IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 March 2026, accepted 28 April 2026, date of publication 1 May 2026, date of current version 7 May 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3689744

RESEARCH ARTICLE

Multi-Agent Reinforcement Learning-Based Variable Speed Limit Control for Mixed Urban Traffic Flow

XUAN FANG[1], ISTVÁN VARGA[1,2], AND TAMÁS TETTAMANTI[1,2]

[1]Department of Control for Transportation and Vehicle Systems, Faculty of Transportation Engineering and Vehicle Engineering, Budapest University of Technology and Economics, 1111 Budapest, Hungary
[2]Systems and Control Laboratory, HUN-REN Institute for Computer Science and Control, 1111 Budapest, Hungary

Corresponding author: Tamás Tettamanti (tettamanti.tamas@kjk.bme.hu)

This work was supported by the European Union within the Framework of the National Laboratory for Autonomous Systems under Grant RRF-2.3.1-21-2022-00002.

ABSTRACT With the acceleration of urbanization and the advent of autonomous driving technology, urban traffic systems are gradually evolving into a mixed traffic flow environment where Human Driven Vehicles (HAVs) and Connected and Automated Vehicles (CAVs) coexist. Conventional urban traffic control is primarily based on traffic lights with fixed or adaptive logic, as other means of actuation are not traditionally in use (except for tolling systems, which are rather indirect regulators). Traditional traffic signal control, however, is limited due to its rigidity; traffic lights are located at fixed spots with a strict operational mechanism (cycle time, phase order, offset), making them less efficient in managing a mixed traffic environment. To address this challenge, this paper follows a novel approach leveraging the Variable Speed Limit (VSL) control in the urban context. The control framework is based on Multi-Agent Reinforcement Learning (MARL), aiming to optimize road sustainability under mixed traffic flow conditions. This study employs a Centralized-Training Decentralized-Execution (CTDE) architecture, utilizing the Multi-Agent Proximal Policy Optimization (MAPPO) algorithm to address the curse of dimensionality and non-stationarity issues in large-scale urban road networks. This framework enables collaborative decision-making in both space and time through the use of shared parameters and global value function evaluation. The method is validated via realistic traffic simulation experiments in Simulation of Urban MObility (SUMO), demonstrating that the proposed control strategy can significantly reduce travel time, pollutant emissions, and improve traffic safety on the studied urban network under different traffic conditions.

INDEX TERMS Connected and automated vehicles, multi-agent reinforcement learning, mixed traffic flow, pollutant emissions, urban traffic control, variable speed limit control.

I. INTRODUCTION

Traffic congestion is a growing problem in cities around the world. It results in increased travel time, environmental pollution, and economic externalities [1], [2], [3], [4]. Essentially, traffic congestion is caused by an imbalance between traffic supply and growing demand. Accordingly, efficient traffic management becomes increasingly important in the context of maximizing the use of limited road facilities [5]. At the same time, urban traffic control in most cities still relies predominantly on fixed-time or traffic-responsive signal plans, which can only change green phases and offsets. Despite decades of development, these systems have a limited adaptivity due to the rigid nature of traffic lights (fixed spots with simple control), i.e., these systems can not directly influence queue formation, headways, or the propagation of stop-and-go shockwaves. While more advanced Reinforcement Learning (RL) based traffic signal controllers have been proposed, their large-scale deployment remains challenging due to safety certification and fail-safe

The associate editor coordinating the review of this manuscript and approving it for publication was Jiachen Yang.

requirements, real-time computational reliability, robustness to non-recurrent events, and integration with legacy signal hardware and operational constraints (e.g., tram priority and coordination plans). To address these limitations, this study proposed a multi-agent deep RL framework for Variable Speed Limit (VSL) control in a mixed urban traffic environment. The proposed control framework enables dynamic regulation of vehicle speeds, thereby improving network performance in ways that conventional signal control cannot.

Related research has applied multi-agent reinforcement learning to traffic control problems such as coordinated signal control and ramp metering, but MARL-based VSL control remains comparatively less explored in complex urban networks. Existing RL-VSL studies often rely on single-agent decision-making across an entire corridor, simplified networks, or homogeneous vehicle fleets.

