Path-Aware Heterogeneous Graph Reinforcement Learning for Green Light Optimal Speed Advisory

Wei Zhao[1]*, Lingchao Wang[2], Qiang Ai[3]
College of Information Engineering, Liaoning Institute of Science and Engineering, Jinzhou, China[1,2]
School of Information Science and Engineering, Lanzhou University, Lanzhou, China[3]

***Abstract*—Green light optimal speed advisory (GLOSA) needs to jointly balance stopping, emissions, travel efficiency, and advice-update overhead. To address the difficulty of jointly optimizing the path traffic state and update frequency in multi-intersection GLOSA, this paper proposes a path-aware heterogeneous graph reinforcement learning method for green-wave speed advice. A heterogeneous traffic graph composed of road segments and signalized intersections is constructed; relation-specific graph attention and a remaining-path readout are used to extract the downstream traffic state of vehicles; and twin delayed deep deterministic policy gradient (TD3) is adopted to jointly output the advised speed and the next update interval. Experiments are conducted on SUMO, with a manual review of 456 unique run records. The results show that, compared with the no-advice baseline, the full model reduces the number of stops and CO_2 emission intensity by 48.0% and 4.2%, respectively; compared with the fixed 5 s update strategy, the advised update frequency is reduced by 13.9%, and the three independently trained models all exhibit consistent trends. The study indicates that the proposed method can effectively fuse vehicle-path and signal-interaction information, improve the traffic operation state while reducing the advice-update frequency, and provide a co-optimization approach for multi-intersection green-wave speed advice.**

***Keywords*—Green Light Optimal Speed Advisory (GLOSA); Reinforcement Learning; Heterogeneous Graph; Graph Attention Network; Speed Advisory; Semi-Markov Decision Process**

I. Introduction

With the continuous growth of urban vehicle ownership and travel demand, stopping, idling, and frequent acceleration/deceleration caused by signalized intersections have become important factors affecting the efficiency of urban road operation and vehicle energy consumption. Green Light Optimal Speed Advisory (GLOSA) provides vehicles with reasonable travel speeds so that they can pass through intersections within the green-time window as much as possible, thereby reducing unnecessary stops and accelerations/decelerations. In recent years, GLOSA research has gradually shifted from traditional rule-based or optimization methods toward data-driven and intelligent decision-making. Yang *et al.* [1] applied deep reinforcement learning to eco-driving near signalized intersections, showing that reinforcement learning can simultaneously consider safety, efficiency, energy consumption, and comfort under limited traffic information. Qin *et al.* [2] further considered the overall energy consumption of platoons and used TD3 for adaptive eco-driving of connected vehicles, extending speed decisions beyond single-vehicle energy consumption to account for cooperative behavior among vehicles. These studies show that reinforcement learning can provide a more flexible continuous-speed decision mechanism for GLOSA.

However, when a vehicle passes through multiple signalized intersections consecutively, its speed decisions are strongly correlated. The passage state and speed at the current intersection affect the spatial correlation during the continuous passage process. Recent studies have begun to extend the research object from a single intersection to consecutive signalized intersections. Chen and Wang [3] used SAC to establish an eco-driving strategy for consecutive intersections, enabling vehicles to pass through two signalized intersections continuously and reduce stopping behavior; Xue *et al.* [4] further considered SPaT uncertainty and mixed-traffic environments, incorporated stochastic signal conditions into eco-driving control, and adopted variable-step stochastic model predictive control to improve real-time performance. Meanwhile, Vignarca *et al.* [5] incorporated actual signal data, bus dwelling times, and timetables into GLOSA control. These studies indicate that GLOSA is gradually evolving from "current-intersection speed advice" toward spatiotemporal decision-making oriented to the continuous passage process.

Beyond traffic efficiency and energy consumption, the actual effect of GLOSA is also influenced by vehicle motion characteristics and control strategies. Chaudhry *et al.* [6] investigated the safety of different GLOSA implementations and found that GLOSA may have different safety impacts under varying connected-vehicle penetration rates and mixed-traffic conditions, indicating that speed advice must coordinate with the vehicle's motion state rather than only considering the signal passage window. Therefore, for multi-intersection GLOSA, the controller needs to perceive the road traffic state, signal state, and the vehicle's future travel direction simultaneously, forming a state representation that reflects the actual traffic topology.

In recent years, graph reinforcement learning has become an important research direction for complex traffic control; consequently, many scholars have attempted to use graph neural networks (GNNs) to address the above issues. Wang *et al.* [7] proposed a multi-layer graph deep reinforcement learning method that extracts the geometric and spatial features of intersections by constructing traffic graphs at different levels, and uses the graph structure to improve the model's adaptability across different road networks. Lin *et al.* [8] further proposed a temporal-difference-aware graph convolutional reinforcement learning method for multi-intersection traffic signal control, strengthening spatial correlation and temporal

*Corresponding author.

sensitivity in the traffic network by using lanes as graph nodes and introducing time-dependent rewards. Devailly *et al.* [9] used graph-based model reinforcement learning for large-scale traffic signal control and verified the potential of graph reinforcement learning to transfer across different road networks and traffic conditions.

As the interactions among different traffic entities and relations in the traffic network become more complex, homogeneous graphs gradually fail to fully describe the interaction process among roads, vehicles, and signalized intersections. Jiang *et al.* [10] proposed a scenario-agnostic reinforcement learning framework that maps intersections of different structures into a unified state and action space to improve the cross-scenario adaptability of the policy; Zhu *et al.* [11] constructed a coordinated-graph-based framework for network traffic signal control, improving multi-agent decision-making by explicitly modeling the cooperative relations among intersections. Further, some scholars introduced heterogeneous graphs into multi-intersection control. Zhai *et al.* [12] constructed a heterogeneous graph attention network that incorporates different types of traffic objects into a unified representation and combines it with multi-agent reinforcement learning for cooperative control; Yoon *et al.* [13] used graph representation to build a distributed, communication-driven multi-agent traffic signal control model, enabling agents to achieve cooperative decisions using network spatial correlations. These studies show that distinguishing traffic entity types and their relations can further enrich the traffic state representation and provide more targeted state information for reinforcement learning decisions in complex traffic networks.

With deeper research, the focus of reinforcement learning is shifting from static topological representation toward dynamic spatial relations and broader spatiotemporal dependence. Wang *et al.* [14] introduced spatiotemporal hypergraphs into multi-agent traffic signal control, describing complex relations among multiple intersections through high-order relations; Ye *et al.* [15] further constructed dynamic heterogeneous graphs and spatiotemporal memory-augmented models, enabling the spatial relations in the traffic network to adjust with traffic states; Yang and Zeng [16] adopted heterogeneous graph distributional reinforcement learning in 2026 to address the out-of-distribution generalization problem in traffic signal control, improving the model's generalization across different traffic networks by enlarging the heterogeneous-graph receptive field and introducing distributed value representation. The development trend of these studies indicates that traffic control models need not only local-neighborhood information but also distal states with decision significance extracted according to the actual spatial relations of the controlled objects.

For GLOSA, vehicles have explicit travel directions and remaining paths, and relying only on local-neighborhood information makes it difficult to fully reflect the influence of downstream road and signal states on the current speed decision during consecutive-intersection passage. Therefore, it is necessary to incorporate the vehicle's remaining path into the traffic state representation to enhance the model's ability to characterize directional spatiotemporal dependence. Meanwhile, existing GLOSA research mainly focuses on speed advice and pays little attention to the advice-update timing, whereas the update interval directly affects the policy's responsiveness to traffic-state changes and the advice-update frequency. Therefore, how to combine the vehicle's downstream path information and further coordinate speed decisions with the update interval is a worthwhile issue for multi-intersection GLOSA.

Based on this, to provide a path-aware reinforcement learning method for GLOSA decision-making that balances vehicle passage performance and advice-update efficiency in a fixed-signal multi-intersection environment, this paper proposes a green-wave speed advice method based on path-aware heterogeneous graph reinforcement learning, extending the reinforcement learning modeling of multi-intersection GLOSA from two aspects: path-aware state modeling and joint speed–update-interval decision-making. The main contributions of this paper are threefold:

- Construct a road-segment–signalized-intersection heterogeneous traffic graph, and extract interaction features among different traffic entities through a relation-specific graph attention mechanism;
- Design a path readout mechanism oriented to the vehicle's remaining travel direction, incorporating downstream road and signal states into the current decision representation;
- Establish a joint action space of advised speed and update interval, use TD3 to learn continuous control policies under different traffic states, and comprehensively evaluate the method from the perspectives of stopping, operation efficiency, traffic emissions, and advice-update frequency.

