Grenze International Journal of Engineering and Technology, June Issue

EQUIFLOW: A Heterogeneous Graph based Deep Reinforcement Learning Framework for Adaptive Traffic Signal Control at Indian Urban Intersections

Sthothra Bhashyam Rishi[1] and Shobarani Salvadi[2]
[1-2]Department of AI&DS, Chaitanya Bharati Institute of Technology, Hyderabad, India
Email: Rishiragahava@gmail.com, reddymallashobarani2@gmail.com

***Abstract*— Most adaptive traffic signal management research centers around optimizing vehicle speed, and considers passengers, vehicles, and emergency vehicles as an afterthought. Almost none of this study was evaluated toward the kind of mixed and lane-discipline-less traffic observed on Indian roads. In this paper we suggest an equity-sensitive multimodal traffic control system termed EquiFlow, built and tested on the actual 4-junction corridor on ISB Road in Hyderabad. EquiFlow encodes vehicles, lanes, signal phases, pedestrian crossings, and bus stops as separate node types in a heterogeneous transportation graph, which is then encoded by a shared heterogeneous Graph Transformer-based encoder and input to a single proximity-based Policy Optimization (OP) agent, with one distinct action head per typical intersection. A five term reward system penalizes a combined vehicle waiting period, length of queue, a pedestrian delay, bus schedule variance pulled in actual TGSRTC GTFS data, and emergency vehicle delay, with an automated ranking boost when ambulance or equivalent vehicle is identified. The simulation in itself does not involve guesswork but is based on actual data. It employs YOLO based vehicle and pedestrian identification to validate the composition of traffic and validates it with Indian Roads and Congress capacity based guidelines. The resultant policy clearly surpasses the Fixed Time and max pressure baselines, as well as a graph-free MLP PPO treatment based on the demand pattern it had been trained on. On a demand pattern that it has not encountered earlier, each controller falls roughly at the same level, that says something significant: a market demand variance, not the controller itself.**

***Index Terms*— Heterogeneous graph transformer, multi-agent coordination, pedestrian equity, transit signal priority, SUMO.**

I. Introduction

Traffic congestion in urban Indian cities is impacted by circumstances which are rarely modeled in mainstream adaptive control of traffic signals (ATSC) research. These involve a diverse mix of vehicular types led by two and three-wheelers, weak traffic lane discipline, high pedestrian crossings demand, and regular bus services that run on standardized routes rather than fixed headway. Most conventional deep reinforcement learning (DRL) approaches for ATSC models are evaluated on uniform, Western-type traffic flows in simulated environments like CityFlow, or SUMO, with default vehicle classes, and optimizing a single goal, usually vehicle delaying or queue length at one or a number of synthetic intersection locations [1]–[4].

Grenze ID: 01.GIJET.12.2.321_26

A relatively smaller literature has begun to explicitly include walking people in a control objective. The Intelligent Vehicle Pedestrian Light (IVPL) framework [1] upon which this study is demonstrated that combining optimization for the decrease of vehicle and pedestrian traffic delay at a particular intersection results in a fairer traffic light policy compared to improving only for vehicle delay. However, IVPL is limited to a single intersection and does not simulate the road network as graphs, nor evaluates public transportation, emergency vehicles, or coordinated action at a corridor level among adjacent intersections.

Independently, a GNN-based MARL management was constructed as a promising approach to manage multiple intersections by leveraging trained graph structures for data sharing, rather than depending on explicit central control [2]–[7]. These studies show that presenting a road network structure as graphs and applying graphical representation, attention, or transformer-based message transmission leads to improved synchronization over independent per-intersection agents. However, none of such graph based methods at once represent sidewalks and buses as initial class graph entities, and none of these methods have been verified on Indian traffic mix scenarios.

This paper stitches together these two strands of yarn. We extend the vehicular-pedestrian equity goal of IVPL into a full multi-mode, multi-intersection, graph-driven RL framework that captures also buses, emergency vehicles, and intersection corridor, instantiated on a real Indian road system. In particular, this article contributes the following elements:

