edu

haozeh@cs.cmu.edu