II. Control Method

This paper adopts a green-wave speed advice framework composed of traffic-state encoding, path-information aggregation, and joint action decision-making. The road-network encoder extracts traffic features under different road relations; the path readout module further aggregates segment information relevant to the ego vehicle's subsequent travel; the policy network then determines the advised speed and the next update interval. The first two parts construct the spatial state representation required for control, and the last part weighs vehicle operation performance against the number of advice updates. The signal timing is kept fixed in the experiments, and the controller does not adjust signal phases or green durations. The overall framework and the actual road network are shown in Fig. 1. In Fig. 1, (a) The five-dimensional features of road segments and signalized intersections are projected to a 32-dimensional space, encoded by two layers of relation-specific graph attention, and then aggregated through global mean readout and remaining-path attention readout to obtain a 32-dimensional global representation and a path representation, respectively; the two are concatenated with the 7-dimensional ego-vehicle observation to form a 71-dimensional state. The policy network jointly outputs the advised speed upper bound and the next appointment update interval, executed by SUMO. The dashed line indicates the observation feedback at the next effective decision. The path readout selects at most four segments after the current segment along the predetermined route and does not change the travel route; all inputs are current states, and the signal timing remains fixed. The cycle and green-ratio channels in the signal input use the historical

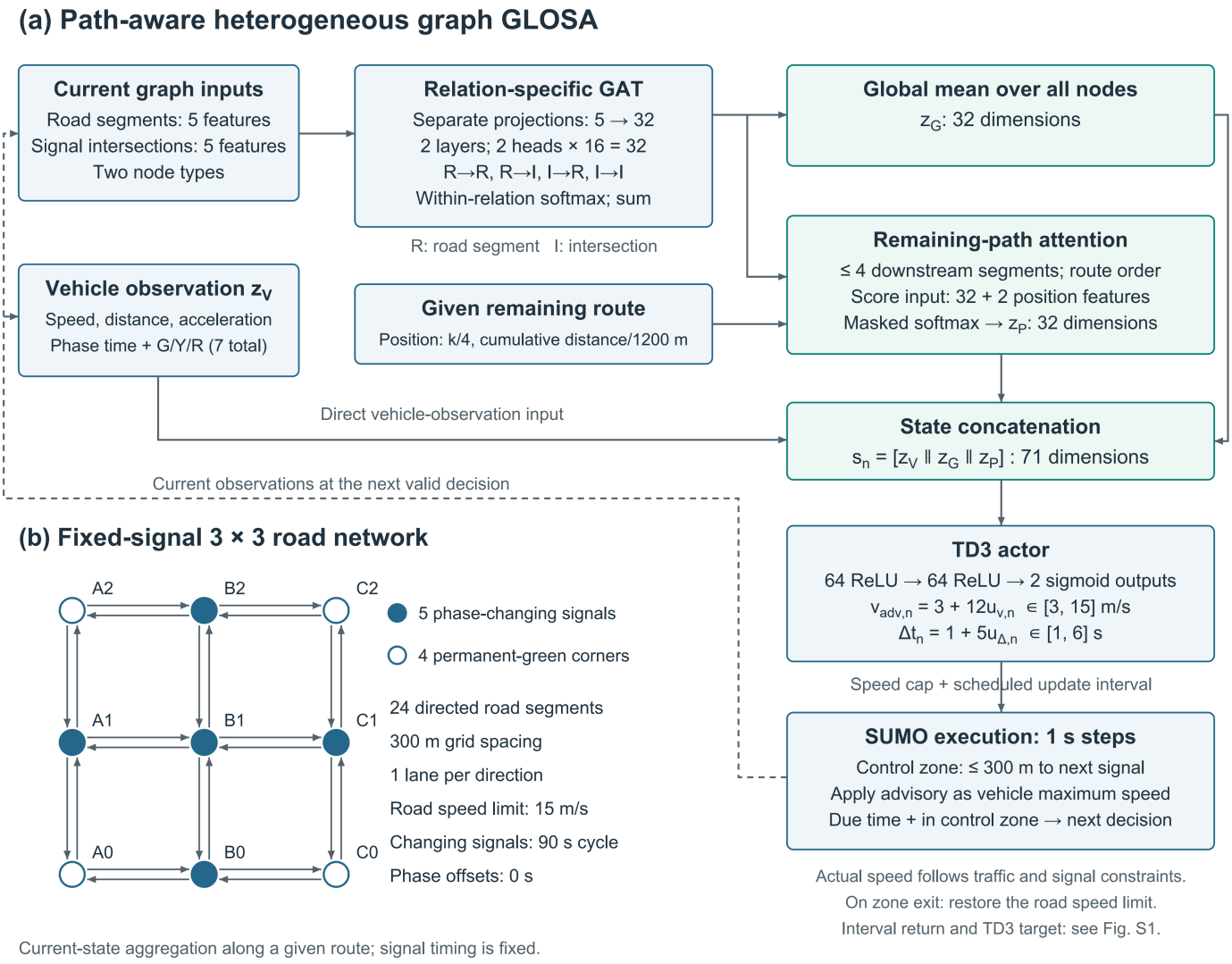

Fig. 1. Framework of path-aware heterogeneous graph green-wave speed advice and the simulation road network.

default values 90 s and 0.5. The masked-attention illustration corresponds to the case where a valid downstream segment exists; see Section II-C for the historical handling defect of completely empty paths. (b) The 3×3 grid road network contains 24 directed segments, with an intersection spacing of 300 m and a single lane in each direction. Solid nodes denote the five phase-changing signals, and hollow corner nodes denote the four permanent-green signals; the changing signals have a cycle of 90 s and a phase offset of 0. Road arrows indicate the actual passage direction and do not represent all message-passing relations in the heterogeneous graph.

A. Road-Network Representation and Vehicle Observation

To represent both segment operation states and intersection signal constraints simultaneously, the road network is constructed as a heterogeneous graph containing directed-segment nodes and signalized-intersection nodes. Each segment uses density, mean speed, stopped-vehicle ratio, number of vehicles, and length as inputs, describing traffic load, operation speed, queuing state, and spatial scale, respectively. Except for the stopped-vehicle ratio, the remaining features are normalized with reference scales of 50 veh/km, 15 m/s, 20 veh, and 300 m, respectively.

Intersections use a five-dimensional input, including normalized phase index, the time ratio to the next phase switch, cycle-length ratio, green-ratio channel, and red-signal-group ratio. Among them, the phase index identifies the current stage in the signal program, the remaining phase time reflects the near-future signal change, and the red-signal-group ratio describes the range of currently restricted passage at the intersection. When the historical implementation reads the static network without loading the complete signal program, the cycle and green-ratio channels use default values of 90 s and 0.5, respectively. Therefore, these two channels do not provide information about real timing differences among intersections, and the green-ratio channel is not equivalent to the actual green-to-cycle ratio in the ego vehicle's passage direction.

To supplement the ego-vehicle passage conditions that the network-aggregation features cannot directly express, a 7-

dimensional vehicle observation is set up, consisting of current speed, distance to the next signal, acceleration estimated from the 1 s speed difference of the adjacent second, remaining phase time, and three-dimensional indicator variables for green, yellow, and red. The light color is read according to the ego vehicle's next signal connection, so that the observation corresponds to the ego vehicle's actual passage direction rather than merely reflecting the phase number of the whole intersection. All inputs come from the current time step; the path-aware representation in this paper aggregates the current traffic state along the downstream spatial position and does not include prediction of future traffic states.

The control range is within 300 m upstream of the next signal. After the vehicle enters this range, the policy output is enforced through the vehicle's maximum speed upper bound; the actual speed is still jointly determined by car-following, acceleration/deceleration capability, and signal passage rules [8]. Therefore, the advised speed is a control input, not a direct assignment of the vehicle's actual speed. This implementation can be used to evaluate the impact of speed advice on vehicle operation, but it cannot replace an independent traffic-safety verification. Since the aggregated observation does not contain all vehicle interactions and historical information, this paper treats it as an approximate state representation for semi-Markov control and does not assume that it strictly satisfies the Markov property.

B. Relation-Specific Graph Attention Encoding

Different types of road relations have different meanings. For example, the upstream-segment state can reflect entering traffic, while the intersection signal state can reflect downstream release conditions. To avoid processing this information entirely with the same set of parameters, the encoder distinguishes four relations: segment-to-segment, segment-to-intersection, intersection-to-segment, and adjacent-intersection, and learns neighbor weights within each relation.

The two types of nodes are first mapped to a 32-dimensional representation space through their respective linear projections, so that different types of raw features can participate in subsequent message passing. The encoder contains two graph-attention layers, each using two attention heads, and each head outputs 16-dimensional features; after concatenation, the dimension is maintained at 32. For the q-th attention head of the ℓ-th layer, the node update is written as

$$h_i^{(\ell+1,q)} = \mathrm{ELU}\left(W_{\mathrm{self}}^{(\ell,q)} h_i^{(\ell)} + \sum_{r\in\mathcal{R}} \sum_{j\in\mathcal{N}_i^r} \alpha_{ij}^{r,\ell,q} W_r^{(\ell,q)} h_j^{(\ell)} \right), \tag{1}$$

and the outputs of the two heads are concatenated to form the next-layer node representation:

$$h_i^{(\ell+1)} = \left[h_i^{(\ell+1,1)} \parallel h_i^{(\ell+1,2)} \right]. \tag{2}$$

where h_i denotes the embedding of node i, $\mathcal{R}$ is the relation set, and $\mathcal{N}_i^r$ is the neighbor set pointing to node i under relation r; W_r is the relation-specific message transformation matrix, and W_{self} is the self-node transformation matrix shared across relations within the same layer and the same attention head. ELU is the activation function. The results of each attention head are concatenated along the feature dimension to obtain the next-layer input. For simplicity, the following formula omits the superscripts of the layer and attention head; the relation-internal attention is defined as

$$b_{ij}^r = \mathrm{LeakyReLU}\big((a_{r,s})^\top W_r h_j + (a_{r,d})^\top W_r h_i \big), \tag{3}$$

$$\alpha_{ij}^r = \frac{\exp(b_{ij}^r)}{\sum_{u\in\mathcal{N}_i^r} \exp(b_{iu}^r)}. \tag{4}$$

where $a_{r,s}$ and $a_{r,d}$ are the attention parameters of the source node and the target node, respectively. Equation (4) is normalized only within the in-neighbors of the same target node under the same relation; the aggregation results of each relation are then added. Therefore, this module can distinguish relation parameters but does not additionally set cross-relation semantic attention weights. Its design draws on the ideas of graph attention and heterogeneous-graph representation [3], [4], and is not a direct reproduction of the meta-path-based hierarchical attention structure in [4].