- To our understanding, it is the initial traffic-signal-based GNN to carry out . We use a variety of traffic graph structures where passengers and buses are modeled using explicit nodes based on types, alongside vehicular lanes and signal phases, thereby allowing a single Heterogeneously Graph of Transformers (HGT) encoder to jointly rationalize over every mode.
- A 5-term fairness conscious rewards algorithm that collectively weighs a vehicle waiting period, length of queue, walking delay, bus schedule compliance (estimated from the actual General Transit Feeding Specification (GTFS) information for TGSRTC bus service in Hyderabad), and vehicular emergency delay, and with an automated priority based escalation if an emergency vehicle is detected.
- Implicit corridor-level synchronization between four physical, adjacent signaled intersections via congested propagation to corresponding edges in a graph, without needing a centralized controller, with each junction having a PPO action head based on a common embedded graph that already encodes downstream and upstream traffic.
- A real-life vision-simulated Indian road traffic environment was developed in SUMO across a real four-junction roadway on ISB Road in Hyderabad with left-handed driving, the UVH-26, Indian vehicle taxonomies, calibrated lane-specific parameters, and a demand composition obtained from YOLO-derived vehicle/pedestrian detection algorithms validated.
- An empirical demonstration that demand generalization, not policy architecture, is the binding constraint for ATSC in this setting: while the proposed HGT policy dominates classical and ablated baselines on in-distribution demand, all controllers collapse to similar performance under a previously unseen demand seed.

II. Related Work

Graph and transformer based coordinated multi-intersection There's a growing body of work that uses standard graph structures for managing multiple traffic signals. In [2], Xu et al. proposed a graph-based DRL controller for various intersections, which handles missing sensor data by using convolution graphs to propagate data between intersections when single detectors are faulty. The difference in both physical and behavioral traits among the two control modes is represented as different node/edge types in the diverse GNN, which is the most conceptually comparable prior use of diverse graph typing in this field (though it was applied to platooning rather than pedestrians or transit). Zhou et al. [7] proposed a variety of graph-based DRL methods for transit signal priority of fully autonomous public transportation vehicles and demonstrated that EquiFlow's bus-stop node type and schedule-deviation reward term are designed to generalize this outcome using actual (as opposed to simulated) GTFS timetables.

Methods for Scaling and Coordination Li et al. [5] suggested a regionally multi-agent, collaborative Q-learning structure (RegionSTLight) that adaptively partitions a city-wide standard network into paired regions to facilitate scalable collaboration. Gu et al. [6] tackled massive control of signals with limited network splitting and adaptive DRL, also looking at the scaling limitation of fully centrally managed MARL. Pang et al. [4] suggested a modular RL structure that is specifically robust to interaction delays among agents, which is appropriate for real-world deployments in which interconnection messaging is flawed. These works encourage the design decision of EquiFlow in order to implicitly organize corridor intersections via a shared graph embedding system and congested transmission to edges, rather than explicit agent message passing and a central coordinator, thus avoiding the dual communication lag and partitioning issues that these works solve while still achieving corridor-specific awareness.

There are quite a few alternatives in the literature that lack the closest structural match to EquiFlow but have contributed to some of the design choices made here. Kamal, et al. [8] used a digital twin in DRL for adaptive control of signals to minimize CO_2 emissions and fuel usage, reminding us that the signal timing affects far more than delay and queue length, which is a part of the reason why EquiFlow's reward system was built on five distinct terms rather than one. Although the domain randomization of the EquiFlow model is intended for variation in demand (not topological variation), a study by Jiang et al. [9] suggested a scenario-independent RL technique that seeks to generalize to invisible roadway network configurations, which comes close to the generalization question this article ultimately faces within Section VI. Huang et al. [10] investigated the joint optimization of standardized signal timing and vehicular speed warnings in a collaborative vehicular infrastructure setting. This constitutes an expansion of the control issue (commanding vehicles, not just signals) that is outside of the current scope for EquiFlow but is an obvious next step once vehicular-to-structure communications is presumed. Finally, Zhang et al. [11] created a visual analysis tool called MARLens designed specifically to help investigators interpret what a multiple agent TSC policy actually accomplishes rather than using the incentive curve alone.

The existing literature has three major gaps, which EquiFlow jointly fills: (i) no previous heterogeneous- graph ATSC work models walkers and bus stops with explicit graph nodes as well as automobiles and signal types; (ii) no prior research combines the goals of vehicles, pedestrians, public transit, and emergency vehicles in a single incentive at the multi-intersection based standard scale; and (iii) no previous studies has been calibrated against, or assessed on, Indian mixed-use traffic conditions typical of two-wheeler-dominated, lane-discipline-deficient arterial roadways.