After the two-layer propagation, the arithmetic mean of the embeddings of all segment nodes and intersection nodes is computed to obtain a 32-dimensional global representation $\bar{z}_G$, which is used to summarize the current road-network operation background. The global mean is the same for all vehicles; therefore, the ego-vehicle observation and the path representation must be combined to distinguish the control requirements of different vehicles.

The historical segment adjacency is constructed from the consistency of the endpoint of the previous segment and the start point of the next segment, without further filtering by actual feasible turns. This relation expresses topological proximity and does not guarantee that messages propagate only along feasible vehicle travel connections. The inter-intersection relation also only expresses spatial adjacency and does not represent signal coordination control. The homogeneous-graph counterpart merges the four types of edges and shares message-passing parameters, but retains the independent input projections of the two node types. Therefore, this counterpart examines the relation-distinguishing manner rather than whether node-type information is used.

C. Remaining-Path Attention Readout

The global mean may weaken the local differences directly relevant to the ego vehicle's subsequent travel. To enable the policy to focus on the segments the ego vehicle will pass through, at most four downstream segments are extracted after the current segment, forming a path sequence according to the current predetermined route. This module does not select or change the travel route.

For the k-th valid path position, the corresponding segment's final embedding is concatenated with a two-dimensional position feature. The position features are, respectively, the path index divided by 4, and the cumulative distance from the first downstream segment divided by 1200 m, used to distinguish the relative order and spatial scale of segments. The path scoring and readout are

$$e_k = w^\top \tanh(W_p[h_{r_k} \parallel p_k] + b_p), \tag{5}$$

$$\beta_k = \frac{\exp(e_k)}{\sum_{u\in\mathcal{P}} \exp(e_u)}, \qquad z_P = \sum_{k\in\mathcal{P}} \beta_k h_{r_k}. \tag{6}$$

where r_k is the k-th downstream segment, p_k is its position feature, and $\mathcal{P}$ is the set of valid path positions; W_p, b_p, and w are trainable parameters, and β_k is the normalized attention weight within the valid positions. z_P maintains 32 dimensions, providing a downstream state representation of the ego vehicle distinct from the global mean. Its role is to selectively aggregate current-path information and cannot directly infer the causal effect of traffic factors from the attention weights.

When the path is shorter than four segments, a mask is applied to invalid positions. The historical implementation has a defect in handling completely empty paths: the padded positions may introduce the embedding of the zeroth segment. The relationship between this situation and the maintenance-version fix and the historical results is given in Section VI-C. This paper retains the implementation description corresponding to the reported experiments and does not retroactively treat the fixed zero-vector handling as the actual behavior of the historical model.

Finally, the 7-dimensional ego-vehicle observation z_V, the global representation $\bar{z}_G$, and the path representation z_P are concatenated to form the input to the policy and value networks:

$$s_n = [z_V \parallel \bar{z}_G \parallel z_P] \in \mathbb{R}^{71}. \tag{7}$$

The full model input is 71-dimensional; after removing the path readout it is 39-dimensional. This design enables the controller to receive the ego-vehicle signal passage condition, the whole-network traffic background, and the remaining-path information simultaneously; whether the module improves vehicle operation performance is tested by the subsequent comparison experiments.

D. Joint Decision of Speed and Update Interval

A fixed update interval uses the same decision frequency for all states, making it difficult to balance traffic-change responsiveness and the number of advice events simultaneously. This paper treats the next update interval and the advised speed as actions together, so that the policy can select both the speed upper bound and the subsequent update time at each decision. Existing AF-GLOSA research used hybrid-action reinforcement learning to control advice updates [11], [12]; this paper examines the specific implementation that combines the path graph representation with the speed and appointment-interval outputs.

The two outputs of the policy network are mapped to the normalized action space through a sigmoid and then linearly scaled:

$$a_n = (v_{\mathrm{adv},n},\, \Delta t_n) = (3 + 12\, u_{v,n},\; 1 + 5\, u_{\Delta t,n}), \tag{8}$$

$$u_{v,n},\, u_{\Delta t,n} \in [0, 1]. \tag{9}$$

where $u_{v,n}$ and $u_{\Delta t,n}$ are the normalized outputs of speed and interval, respectively; the unit of $v_{\mathrm{adv},n}$ is m/s and that of Δt_n is s. The speed range is 3–15 m/s, and the selected interval range is 1–6 s. This lower bound restricts the policy advice; the vehicle can still decelerate to a stop due to a red light or the constraint of a preceding vehicle.

The simulation advances with a 1 s step. The next decision is executed only when the appointment update time is reached and the vehicle is still in the control zone. Therefore, the continuously output interval is quantized to integer seconds in actual execution; when the vehicle remains in the control zone, according to the triggering rule, the execution interval usually corresponds to the ceiling of Δt_n in seconds. After leaving the control zone, the vehicle restores the road speed limit, but the reward of the current transition continues to accumulate until the next effective decision or vehicle termination. Let the actual duration between adjacent decisions be T_n, which may be larger than the appointment interval.

This joint action-and-time modeling requires computing the interval return and the subsequent value discount with the actual duration, rather than using a single-step discount for all decisions [6]. Meanwhile, integer-second execution makes the true control response in the interval direction piecewise, and the continuous TD3 is an approximate optimization for this action dimension; accordingly, this paper interprets the interval learning as approximate control with continuous parameterization and discrete-time execution.

E. Multi-Objective Interval Return

To form an explicit trade-off between vehicle operation and the number of advice events, the reward considers travel progress, continuous operation time, stopping, acceleration, and CO_2 emission simultaneously, and charges a decision cost at each advice occurrence. The reward per simulation second is

$$r_t = 0.02\, v_t - 0.10 - 5\, I_{\mathrm{stop},t} - 0.02\, a_t^2 - 0.005\, \dot{m}_{\mathrm{CO}_2,t}. \tag{10}$$

where v_t is the actual speed, a_t is the acceleration estimated from the 1 s speed difference, and the CO_2 emission rate is in g/s. The stopping indicator variable is 1 when the speed changes from greater than 0.1 m/s to not greater than 0.1 m/s, and 0 otherwise. The speed term provides a travel-progress incentive, the constant term counts the continuous-operation cost, the stopping term suppresses new stopping events, the squared-acceleration term suppresses large accelerations/decelerations, and the emission term counts the environmental cost. The acceleration penalty is not equivalent to a jerk penalty, so an improvement in comfort metrics cannot be presumed from it. The coefficients are reward weights under specified units, used to set the optimization preference, and do not correspond to calibrated economic costs or fixed exchange rates among metrics.

The interval return corresponding to the n-th decision is

$$R_n = -\lambda_{\mathrm{dec}} + \sum_{k=1}^{T_n} \gamma^k\, r_{t_n+k}. \tag{11}$$

where $\gamma = 0.995$ is the per-simulation-second discount factor and $\lambda_{\mathrm{dec}} = 0.03$ is the cost per decision. Equation (11) preserves the historical implementation's convention of discounting traffic rewards from $k = 1$, and the decision cost is charged at the start of the interval. A longer interval reduces the cost per unit time while possibly delaying the response to traffic-state changes; the two jointly affect the policy choice. The actual duration includes the default-driving phase after leaving the control zone, so the interval return evaluates the advised action and its subsequent environment evolution, not