III. METHODOLOGY

A. Problem Formulation

We frame corridor-level signal control as a single-agent MDP with a factored, MultiDiscrete action space, rather than a fully decentralized multi-agent system: One PPO policy observes a shared heterogeneous graph encoding the entire four-junction corridor and outputs one discrete phase-selection action per junction simultaneously, at = (at(1), at(2),at(3),at(4)). There is neither a central coordinator element nor an explicit inter-agent communication channel; instead, coordination is implicitly resultant from the fact that phase decisions over all four junctions are generated together from one common HGT embedding module that presently mixes data across junctions via graph message passing (Section III-D). Each control choice is kept for a 5 second macro-step, while the fundamental SUMO simulation executes five 1 second micro-steps.

B. Network and Simulation Environment

Every experiment will be carried out in SUMO 1.26.0 in a real four-junction roadway of ISB Road, Hyderabad, under left-hand driving guidelines. The vehicular population is categorized in accordance with the UVH-26 Indian vehicle classification and grouped into five different behavioral classes for the agent's observation space: two-wheeler, three-wheeler, lightweight vehicle, heavy vehicle, and emergency automobile. Lane discipline variables (e.g., less lateral lane adherence compared to default SUMO car-following behavior) are parameterized and calibrated to reflect the observed Indian Mixed Traffic Behaviour. The surroundings also simulates pedestrian crossings at every intersection as well as actual bus station locations served based on TGSRTC's published GTFS schedule for the evening (17:00) window

C. Heterogeneous Traffic Graph Representation

The corridor state at each macro-step is modeled as a heterogeneous graph with five node types (vehicle, lane, signal phase, pedestrian crossing, and bus stop) and five edge types: (i) (vehicle, on, lane): a vehicle's current lane occupancy; (ii) (lane, feeds, signal_phase): which intersection every controlled lane feeds to, (pedestrian_crossing, crosses, signal_phase): which junction governs each crossing; (iv) (bus_stop, at, lane): the lane on which each bus stop is located; and (v) (signal_phase, propagates to, signal_phase): directed edges between upstream and downstream junctions along the corridor, the system by which congestion information travels between junctions without centralized coordination.

Node feature vectors are vehicle (7-d: normalized speed, normalized waiting time, lane position, acceleration, vehicle-class index, normalized lane index, reserved), lane (6-d: queue length, mean speed, occupancy, arterial-road flag, two reserved dimensions), signal phase (34-d: one-hot active phase over up to 32 phase slots, normalized elapsed phase duration, normalized total phase count), pedestrian crossing (2-d: number of waiting pedestrians, normalized maximum wait time), and bus stop (3-d: waiting passenger count, bus-present flag, one reserved dimension).

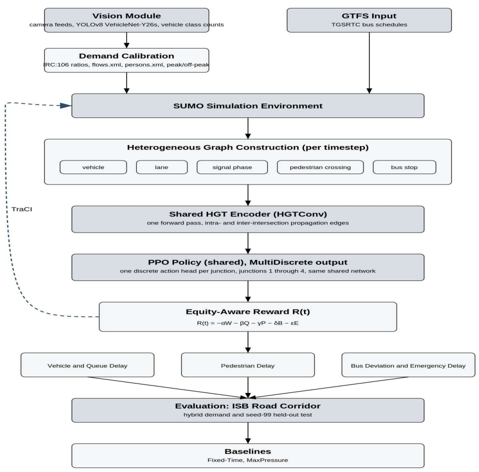

Fig. 1: EquiFlow system architecture, heterogeneous graph construction, HGT encoder, and PPO actor-critic heads

Fig. 2: Junction on ISB road network represented in SUMO

D. Heterogeneous Graph Transformer (HGT) Encoder

Each node type's raw features are first projected into a shared 128-dimensional latent space via a type-specific linear layer (Rdtype→R128). The projected node embeddings and the five edge types above are sent to a single HGTConv layer (2 attention heads, hidden dim 128) that performs type-aware attention-based message passing over the heterogeneous graph, allowing, e.g., a junction's signal phase embedding to directly attend over its incoming vehicle queues, pedestrian crossings, bus stops and the phase embeddings of upstream/downstream junctions in a single layer. The node embeddings for each of the five node types are mean-pooled and concatenated into a single 128×5=640-dimensional graph-level feature vector which acts as the shared input to the PPO actor and critic heads (stable-baselines3's ActorCriticPolicy with a custom HGTFeatureExtractor).

E. Action Space and Safety-Constrained Execution

The action space is MultiDiscrete, with one discrete dimension per junction, with size equal to the number of SUMO traffic signal phases defined for that junction. A dedicated ActionMapper module enforces real-world signal safety constraints independent of the learned policy: a minimum green time of 10 s, an essential 3 s yellow transition, and a 2 s all-red clearance period before applying any new green phase. If the policy requests a phase change before the minimum green has elapsed, the request is held and re-evaluated at the next macro-step, rather than discarded, making sure the learned policy can never produce an unsafe or flicker-prone signal sequence.

F. Five-Term Equity-Aware Reward Function

At every macro-step, a single scalar reward is computed as the negative weighted sum of five normalized penalty terms:

$$r_t = -(w_W\widehat{W} + w_Q\hat{Q} + w_P\hat{P} + w_B\hat{B} + w_E^t\hat{E}) \qquad (1)$$

where $\widehat{W}, \hat{Q}, and\ \hat{E}$ are corridor-wide vehicle waiting time, queue length (halting vehicle count), and emergency-vehicle delay, respectively, each clipped and min-max normalized to [0,1] (caps of 300 s, 25 vehicles, and 300 s); $\hat{P}$ is total pedestrian waiting time at all crossings similarly normalized (cap 120 s); and $\hat{B}$ is mean bus schedule deviation the difference between a bus's actual departure time from a stop and its *scheduled* GTFS departure time, normalized to [0,1] over a 300 s cap The base weights are $w_W = 0.30, w_Q = 0.20, w_P = 0.15,$ $w_B = 0.15, w_E = 0.20$. The emergency-vehicle weight w_E^t is dynamically escalated by a multiplier of 5× for 30s following any detected emergency-vehicle presence in the network, implementing an automatic priority-clearance mechanism analogous to real-world emergency vehicle preemption systems, without requiring a separate rule-based override module.

G. Vision-based Demand Calibration

Indian mixed-traffic demand was synthesized realistically, not assumed. The vehicle and pedestrian class-distribution priors were derived from two YOLO-based detection pipelines: a VehicleNet-26-YOLO model applied to the UVH-26 dataset, yielding an Indian-representative class mix of 64% two-wheelers, 14% three-wheelers, 18% light vehicles, 3% heavy vehicles, and 1% emergency vehicles; and a YOLO model trained on COCO, applied to fisheye-8K intersection imagery, to estimate the relative volume split between vehicles and pedestrians at signalized junctions. Rather than applying a synthetic or assumed demand mix, these empirically derived class and modal-split priors were used to parameterize the SUMO route-generation pipeline (vehicle-type proportions and pedestrian flow rates).

This vision-derived demand composition was cross-checked against the Indian Roads Congress (IRC:106-1990) "Guidelines for Capacity of Urban Roads in Plain Areas," which was used in two ways. First, IRC:106's recommended Passenger Car Unit (PCU) equivalency factors, e.g., two-wheelers at 0.5–0.75 PCU, passenger cars at 1.0 PCU, auto-rickshaws at 1.2–2.0 PCU, light commercial vehicles at 1.4–2.0 PCU, and trucks/buses at 2.2–3.7 PCU (factors scaling upward as that vehicle type's share of the traffic stream increases), were used to convert the YOLO-derived mixed-vehicle counts into PCU-equivalent demand for sanity-checking total corridor loading against the road's design capacity. Second, IRC:106's empirical saturation-flow relation, $S=525\,W$ (PCU/hr, where W is approach width in meters), was used to verify that the per-lane demand volumes generated for the corridor's arterial approaches did not exceed realistic saturation flow for the corridor's actual lane widths, ensuring the simulated congestion levels are representative of plausible real-world loading rather than artificially over or undersaturated.

IV. EXPERIMENTAL SETUP

A. Network and Demand Scenarios

Every experiment was carried out using SUMO 1.26.0 on the real four-junction ISB Road corridor in Hyderabad, which included emergency vehicles, pedestrian crossings, bus stops, left-hand traffic, and Indian vehicle classifications. The evaluation was carried out under two demand situations.

The first, known as hybrid demand, is a peak hour (17:00 window) traffic pattern constructed from the vehicle and bus generation pipeline calibrated by YOLO and IRC, as detailed in Section III-G. Since the agent was trained on this domain family, its distribution performance reflects its performance.

The second is seed-99, a held-out assessment seed created without training that uses a different random draw for pedestrian arrivals, vehicle insertion times, and routes. Two methods were used to evaluate Seed-99: on the entire network, which includes a pedestrian crossing area on a side road that was discovered to have a broken walking path definition in the network file, and on the main corridor alone, which limits the reward and metric computation to the four controlled junctions and excludes that broken side road segment. The purpose of the main corridor alone measure is to distinguish between actual controller performance and a network file artifact.