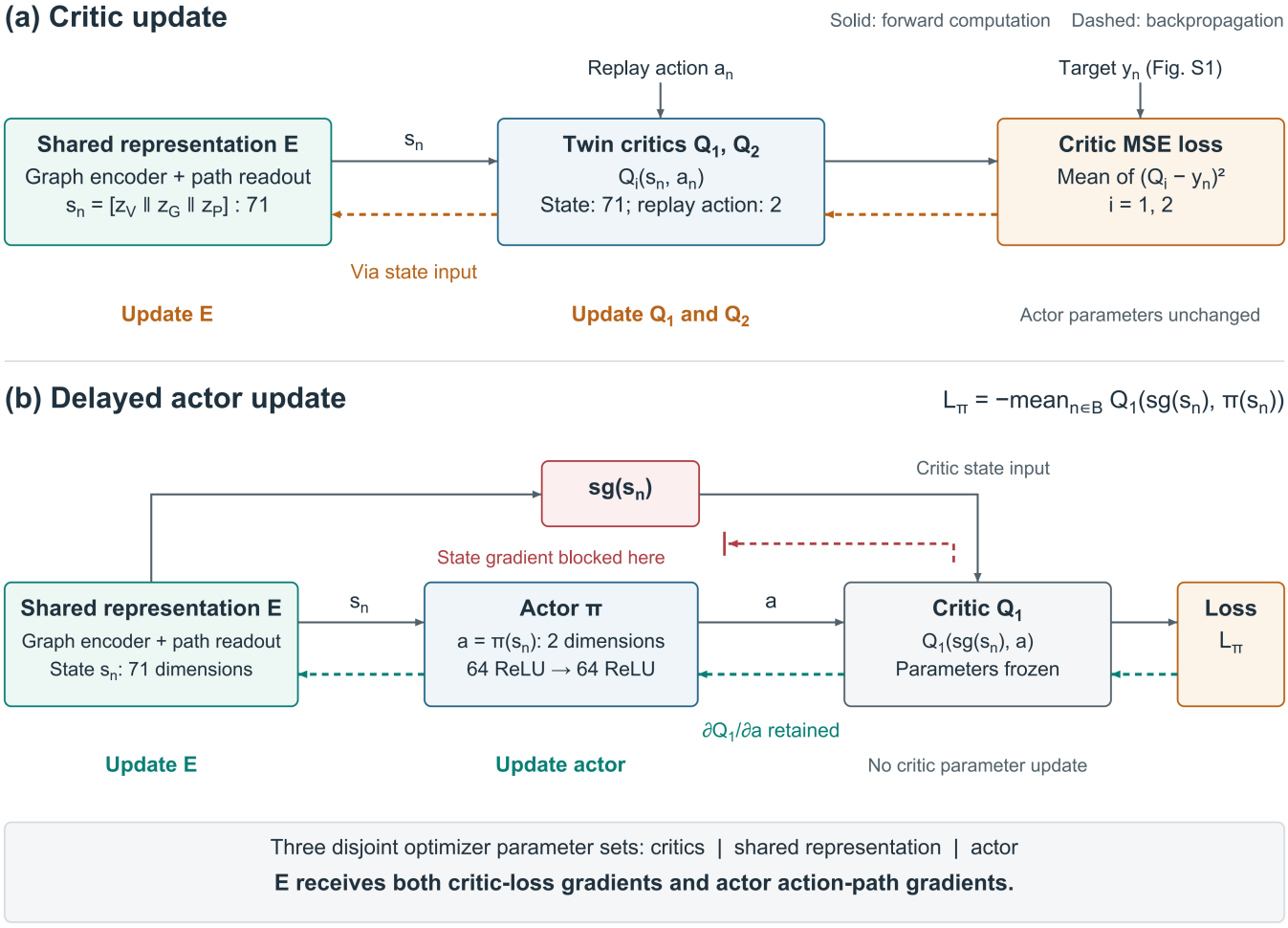

Fig. 2. Gradient paths of shared-representation updates in value and policy updates.

only the phase where the advised speed upper bound is continuously applied.

To adjust the return scale of transitions with different durations, historical training also used a return baseline updated with the batch. Define the discounted-duration coefficient H_n, and update the baseline ρ according to the replay batch:

$$H_n = \frac{1-\gamma^{T_n}}{1-\gamma}, \tag{12}$$

$$\rho \leftarrow \rho + 0.01\Big(\frac{1}{|\mathcal{B}|}\sum_{n\in\mathcal{B}}\frac{R_n}{H_n} - \rho\Big), \tag{13}$$

$$\tilde{R}_n = R_n - \rho H_n. \tag{14}$$

where $\mathcal{B}$ is the sampled batch and 0.01 is the baseline update step. R_n already includes the decision cost, so ρ represents the mean estimate of the interval return scaled by H_n, rather than the average of pure traffic instantaneous rewards. This operation changes the reward level fed into value learning; its design intent is numerical centering, but there is no independent ablation proving its impact on training performance.

When ρ is fixed and the total duration is fixed, the H_n term discounted at the decision time can be merged into a quantity independent of the interval division; when the vehicle termination time varies with the policy, this correction generally still changes the return comparison among different policies. In addition, ρ introduces non-stationarity of the reward target as training updates. Therefore, this paper treats this operation as empirical training and does not endow it with a general policy-invariance guarantee.

F. TD3 Optimization and Shared-Representation Update

The advised speed and appointment interval use continuous-parameter outputs, so they are trained with the twin delayed deep deterministic policy gradient algorithm, TD3 [5]. TD3 mitigates the impact of value-estimation error on policy learning through double value networks, target-action smoothing, and delayed policy updates. This paper uses the actual duration T_n in the target value to adapt to unequal-interval decisions:

$$y_n = \tilde{R}_n + \gamma^{T_n}(1-d_n)\min_{i=1,2} Q_i'(s_{n+1}, \tilde{a}_{n+1}). \tag{15}$$

where Q_1' and Q_2' are the target value networks and d_n is the termination indicator variable; the target action of the next state is generated by the target policy. The target noise is first

added and clipped in the normalized action space, restricting the action to $[0, 1]$ before mapping to physical units. This avoids applying the same numerical noise directly to actions in the two different units of speed and time.

The two value networks are trained through the mean-squared error between the target value and the predicted value. The shared representation is jointly generated by the graph encoder and the path module, and the value loss can be backpropagated to these representation modules through the state input. The policy loss uses the first value network to evaluate the current policy action; its computational form is

$$\mathcal{L}_\pi = -\frac{1}{|\mathcal{B}|} \sum_{n \in \mathcal{B}} Q_1\big(\mathrm{sg}(s_n),\, \pi(s_n)\big). \tag{16}$$

where sg denotes stop-gradient and π is the policy network. The stop-gradient acts only on the state input of the value network, aiming to block the path along which the policy loss directly changes the state representation; the policy action is still generated from the untruncated s_n, so the action path can still update the shared representation. This training arrangement cannot be described as "freezing the encoder during policy updates."

Both the policy network and the two value networks use two 64-unit ReLU hidden layers. Three optimizers with disjoint parameter sets separately manage the value networks, the shared-representation module, and the policy-output network. The shared representation can receive gradients from both the value loss and the policy-action path, while the policy-output parameters are updated only by the policy loss. This arrangement is the shared-representation training implementation of this paper and is not automatically guaranteed to be stable by standard TD3; the gradient paths are shown in Fig. 2, and the specific learning rates, delayed updates, and soft updates are given in Table II. Solid lines denote forward computation; dashed lines denote gradient backpropagation. (a) The value update computes the mean-squared-error loss of the two value networks using replay actions and target values; the gradient updates the value networks and the shared graph encoder and path readout module through the state input. (b) The delayed policy update applies a stop-gradient operation only to the direct state input of the first value network; the policy network still receives the untruncated state, so the policy loss can update the policy network and the shared representation along the action path. At this time the value-network parameters are not updated, but their derivatives with respect to the action are retained. Both the policy network and each value network contain two 64-unit ReLU hidden layers; three optimizers separately manage the disjoint parameter sets of the value networks, the shared representation, and the policy network. The raw ego-vehicle observation is a concatenated input and is not a trainable parameter.

In addition, the appointment-interval execution, interval-return centering, and duration-aware TD3 target construction are shown in Fig. 3. In Fig. 3, (a) The continuously output appointment update interval is executed on a 1 s simulation grid; the next decision is triggered only when the appointment time is reached and the vehicle is within the control zone. Therefore, the actual duration between consecutive effective decisions may exceed the appointment interval and may include the default driving phase after leaving the control zone; the time axis is not to scale. (b) The interval return charges the decision cost at the start and accumulates discounted traffic rewards from the first second. Historical training updates a baseline from the mean ratio of each sample's interval return to its discounted-duration coefficient within the batch, yielding a centralized return for value learning. (c) The target policy output first adds clipped noise in the normalized action space and is restricted to $[0, 1]$, then mapped to the physical units of speed and time. The TD3 target consists of the centralized return and the smaller estimate of the two target value networks; subsequent values are discounted by the actual duration and controlled by the termination flag whether to retain.

III. Experimental Setup and Statistical Analysis

A. Evaluation Data and Simulation Road Network

The evaluation data include 456 unique run records. The original file contains 474 rows in total; keeping the last record for each combination of method, demand level, and demand seed removes 18 duplicate items; the difference among duplicates is mainly at the floating-point precision level of emission calculation. The main comparison consists of seven methods, three demand levels, and eight demand seeds, for a total of 168 runs; the five learning variants each add two groups of training models, adding 240 runs; the nominal 45-episode checkpoints of the full model and the homogeneous-graph model each contain 24 evaluations, adding 48 runs. All results in this paper correspond to the above historical evaluation scope.

The road network is generated by netgenerate and netconvert as a 3×3 grid with a spacing of 300 m, a single lane in each direction, a road speed limit of 15 m/s, and 24 directed segments in total. Among the nine signal nodes, five use changing phases and four corner nodes maintain permanent green. The changing-phase signals have a cycle of 90 s and a phase offset of 0. This setting represents a speed-advice scenario under fixed-signal conditions and does not constitute an optimized arterial green-wave coordination scheme.