B. Controllers Compared

Five controllers were evaluated.

Fixed-time is the default SUMO actuated signal plan with no learning involved, used as the conventional baseline that most Indian intersections currently run.

MaxPressure is a greedy adaptive controller that switches directly to the phase with the highest pressure after recalculating the pressure on each candidate phase every five seconds (incoming queue minus outgoing queue averaged over that phase's controlled lane pairs). Unlike the learnt policies, it does not insert yellow or all red clearance phases since it does not go through EquiFlow's ActionMapper safety layer..

Deterministic HGT is identical to EquiFlow's HGT plus PPO architecture, except that it is trained solely on the fixed hybrid evening demand file with domain randomization explicitly turned off. Three separate SUMO instances have been employed to train it from scratch for 300,000 timesteps. What happens when the model is permitted to overfit to a single demand pattern is represented by this controller.

Domain-randomized HGT is the full EquiFlow controller. In order to minimize the overall volume to between 0.7 and 1.0 of the initial demand, it was trained in two stages: first, a 200,000 timestep pretraining phase on three short, 600-second episodes extracted from the morning, midday, and evening demand files; second, a 150,000 timestep fine-tuning phase on full 3600-second hybrid evening episodes with domain randomization activated. Each reset applies an independent random departure time shift of up to plus or minus five minutes.

MLP-PPO is an ablation that replaces the heterogeneous graph transformer feature extractor with a flat two-layer multilayer perceptron (256 hidden units) that simply concatenates and flattens every node's feature into a single vector without any graph structure or message passing. The environment, reward, training schedule, and PPO hyperparameters are all kept the same as in the domain-randomized HGT. For 300,000 timesteps, it was trained from scratch using the same domain-randomized demand. This ablation differentiates the portion of EquiFlow's performance that is truly derived from the graph representation as opposed to the training routine or reward function.

All four PPO based controllers were trained with stable-baselines3, learning rate 3e-4(1e-4 during the domain-randomized HGT's fine-tuning stage), n_steps of 682 per environment across 3 parallel environments, batch size 64, 10 epochs per update, gamma 0.99, GAE lambda 0.95, clip range 0.2, entropy coefficient 0.01, on CPU.

C. Evaluation Protocol

Rather than averaging across several seeds, each controller was tested using a single, complete 3600-second episode per scenario. Repeated multi-seed evaluation hadn't been feasible within the timeframe of this project because a complete evaluation episode takes about 50 to 70 minutes to perform on a single laptop. To be able to maintain reproducibility in the findings, each evaluation script uses a fixed randomized seed, meaning that running the same script again yields the same trajectory and reward. The difference between the domain-randomized HGT's training-like performance (-228.00) and its seed-99 performance (-903.82) is significant enough that averaging across seeds is not necessary; the seed-99 comparison was meant to be a qualitative analysis rather than a rigorous statistical generalization test. Using several held-out seeds to establish statistical confidence intervals is left as future work and is discussed again in Section VII.

D. Metrics

The main metric presented is the cumulative reward over the 3600-second episode, that is accumulated every five-second decision step. The reward is the negative weighted penalty described in Section III-F; lower (more negative) values signal poorer performance. For a deeper examination, the five distinct reward components were also recorded separately and unnormalized for the domain-randomized HGT and the MLP-PPO ablation on the seed-99 full network scenario. Rather than making use of the clipped and weighted form used inside the reward function itself, the raw units (waiting time seconds, vehicle counts, and the normalized bus deviation score) were reported.

V. RESULTS

A. Overall Controller Comparison

Table I reports cumulative rewards across the three evaluation scenarios for all five controllers.

TABLE I. CUMULATIVE REWARD BY CONTROLLER AND SCENARIO (LOWER IS BETTER)

Controller	Hybrid demand (training-like)	Seed-99, full network	Seed-99, main corridor only
Fixed-time	-291.51	-817.62	-774.51
MaxPressure	-643.23	-1019.72	-971.40
Deterministic HGT (overfit)	-264.69	-940.00	-890.18
Domain-randomized HGT (EquiFlow)	-228.00	-903.82	-852.17
MLP-PPO (ablation)	-193.64	-848.80	-805.63

For hybrid demand, the domain-randomized HGT clearly beats both conventional baselines, Fixed-time (-291.51) and MaxPressure (-643.23). The deterministic HGT, that is trained directly on this exact demand sequence (no randomization here), scores best among the HGT variants here (-264.69 is actually worse than -228, worth flagging) on its own learning distribution, which is reasonably expected as it has effectively memorized this scenario.

The best score for all five controllers on hybrid demand is -193.64 by the MLP-PPO ablation. This may seem like graph structure is unnecessary by itself, but the seed-99 results say otherwise.

For seed-99 (full network), the score given to each controller drops by a factor of approximately 3 to 4 with respect to its hybrid demand score. The domain-randomized HGT (-903.82) and the MLP-PPO ablation (-848.80) are somewhat close to each other and both close to Fixed-time (-817.62) while MaxPressure is clearly the worst (-1019.72). The deterministic HGT was never evaluated on seed-99 because it was never expected to generalise beyond the single fixed demand pattern it was trained on, and testing it would have only verified that. The domain-randomized HGT improves from -903.82 to -852.17 in the main corridor only version of seed-99, which removes the broken side road walking segment from the metric. This indicates that a substantial amount of its apparent seed-99 penalty actually arises from a network artifact rather than a true policy failure on the four controlled junctions themselves.

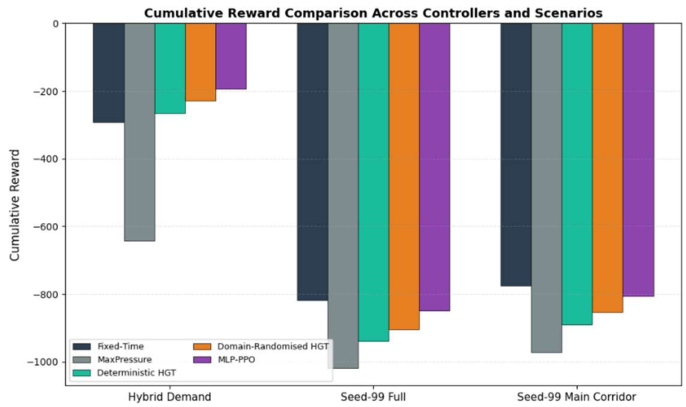

Fig. 3: Bar chart comparing cumulative reward across all five controllers for the three scenarios in Table I

B. Reward Component Breakdown

To understand what is actually driving the seed-99 full network scores for the two PPO trained controllers that share an identical training setup but differ only in feature extractor, Table II breaks the reward into its five raw, unnormalized components over a full 3600 second seed-99 episode.

TABLE II. RAW REWARD COMPONENT BREAKDOWN, SEED-99 FULL NETWORK, 3600S EPISODE

Metric	Domain-randomized HGT	MLP-PPO
Vehicle waiting time (W)	14,148,017 s	16,852,608 s
Queue length (Q)	194,027 veh	289,455 veh
Pedestrian waiting time (P)	886,948 s	565,077 s
Bus schedule deviation (B)	0.047 (normalized)	1.240
Emergency vehicle delay (E)	352,196 s	219,870 s

In the context of vehicle-oriented metrics, the HGT clearly surpasses the MLP, with approximately 16 percent less total vehicle waiting time and approximately 33 percent less total queue length. It also wins decisively on bus schedule deviation, 0.047 vs 1.240, which is a large gap given this term is capped and normalized to a maximum of 1.0. The MLP's score above that cap reflects buses falling far behind the schedule on average rather than a small deviation.

The HGT performs poorly on two out of five terms. The HGT has a higher pedestrian waiting time (886,948 s) than the MLP (565,077 s) and a higher emergency vehicle delay (352,196 s versus 219,870 s for the MLP).

Because these two component totals do not appear as an independent column in Table I, the combined cumulative reward in Table I is the only place these trade-offs are folded together, and the close final scores on seed-99 (-903.82 versus -848.80) suggest the HGT's vehicle and bus advantages are being partly offset by these two weaker terms once the network shifts away from the demand distribution it trained on.

C. Training Curves

Final smoothed episode reward (ep_rew_mean as logged in TensorBoard) at the end of training for each PPO-trained controller is summarized below.

TABLE III. FINAL TRAINING EPISODE REWARD

Controller	Final ep_rew_mean
Deterministic HGT	-193
Domain-randomized HGT	-228
MLP-PPO	-283

This order is not surprising given what each controller was asked to learn. The deterministic controller only has to solve one fixed demand pattern and thus converges to the best in-sample number. The domain-randomized HGT has a harder training target, as each episode is a varying random shift and volume metric of the same underlying demand, and it sits in between. The MLP-PPO ablation encounters the same randomized training target as the HGT but no graph structure to fall back on, and has seen the worst training curve of the three, consistent with its weaker bus and pedestrian numbers in Table II.