The three demand levels are 500, 1000, and 2000 veh/h. The demand-generation seeds are 11, 22, 33, 44, 55, 66, 77, and 88, and the SUMO internal random-number seed is fixed at 42. Therefore, the eight demand repetitions are mainly used to reflect variations in vehicle-demand generation and do not represent independent sampling of all random factors within the simulation.

B. Comparison Methods and Comparison Scope

Seven methods are set up: no advice, rule-based GLOSA, local MLP, fixed 5 s interval, homogeneous GNN, remove-path-readout, and full model, with the configuration given in Table I. The fixed-interval variant retains the heterogeneous graph and path readout and trains the speed policy independently, used to compare the results of allowing the learned interval with the fixed 5 s update. This comparison also includes differences in training implementation and cannot be interpreted as changing only the interval when the same policy is deployed.

The local MLP uses 11-dimensional local features, including the average speed of the next segment, and outputs

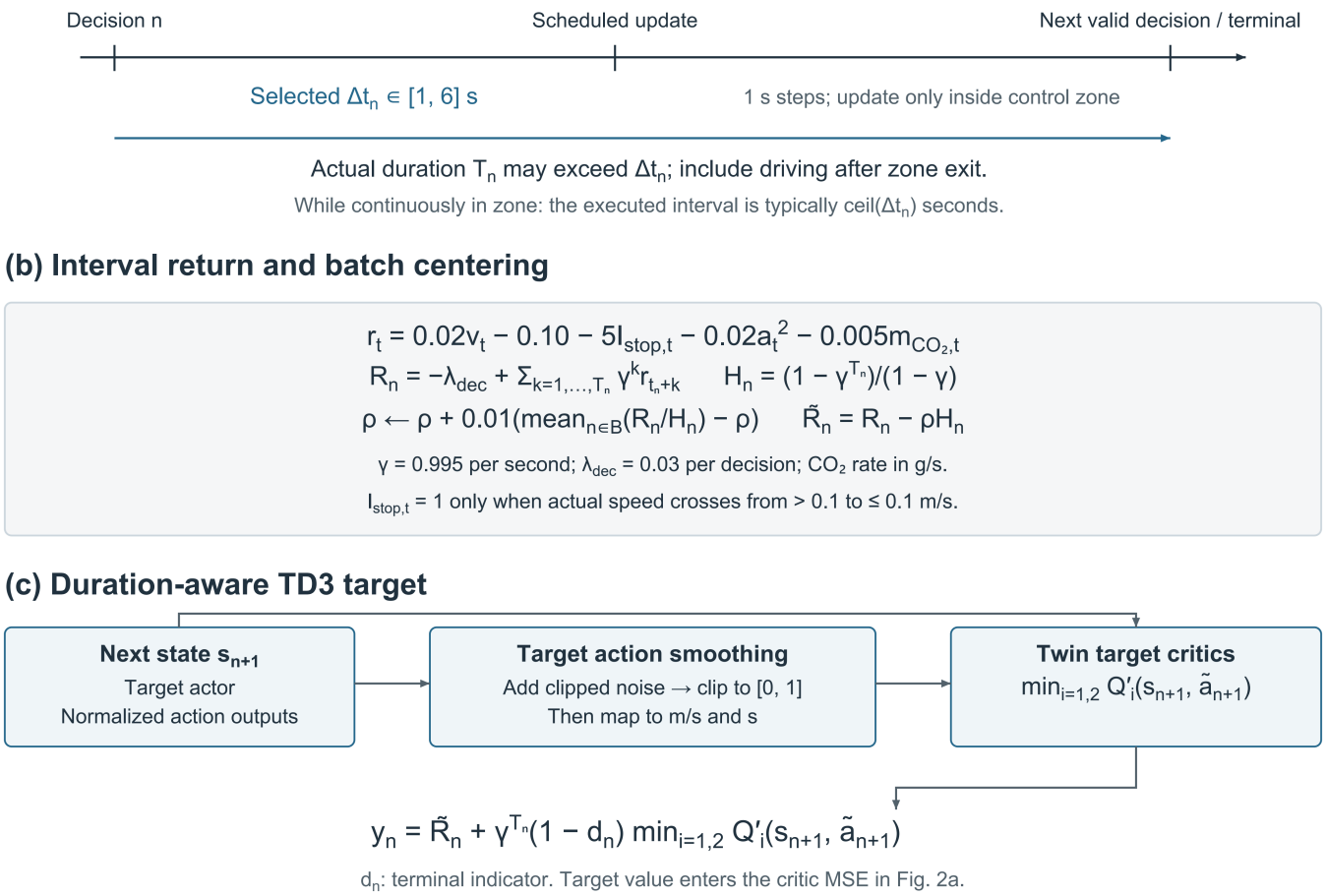

Fig. 3. Appointment-interval execution, interval return, and duration-aware TD3 target construction.

TABLE I. METHOD CONFIGURATION OF THE MAIN COMPARISON.

Method	Input & Representation	Update Interval
No advice (NA)	Native SUMO driving behavior	No advice event
Rule-based GLOSA	Nearest signal and ego arrival info	Fixed 5 s
Local MLP	11-dim local features, incl. next-segment mean speed	Fixed 5 s
Fixed 5 s	Heterogeneous graph & path readout	Fixed 5 s
Homogeneous GNN	Merged relations, type projection & path readout	Learned 1–6 s
Remove-path-readout	Heterogeneous graph & global mean	Learned 1–6 s
Full model	Heterogeneous graph & path readout	Learned 1–6 s

advice at a fixed 5 s interval. Because its observation range, representation structure, and action dimension all differ from the full model, this baseline is used to compare different control configurations and cannot serve as a strict single-factor graph-encoder ablation.

Rule-based GLOSA advises the maximum speed when the current green can be passed, otherwise computes the arrival speed from the remaining phase time and the default green ratio, restricted to 3–15 m/s. The historical implementation approximates the next switch after the current red or yellow ends as the start of green, but this correspondence does not apply to all signal connections. Therefore, this method represents only the rule-based baseline of this implementation and cannot represent a GLOSA controller that fully scans the future passable green window [1].

C. Evaluation Metrics and Vehicle-Record Scope

Metrics are computed for vehicles with a recorded time of no less than 5 s and a recorded distance of more than 50 m. The historical aggregation does not distinguish whether vehicles arrived completely; at the end of the 2400 s simulation, some vehicles that had not yet completed their trips may also satisfy the above inclusion criteria. To keep the metric names consistent with the actual data, the original ATT column is recorded as mean recorded duration $\bar{T}_{\text{obs}}$.

$$\bar{T}_{\text{obs}} = \frac{1}{N}\sum_{i=1}^{N} n_i, \tag{17}$$

$$C = \frac{1}{N}\sum_{i=1}^{N} \frac{M_i}{D_i/1000}, \tag{18}$$

$$\mathrm{DF} = \frac{1}{N}\sum_{i=1}^{N}\frac{N_i^{\mathrm{dec}}}{n_i/60}. \tag{19}$$

where C corresponds to the CO_2 intensity in the table and DF is the advice frequency. Both are first computed as intensity or frequency per vehicle and then averaged over vehicles. Therefore, CO_2 intensity is not the ratio of total network emissions to total distance, and DF is not the ratio of total network advice count to total vehicle time.

Mean speed AS is the average of each vehicle's recorded distance divided by recorded duration. The number of stops NS counts stopping events with a continuous speed below 0.1 m/s for at least 2 s, which is not entirely consistent with the instantaneous threshold-crossing events in the reward. Speed fluctuation SF is the average of the per-vehicle speed standard deviation. AV is the average of the absolute value of the second-order speed difference under 1 s sampling, corresponding to the mean absolute jerk, with unit m/s^3.

Fuel intensity FC is converted from fuel mass at a density of 0.75 g/mL, and energy intensity EC is further converted at 34.2 MJ/L. Therefore, EC is a linear derivative indicator of FC, and the two do not provide mutually independent energy-saving evidence. AS is usually better when larger, and NS, CO_2, FC, EC, SF, AV, and DF are usually better when smaller, but the trade-offs among metrics should be examined simultaneously. $\bar{T}_{\mathrm{obs}}$ only characterizes the length of the recorded interval, and its reduction cannot be directly interpreted as a faster full trip.

The historical calculation did not implement 120 s warm-up data removal nor empty the road network after demand stopped. The existing per-run summary data cannot reconstruct the arrival state or eliminate the right-censoring of records; complete trip time and completion rate require re-evaluation combined with per-vehicle arrival records. SUMO's TripInfo interface can record arriving-vehicle information and can enable incomplete-vehicle output [9].

D. Training Configuration and Continuation Manner

The policy is trained under the 1000 veh/h demand, with a main budget of 25 episodes and at most 1200 s per episode. All vehicles share the policy parameters and replay buffer; batch inference is only used to process multiple vehicle observations simultaneously and does not constitute per-vehicle independent learning. The training seeds of the first group of models are, in order, full model 101, remove-path 202, homogeneous graph 303, fixed interval 404, and local MLP 505; the latter two groups add 1000 and 2000 to the corresponding seeds, respectively. Because different variants do not use the same numerical seed, the three groups of results are compared as independent training realizations and are not interpreted as paired experiments that strictly share training random numbers.