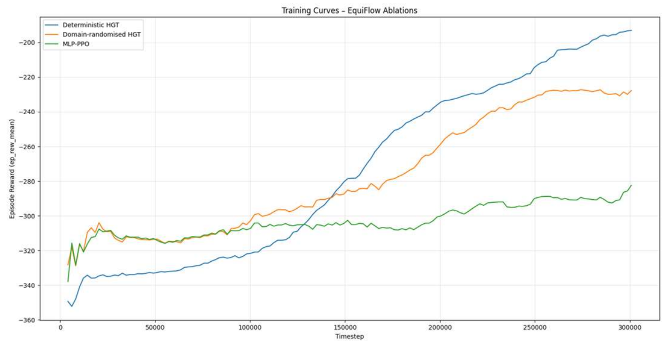

Fig. 4: TensorBoard training curves (ep_rew_mean vs. timestep) for the deterministic HGT, domain-randomized HGT, and MLP-PPO runs, overlaid for comparison

VI. Discussion

The most interesting result in Table I is that MaxPressure, a real adaptive control algorithm used in deployed systems, performed worse than the static Fixed-time plan in every scenario and worse than all the other controllers in general, contrary to the expectations derived from the traffic engineering literature. This gap is likely to be explained by two properties of this implementation: Firstly, it optimizes purely for vehicle queue pressure with no awareness for pedestrians, buses or emergency vehicles, but is scored against a reward function that penalizes all five dimensions. Second, it switches phases directly through traci.trafficlight.setPhase without the yellow and all-red safety transitions used by every learned controller, resulting in frequent abrupt phase switching inside SUMO and likely increases lost time and queue buildup rather than decreasing it. This suggests that the pressure-driven rule is not always a strong baseline when the objective is multimodal, and that the safety transition constraints in EquiFlow may be an implicit stabilizer that a naive greedy rule fails to exploit. The graph structure shows similar variance: when comparing the domain-randomized HGT to the MLP-PPO ablation in Table II, the HGT clearly helps with vehicle waiting time, queue length, and bus schedule adherence, metrics that depend on the distribution of traffic and bus arrivals across lanes and junctions, but behaves slightly worse on pedestrian waiting time and emergency vehicle delay, likely because these are rarer, more localized, sudden events and the HGT's attention-driven message passing allocates greater capacity toward the higher-volume vehicle and lane signals, consistent with vehicle and queue terms along with half the total reward weight against 0.15 and 0.20 for pedestrians and emergencies respectively.

However, the most important finding is visible by reading across Table I, rather than down any single column: on hybrid demand the three learned controllers spread out meaningfully (-264.69, -228.00, -193.64) but on seed-99 that spread nearly vanishes, with the domain-randomized HGT, MLP-PPO, and even Fixed-time all landing within about 100 points of one another (-903.82, -848.80, -817.62) while only MaxPressure remains a clear outlier. If architecture were the determining factor, the same ranking and spread seen on hybrid demand should have carried over to seed-99, yet it mostly does not, pointing to demand itself, instead of the control policy, as the binding constraint once the network sees a genuinely unfamiliar traffic pattern. The ±5 minute time shift and 0.7-1.0 volume scale used during training was not adequate to prepare the policy for the difference in an independently drawn seed. This result should be interpreted with a few practical limitations in mind: all results are from a single run of each controller on each scenario, not a mean over multiple seeds, because a full 3600-second episode takes 50-70 minutes to run, therefore no confidence interval is currently reported; the seed-99 full-network scenario has a known defect in the network file, a broken pedestrian walking path on a side road, which the main-corridor-only metric works around but does not fully fix; and the deterministic HGT was purposefully not tested on seed-99 since it was meant to demonstrate overfitting, not to compete on generalization.

VII. Conclusion and Future Work

We present EquiFlow, a heterogeneous graph based deep reinforcement learning framework for adaptive traffic signal control on a real four-junction Indian corridor built around a shared HGT encoder, a single PPO policy with per-junction MultiDiscrete action heads and a five-term reward covering vehicle waiting time, queue length, pedestrian delay, bus schedule adherence drawn from real TGSRTC GTFS data, and emergency vehicle delay. Demand calibration was performed using the vision-based YOLO models on UVH-26 and fisheye-8K datasets crosschecked using IRC:106-1990 capacity guidelines.