The number of transitions is re-accumulated each episode, and parameter updates begin only after more than 2000 transitions; when training is not interrupted, the buffer is retained across episodes. This threshold specifies the start time of learning and does not mean that the first 2000 decisions use a uniform random policy; the historical implementation still generates actions with the current policy plus exploration noise.

TABLE II. LEARNING CONFIGURATION.

Item	Value / Implementation
Graph hidden dim / layers / heads	32 / 2 / 2
Path length upper bound	4 downstream segments
Actor & critic hidden layers	64, 64; ReLU
Learning rate	actor, critic, representation module all 0.0003
Batch & buffer capacity	256 / 60000 transitions
Per-second discount γ	0.995
Target soft-update τ	0.005; executed every two critic updates
Exploration / target noise std	0.25 / 0.20, both in normalized action space
Target noise clip range	0.40
Update-start condition	accumulate $>$2000 transitions per episode
Update trigger frequency	once per 3 simulation steps after condition met
Gradient clipping	not used

The nominal 45-episode experiment continues training from the 25-episode checkpoint, restoring only the model weights and the return baseline, not the optimizer state, replay buffer, or random-number state. Therefore, this result describes the variation under a specific continuation setting and cannot be equated with continuous 45-episode training that fully retains the training state.

E. Statistical Analysis

The main comparison reports the mean and sample standard deviation of 24 runs for each method; the standard deviation includes variations in both demand level and demand seed. The training-repetition analysis first computes the mean of each training model over 24 evaluations, then counts the mean and standard deviation among the three model means. The two types of standard deviation describe evaluation-scenario variation and training-model variation, respectively, and are not interpreted by mixing.

The paired analysis first averages the three demand levels within each demand seed, then computes the method difference from the eight seed-level results. A two-sided Wilcoxon signed-rank test is used, enumerating the sign combinations of non-zero differences to compute the exact p-value. With $n = 8$ and no zero differences, the minimum reachable two-sided p-value is 0.0078125. For each comparison, Holm correction is applied to the four main metrics: relative to the no-advice baseline, $\bar{T}_{\mathrm{obs}}$, NS, CO_2, and AV are tested; for other comparisons, $\bar{T}_{\mathrm{obs}}$, NS, CO_2, and DF are tested. The confidence intervals of the seed-level differences are obtained from 20000 resampling iterations.

This statistical scheme is used for exploratory analysis of existing data. The multiple correction within each comparison does not equal controlling the overall error rate constituted by all methods and all metrics, and the demand-seed-level pairing cannot eliminate the uncertainty brought by training-model selection [7].

IV. MAIN RESULTS AND TRAINING-REPETITION ANALYSIS

A. Overall Performance and Traffic-Demand Differences

Tables III and IV give the nine metrics of the main comparison. Relative to the no-advice baseline, the full model

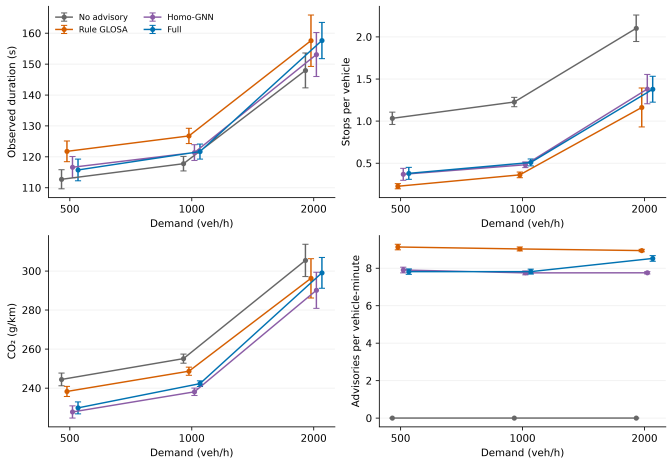

Fig. 4. Operation metrics of four representative methods under different demand levels. Points are the mean over eight demand seeds, and error bars are sample standard deviations. The vertical axes are, in order, mean recorded duration, stops per vehicle, CO_2 intensity, and advice frequency. Lines connect only the discrete demand scenarios.

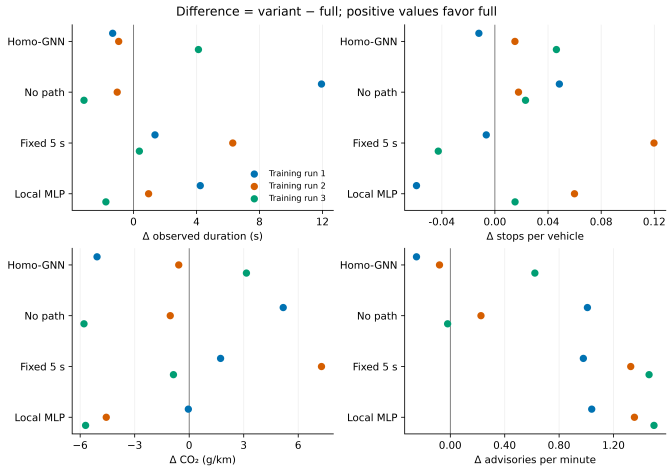

Fig. 5. Component comparison of three training realizations. The difference is defined as variant minus full model; each point corresponds to the mean difference of one training model over 24 evaluations. Colors distinguish training realizations. A positive value indicates a lower value for the full model, where the recorded-duration difference does not directly represent a full-trip efficiency difference.

reduces the number of stops and CO_2 emission intensity by 48.0% and 4.2%, respectively, while increasing the mean recorded duration by 4.4% and the mean absolute jerk by 23.4%. This result indicates that the improvement in stopping and emissions is accompanied by changes in other operation metrics, and the full model is not superior on all metrics.

The no-advice baseline does not uniformly impose the 15 m/s vehicle maximum-speed upper bound as the advice methods do, while the SUMO speed factor allows different vehicles to have different desired speeds [10]. Therefore, the above differences may include both the influence of the control strategy and the speed-limit implementation manner, and the existing comparison cannot attribute all improvements to the learning strategy.

Relative to rule-based GLOSA, the full model's $\bar{T}_{\mathrm{obs}}$, CO_2 intensity, and DF decrease by about 2.7%, 1.5%, and 10.9%, respectively, while NS increases by about 29.5% and SF is also higher. Under this comparison, the full model shows lower emission intensity and advice frequency but more stops. Considering that the rule baseline uses an approximate signal window, this conclusion applies only to this specific implementation and cannot be generalized to being superior to existing GLOSA methods overall.

Fig. 4 shows the differences among the four representative methods under low, medium, and high demand. As demand increases, the recorded duration, stops, and CO_2 intensity of each method generally rise, indicating that traffic load has a clear impact on operation performance. The cross-demand merged mean reflects the average performance of the three selected scenarios, and its standard deviation also includes variations among scenarios and cannot replace the separate judgment of each demand level.

B. Consistency across Different Training Models

To examine whether the main comparison depends on a single training, Table V summarizes the results of three training models. The full model's mean $\bar{T}_{\mathrm{obs}}$ and NS are lower than the listed learning comparisons, but its CO_2 intensity is not better than the local MLP, homogeneous GNN, or remove-path variant. Therefore, the cross-model averaged result still reflects the trade-off among different metrics.

Tables VI and Fig. 5 give the differences of the three training realizations as "variant minus full model." For metrics where a smaller value is usually more favorable, a positive value indicates that the full model is lower. The three models are insufficient to determine the performance distribution of the training population; the differences are mainly used to check whether the comparison direction is consistent.

The fixed 5 s variant has a higher DF in all three trainings, averaging 9.04 /min, while the full model averages 7.78 /min, a reduction of about 13.9%. The $\bar{T}_{\mathrm{obs}}$ difference also remains positive, but the CO_2 difference becomes negative in the third group, and NS does not remain consistent across the three groups. Therefore, the more consistent evidence is the reduction in advice frequency relative to this fixed 5 s configuration; it cannot be simultaneously inferred that all traffic and emission metrics improve stably.

The remove-path-readout $\bar{T}_{\mathrm{obs}}$ differences are $+11.93$, -1.04, and -3.14 s in sequence; the path benefit observed in the first group is not reproduced in the other two groups. The CO_2 difference of the homogeneous graph relative to the full model also changes sign. The above results are insufficient to establish the independent performance advantage of relation-specific encoding or path readout. The local MLP has lower CO_2 intensity than the full model in all three groups, but because its inputs and action configuration differ, this result also cannot be used alone to determine the role of the graph representation.

C. Paired Comparison of a Single Training Model

Table VII summarizes the paired test of the main model relative to six comparisons; a positive relative difference indi-

TABLE III. MAIN METRICS: MEAN ± SAMPLE STANDARD DEVIATION OVER 24 RUNS.