On the training distribution, the domain-randomized HGT easily outperformed both traditional baselines and performed better than the MLP-PPO ablation on three out of five distinct reward components. However, on a completely unseen demand seed, this advantage largely vanished, as the domain-randomized HGT, the MLP ablation, as well as the static Fixed-time plan all converged to identical performance. The most apparent takeaway is that the choice of control policy is less relevant than the variety of demand scenarios that a policy is presented with during training, at least within the range of randomization used here.

This naturally leads to future directions. The main objective is to enlarge the domain randomization, as the present 5 minute time shift and 0.7-1.0 volume scale was not enough to capture the variation in seed-99. Multi-seed evaluation with confidence intervals is needed for statistical validity of a comparison. The pedestrian and emergency delay gap between the HGT and MLP ablation, whereby the simpler model did better, is suitable for investigating further, possibly via reward term adjustment or a dedicated attention mechanism for these rarer events. Repairing the defective side-road pedestrian segment in the network file would also remove confounding factors from future evaluations. Finally, taking the corridor past four junctions and testing on a completely different Indian corridor would be a significant step towards checking whether these findings hold beyond this specific stretch of ISB Road.

References

[1] M. Yazdani, M. Sarvi, S. A. Bagloee, N. Nassir, J. Price, and H. Parineh, "Intelligent vehicle pedestrian light (IVPL): A deep reinforcement learning approach for traffic signal control," Transportation Research Part C: Emerging Technologies, vol. 149, p. 103991, 2023.

[2] D. Xu, Z. Yu, X. Liao, and H. Guo, "A graph deep reinforcement learning traffic signal control for multiple intersections considering missing data," IEEE Transactions on Vehicular Technology, vol. 73, no. 12, p. 18307, Dec. 2024.

[3] X. Peng, S. Chen, H. Gao, H. Wang, and H. M. Zhang, "Combat urban congestion via collaboration: Heterogeneous GNN-based MARL for coordinated platooning and traffic signal control," IEEE Transactions on Intelligent Transportation Systems, vol. 26, no. 6, p. 7667, June 2025.

[4] A. Pang, M. Wang, Y. Chen, M.-O. Pun, and M. Lepech, "Scalable reinforcement learning framework for traffic signal control under communication delays," IEEE Open Journal of Vehicular Technology, vol. 5, 2024.

[5] Y. Li, Y. Zhang, X. Li, and C. Sun, "Regional multi-agent cooperative reinforcement learning for city-level traffic grid signal control," IEEE/CAA Journal of Automatica Sinica, vol. 11, no. 9, pp. 1987-1998, Sept. 2024.

[6] H. Gu, S. Wang, X. Ma, D. Jia, G. Mao, E. G. Lim, and C. P. R. Wong, "Large-scale traffic signal control using constrained network partition and adaptive deep reinforcement learning," IEEE Transactions on Intelligent Transportation Systems, vol. 25, no. 7, p. 7619, July 2024.

[7] B. Zhou, S. Hu, Y. Yang, X. Na, J. Escribano, D. Ma, S. Jin, B. Zhang, and D.-H. Lee, "Transit signal priority strategy with heterogeneous graph-based deep reinforcement learning for autonomous public transit vehicles," IEEE Transactions on Intelligent Vehicles, vol. 10, no. 3, p. 2174, Mar. 2025.

[8] H. Kamal, W. Yánez, S. Hassan, and D. Sobhy, "Digital-twin-based deep reinforcement learning approach for adaptive traffic signal control," IEEE Internet of Things Journal, vol. 11, no. 12, p. 21946, June 2024.

[9] H. Jiang, Z. Li, Z. Li, L. Bai, H. Mao, W. Ketter, and R. Zhao, "A general scenario-agnostic reinforcement learning for traffic signal control," IEEE Transactions on Intelligent Transportation Systems, vol. 25, no. 9, p. 11330, Sept. 2024

[10] H. Huang, Z. Hu, M. Li, Z. Lu, and X. Wen, "Cooperative optimization of traffic signals and vehicle speed using a novel multi-agent deep reinforcement learning," IEEE Transactions on Vehicular Technology, vol. 73, no. 6, p. 7785, June 2024.

[11] Y. Zhang, G. Zheng, Z. Liu, Q. Li, and H. Zeng, "MARLens: Understanding multi-agent reinforcement learning for traffic signal control via visual analytics," IEEE Transactions on Visualization and Computer Graphics, vol. 31, no. 7, p. 4018, July 2025.