Method	$\bar{T}_{\text{obs}}$ (s)	NS (/veh)	CO_2 (g/km)	AV (m/s^3)	DF (/min)
No advice (NA)	126.13±16.33	1.454±0.486	268.30±27.65	0.803±0.017	0.00±0.00
Rule GLOSA	135.36±16.97	0.583±0.441	261.04±26.44	1.041±0.016	9.03±0.13
Local MLP	135.92±17.34	0.696±0.528	257.00±27.40	0.977±0.023	9.09±0.16
Fixed 5 s	133.04±20.80	0.749±0.466	258.77±33.98	0.984±0.014	9.03±0.14
Homo. GNN	130.35±17.16	0.743±0.474	251.98±28.42	0.980±0.016	7.80±0.13
Remove-path	143.61±27.73	0.804±0.630	262.22±39.79	0.968±0.019	9.06±0.32
Full model	131.68±19.32	0.755±0.463	257.05±31.14	0.991±0.015	8.05±0.37

TABLE IV. OTHER METRICS: MEAN ± SAMPLE STANDARD DEVIATION OVER 24 RUNS.

Method	AS (m/s)	FC (mL/km)	EC (MJ/km)	SF (m/s)
No advice (NA)	9.33±0.98	115.97±11.95	3.97±0.41	5.20±0.25
Rule GLOSA	8.62±0.87	112.84±11.43	3.86±0.39	3.96±0.21
Local MLP	8.56±0.85	111.09±11.84	3.80±0.41	4.17±0.23
Fixed 5 s	8.86±1.06	111.85±14.69	3.83±0.50	4.14±0.29
Homo. GNN	8.97±0.92	108.92±12.28	3.72±0.42	4.08±0.29
Remove-path	8.32±1.20	113.34±17.20	3.88±0.59	4.05±0.22
Full model	8.92±1.03	111.11±13.46	3.80±0.46	4.16±0.28

TABLE V. MEAN ± STANDARD DEVIATION AMONG THE THREE TRAINING-MODEL MEANS.

Method	$\bar{T}_{\text{obs}}$ (s)	NS (/veh)	CO_2 (g/km)	DF (/min)
Full model	133.35±2.29	0.722±0.031	258.03±1.49	7.78±0.24
Homo. GNN	133.97±5.32	0.738±0.027	257.20±5.47	7.88±0.30
Remove-path	135.94±6.68	0.752±0.047	257.48±4.27	8.19±0.77
Fixed 5 s	136.03±2.85	0.746±0.070	260.74±3.33	9.04±0.01
Local MLP	134.50±1.30	0.727±0.029	254.59±2.18	9.08±0.02

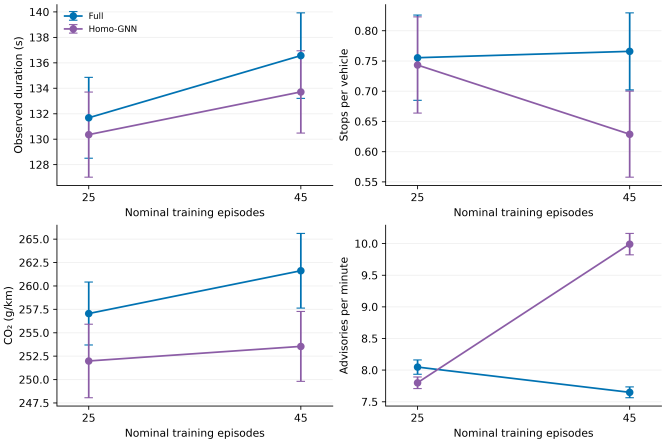

Fig. 6. Results of the full model and the homogeneous GNN under two nominal training budgets. Error bars are sample standard deviations over eight seed-level means, with the three demand levels averaged within each seed. Lines compare only the two checkpoints; the 45-episode result is from continued training that did not restore the full training state.

cates a higher value for the full model. Relative to the fixed 5 s variant, the Holm-corrected p-value of DF is 0.0312; $\bar{T}_{\text{obs}}$, NS, and CO_2 do not reach the 0.05 threshold within this four-metric family. This result supports the current main model's frequency difference but does not constitute evidence of traffic-performance equivalence or non-inferiority.

The four differences relative to the no-advice baseline all pass the correction threshold within their metric family, including the increases in $\bar{T}_{\text{obs}}$ and AV, indicating that the statistical difference involves both benefits and costs. The significant differences in the homogeneous-graph and path comparisons correspond only to the main training model; combined with the direction change in Table VI, they should not be interpreted as module contributions that hold across training realizations.

V. CONTINUATION-TRAINING RESULT ANALYSIS

To examine the relationship between the existing training-budget variation and the operation metrics, the results of the full model and the homogeneous GNN at the 25-episode and nominal 45-episode checkpoints are compared. Each variant has only one continuation trajectory; the checkpoint variation is shown in Fig. 6. The standard deviation in Table VIII is computed from eight seed-level means, with the three demand levels already averaged within each seed, so it is not directly comparable to the 24-run mixed standard deviation in Table III.

After the full model's continuation training, $\bar{T}_{\text{obs}}$ and CO_2 intensity rise and DF falls; the homogeneous-graph model's NS falls while its DF rises. At the nominal 45-episode checkpoint, the full model's DF is about 23.45% lower than the homogeneous graph, but its $\bar{T}_{\text{obs}}$, NS, and CO_2 intensity are all higher. These changes do not show that all metrics improve synchronously with the increase in training budget.

Both variants use a variable interval, so their frequency difference cannot prove that a variable interval is superior to a fixed interval. Because there is no same-budget fixed-interval comparison and the continuation did not restore the full training state, the existing results only describe the operation trade-off between the selected checkpoints and cannot serve as evidence of algorithm convergence or the advantage of heterogeneous encoding.

VI. DISCUSSION AND LIMITATIONS

A. Role and Evidence of Update-Interval Learning

The relatively consistent result of this paper is that the full model has a lower advice frequency than the independently

TABLE VI. DIFFERENCES OF THREE TRAINING REALIZATIONS: VARIANT − FULL MODEL (THREE GROUPS).

Variant − Full	$\Delta\bar{T}_{\mathrm{obs}}$ (s)	ΔCO_2 (g/km)	ΔDF (/min)
Homo. GNN	−1.32 / −0.94 / +4.12	−5.07 / −0.57 / +3.15	−0.25 / −0.08 / +0.62
Remove-path	+11.93 / −1.04 / −3.14	+5.17 / −1.04 / −5.79	+1.01 / +0.23 / −0.02
Fixed 5 s	+1.36 / +6.30 / +0.37	+1.73 / +7.28 / −0.87	+0.98 / +1.33 / +1.46
Local MLP	+4.24 / +0.95 / −1.76	−0.05 / −4.56 / −5.70	+1.04 / +1.35 / +1.50

TABLE VII. PAIRED SIGNED-RANK TEST UNDER A SINGLE TRAINING MODEL ($n = 8$).

Comparison	**Metric**	**Rel. Diff (%)**	**Raw p**	**Holm p**
No advice (NA)	$\bar{T}_{\mathrm{obs}}$	+4.40	0.0078	0.0312
No advice (NA)	NS	−48.04	0.0078	0.0312
No advice (NA)	CO_2	−4.20	0.0078	0.0312
No advice (NA)	AV	+23.42	0.0078	0.0312
Rule GLOSA	$\bar{T}_{\mathrm{obs}}$	−2.72	0.0078	0.0312
Rule GLOSA	NS	+29.50	0.0078	0.0312
Rule GLOSA	CO_2	−1.53	0.0078	0.0312
Rule GLOSA	DF	−10.88	0.0078	0.0312
Local MLP	$\bar{T}_{\mathrm{obs}}$	−3.12	0.0078	0.0312
Local MLP	NS	+8.48	0.0391	0.0781
Local MLP	CO_2	+0.02	0.7422	0.7422
Local MLP	DF	−11.45	0.0078	0.0312
Fixed 5 s	$\bar{T}_{\mathrm{obs}}$	−1.02	0.3125	0.9375
Fixed 5 s	NS	+0.87	0.7422	0.9375
Fixed 5 s	CO_2	−0.67	0.3125	0.9375
Fixed 5 s	DF	−10.84	0.0078	0.0312
Homo. GNN	$\bar{T}_{\mathrm{obs}}$	+1.02	0.1953	0.3906
Homo. GNN	NS	+1.63	0.5469	0.5469
Homo. GNN	CO_2	+2.01	0.0156	0.0469
Homo. GNN	DF	+3.18	0.0078	0.0312
Remove-path	$\bar{T}_{\mathrm{obs}}$	−8.31	0.0078	0.0312
Remove-path	NS	−6.04	0.2500	0.2500
Remove-path	CO_2	−1.97	0.0234	0.0469
Remove-path	DF	−11.13	0.0078	0.0312

TABLE VIII. CONTINUATION CHECK: MEAN ± STANDARD DEVIATION OVER EIGHT SEED-LEVEL MEANS.

Method	$\bar{T}_{\mathrm{obs}}$ (s)	NS (/veh)	CO_2 (g/km)	DF (/min)
Full model	131.68±3.18	0.755±0.071	257.05±3.36	8.05±0.11
Full model 45 ep	136.56±3.36	0.766±0.064	261.62±3.99	7.65±0.09
Homo. GNN	130.35±3.34	0.743±0.080	251.98±3.93	7.80±0.09
Homo. GNN 45 ep	133.71±3.23	0.629±0.071	253.54±3.74	9.99±0.17

trained fixed 5 s variant. This direction is consistent with the design of jointly outputting the interval and charging a decision cost per occurrence, but the existing experiments cannot yet distinguish the respective contributions of state responsiveness, interval upper bound, and the cost term.

DF uses the entire vehicle recorded duration as the denominator, while decisions occur only within the control zone. Therefore, DF is simultaneously affected by the in-zone update interval, the time ratio of the vehicle being in the control zone, and the record length, and is not directly equal to the reciprocal of the selected interval. The fixed 5 s method's DF being lower than 12 /min is compatible with this statistical range; the in-control-zone update frequency and the actual interval distribution still need verification by per-event records.

Existing research has already combined advice-update selection with continuous driving actions [11], [12]. This paper adopts a path graph representation and appointment-interval output, which is a different implementation under this research direction. There is currently a lack of independent training comparisons with a fixed 6 s, removed decision cost, and removed return centering, so it cannot yet be proved that state-dependent interval selection is superior to a longer fixed interval. The DF reduction also only indicates a decrease in the advice-event frequency and does not directly correspond to reductions in communication data volume, end-to-end latency, or computing energy consumption; in particular, the single-inference cost of the graph encoder needs to be measured separately.

B. Role of Relation Modeling and Path Information

Relation-specific encoding provides a structured aggregation manner for segment and signal states, and path readout supplements the global traffic background with the ego-vehicle downstream-related representation. Both designs have traffic-semantic justification, but the design justification and the experimental advantage belong to different levels of conclusion. The sign change of differences in the three trainings indicates that the existing data are insufficient to confirm their stable benefit.

The small-scale road network, fixed signals, independent projection of the two node types, and limited training budget may weaken or change the differences among different encoders, but these explanations have not yet been separated by independent experiments. The segment adjacency is not filtered by feasible turns and may also introduce information inconsistent with the ego vehicle's feasible travel path. Therefore, this paper positions the role of relation modeling and path readout as state-representation design and discusses their performance differences based on the existing repetition results, without generalizing the improvement in a single training to a stable module benefit.

C. Implementation Consistency and Result Interpretation

The results of this paper correspond to the historical implementation and its checkpoints. The historical path module may read in the segment embedding corresponding to the padded position when no valid downstream segment exists; the maintenance version sets the empty-path readout to zero and corrects the static-graph edge cache, but has not yet completed the retraining and re-evaluation corresponding to the results of this paper with the maintenance version. Therefore, the repair record does not change the historical performance in the tables, nor can it be used to prove that the repaired policy has the same performance.

Return centering, integer-second execution of the continuous interval, and the dual gradient source of the shared representation are all training designs that require special verification. The duration discount in the semi-Markov target is compatible with unequal-interval control, but does not automatically solve the problems of incomplete observation, action quantization, or training non-stationarity. For transitions that reach the simulation time limit, the distinction between true task termination and sampling truncation should also be made, and the termination flag and subsequent value handling should be consistent with the task definition [13].

Inconsistent baseline speed limits, approximate rule signal windows, and right-censoring of vehicle records affect the interpretation of method comparison. Speed-limit consistency and arrival state cannot be traced and corrected from the existing summary tables and require re-running with unified settings. The stopping events in the original reward differ in definition from the evaluation metrics, and the squared-acceleration penalty does not directly constrain jerk; these differences help explain why the optimization objective and some evaluation results are not in the same direction, but the specific impact still needs ablation verification.

D. Scope of Application

All experiments are conducted in the same 3×3 fixed-signal network, with vehicles accepting advice as set. This paper does not verify different topologies, mixed penetration rates, communication delay, actuated signals, real driving response, or different emission models. The default vehicle type and speed factor also limit the extrapolation to other vehicles and driving populations. Therefore, the current conclusions apply to the operation and update-frequency trade-off under the reported simulation conditions and do not constitute direct evidence of real-road deployment effects or safety.

VII. Conclusion

This paper constructs a GLOSA framework combining relation-specific graph encoding, remaining-path readout, and joint decision of speed and update interval, and analyzes the trade-off between vehicle operation performance and advice frequency based on 456 historical unique evaluation records. Relative to the no-advice baseline, the single-training model reduces the number of stops and CO_2 emission intensity by 48.0% and 4.2%, respectively, while increasing the mean recorded duration and mean absolute jerk by 4.4% and 23.4%. Because the baseline speed-limit settings are inconsistent and the vehicle records are censored, these results cannot be directly attributed to the strategy's improvement in full-trip efficiency.

Relative to the independently trained fixed 5 s variant, the advice-frequency reduction remains consistent across the three trainings, with the cross-training mean decreasing from 9.04 to 7.78 /min. The performance differences of relation-specific encoding and path readout change direction with the training realization, and the continuation comparison does not establish the advantage of the heterogeneous encoder. Therefore, the existing evidence supports treating the advice frequency as an explicit optimization factor for GLOSA, but is insufficient to prove the general superiority of the adopted graph structure or state-dependent update strategy.

The main value of this paper is to give the complete connection manner of path graph representation and unequal-interval speed advice, and to reveal the trade-off between operation performance and update frequency through different demands and training realizations. The above conclusions are limited to the implemented baselines, training configurations, and fixed-signal road network, providing a simulation basis for understanding the combined influence of representation design and advice update in GLOSA.

References

[1] Z. Yang, Z. Zheng, J. Kim, and H. Rakha, "Eco-driving strategies using reinforcement learning for mixed traffic in the vicinity of signalized intersections," *Transportation Research Part C: Emerging Technologies*, vol. 165, p. 104683, 2024.

[2] Y. Qin, Y. Huang, S. Yan, and Y. Han, "Adaptive formation eco-driving framework for connected automated vehicles at signalized intersections: A deep reinforcement learning approach," *Expert Systems with Applications*, vol. 279, p. 127509, 2025.

[3] Q. Chen and M. Wang, "Deep reinforcement learning for eco-driving control of connected and automated vehicles at consecutive signalized intersections," *Transportation Research Record*, p. 03611981261431730, 2026.

[4] Y. Xue, D. Bai, Y. Zhou, C. Ding, H. L. Vu, and B. Yu, "Chance-constrained eco-driving control of connected autonomous vehicles in mixed traffic environment at signalized intersections with uncertain signal timings," *Transportation Research Part C: Emerging Technologies*, vol. 183, p. 105460, 2026.

[5] D. Vignarca, S. Arrigoni, U. Montanaro, E. Sabbioni, and D. Tavernini, "Design and implementation of an nmpc-based green light optimal speed advisory for buses: A simulation platform," *Control Engineering Practice*, vol. 172, p. 106820, 2026.

[6] A. Chaudhry, R. Haouari, E. Papazikou, M. K. Singh, H. Sha, A. Tympakianaki, L. Nogues, M. Quddus, W. Weijermars, P. Thomas *et al.*, "Examining road safety impacts of green light optimal speed advisory (glosa) system," *Accident Analysis & Prevention*, vol. 200, p. 107534, 2024.

[7] T. Wang, Z. Zhu, J. Zhang, J. Tian, and W. Zhang, "A large-scale traffic signal control algorithm based on multi-layer graph deep reinforcement learning," *Transportation Research Part C: Emerging Technologies*, vol. 162, p. 104582, 2024.

[8] W.-Y. Lin, Y.-Z. Song, B.-K. Ruan, H.-H. Shuai, C.-Y. Shen, L.-C. Wang, and Y.-H. Li, "Temporal difference-aware graph convolutional reinforcement learning for multi-intersection traffic signal control," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 1, pp. 327–337, 2023.

[9] F.-X. Devailly, D. Larocque, and L. Charlin, "Model-based graph reinforcement learning for inductive traffic signal control," *IEEE Open Journal of Intelligent Transportation Systems*, vol. 5, pp. 238–250, 2024.

[10] 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, pp. 11 330–11 344, 2024.

[11] H. Zhu, F. Sun, K. Tang, T. Han, and J. Xiang, "A coordination graph based framework for network traffic signal control," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 10, pp. 14 298–14 312, 2024.

[12] Z. Zhai, R. Hao, B. Cui, and S. Wang, "Hgat and multi-agent rl-based method for multi-intersection traffic signal control," *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 5, pp. 6848–6864, 2025.

[13] J. Yoon, K. Ahn, K. Lee, J. Park, and H. Yeo, "Decentralized and communication-based multi-agent traffic signal control model employing a graph representation for the state," *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 5, pp. 5947–5960, 2025.

[14] K. Wang, Z. Shen, Z. Lei, X. Liu, and T. Zhang, "Towards multi-agent reinforcement learning based traffic signal control through spatio-temporal hypergraphs," *IEEE Transactions on Mobile Computing*, 2025.

[15] B.-L. Ye, P. Wu, L. Li, W. Wu, B. Song, and X. Zhang, "Multi-intersection traffic signal control based on dynamic spatiotemporal memory enhanced learning," *Control Engineering Practice*, vol. 165, p. 106606, 2025.

[16] S. Yang and Z. Zeng, "Heterogeneous graph distributional reinforcement learning for out-of-distribution traffic signal control," *Expert Systems with Applications*, p. 132707, 2026.