The effective control of intersections (being the main bottlenecks) is the key to urban traffic management. Some effective control measures have been applied to urban traffic networks to alleviate traffic congestion and air pollution, such as traffic signal control [6], VSL control [7], and dynamic routing [8]. With the massive accumulation of data from various sources (traffic sensors, navigation), a large number of advanced data-driven machine-learning algorithms have been proposed for the traffic control methods mentioned above [9]. For example, the traffic light control with the classical Q-learning RL [10], [11], Policy Gradient algorithm [12], [13], or deep RL algorithm [14], [15]. Still, the practical limitation of AI-based traffic lights is that upgrading existing conventional traffic light controllers capable of RL is not straightforward (e.g., technical standards, processing times), and it definitely requires a large financial budget. VSL control is another (not yet widespread) feasible solution for traffic control in cities. VSL has been widely used in highway traffic control systems capable of effectively alleviating traffic congestion and improving traffic efficiency [16]. Few studies have also conducted VSL control on signalized intersections on urban roads to improve traffic performance. Classical VSL controllers in urban settings typically rely on rule-based or optimization schemes, e.g., fuzzy-logic [17], [18], optimal control [19], [20], and nonlinear model-predictive control (MPC) [21], [22]. Reference [23] proposed a nonlinear MPC with a macroscopic traffic model to dynamically set speed limits in a synthetic urban network. The referred VSL methods, however, generally treat each control location in isolation and assume accurate traffic models without limiting their applicability in complex, mixed-traffic urban networks.

By contrast, data-driven RL-based methods have been increasingly explored for VSL. RL automatically perceives the state of the environment based on real-time data from environmental feedback and interacts with the environment through a continuous trial-and-error mechanism to learn the optimal control decision. Early work applied tabular Q-learning agents to control speed limits or vehicle speeds near intersections. Reference [24] proposed a speed agent based on Q-learning to minimize fuel consumption by controlling the speed of platoons and individual CAVs crossing signalized intersections. The results show that the proposed speed agent solution improves the average fuel efficiency by 13.6%.

Although the Q-learning-based VSL control can solve the shortcomings of the traditional VSL control strategy, such as poor flexibility, slow control action, and insignificant effect, it uses a two-dimensional table (Q table) to store the expected cumulative reward (Q value) of each ''state-action pair''. When the environment is more complex, the rapid increase in the dimensions of the action and state space will lead to the curse of dimensionality problem. Therefore, Q-learning-based VSL control is only beneficial under simple road environment conditions (e.g., a single intersection). Moreover, many RL-VSL studies assume a single decision-maker for an entire segment (single-agent RL), which cannot capture interactions among multiple VSL signs or adjacent intersections. Previous research has also demonstrated the limitations of a single agent [25].

In conclusion, the existing VSL strategies lack coordination among multiple controllers or require accurate models. These gaps motivate a multi-agent RL approach that can jointly optimize speed limits in a heterogeneous urban network. Deep Reinforcement Learning (DRL) combines the perception capabilities of deep learning and the decision-making capabilities of RL. Based on the advantages of DRL in solving VSL control problems and the controllability of CAV, this paper proposes a Multi-Agent Proximal Policy Optimization (MAPPO) VSL control strategy for large-scale urban networks in mixed traffic environments.

The main contributions of our research are outlined below:

- We develop and evaluate a multi-location MARL-based VSL control framework on a large-scale urban network under free-flow, capacity-flow, and extreme, perturbed flow conditions.
- We construct a SUMO-gym training and evaluation environment, explicit modeling of mixed CAVs/HDVs traffic based on real-world traffic data.
- We formulate a CTDE-based MAPPO algorithm with parameter sharing to improve scalability, and design a multi-objective reward that jointly considers emissions, efficiency, and safety-related surrogate measurements.
- We analyze an Independent Proximal Policy Optimization (IPPO)-based VSL baseline as an ablation of MAPPO that removes centralized critic coordination.

II. METHODOLOGY

A. MICROSCOPIC TRAFFIC SIMULATION

A reliable traffic simulation platform is essential for testing and validating reinforcement learning-based traffic control

systems. RL relies on trial-and-error to optimize decisions and requires dynamic and accurate representations of real-world scenarios. To this end, the microscopic traffic simulation platform SUMO (Simulation of Urban MObility) [26] together with its embedded Traffic Control Interface (TraCI), as well as the Python programming language, were exploited to simulate the mixed traffic dynamics in large-scale urban scenarios with VSL control.

B. CONNECTED AND AUTOMATED VEHICLES

In the actual application of variable speed limit control, due to the driver's compliance with the speed limit, there is a certain difference between the actual vehicle speed and the published speed limit. Driver compliance with the speed limit is a non-negligible issue in the study of variable speed limit control. Researchers have analyzed the impact of driver compliance on the effect of variable speed limit control. The results show that when the driver does not comply with the speed limit, the effect of variable speed limit control is significantly reduced [27], [28]. Additionally, the layout of traffic flow detection equipment and variable speed limit signs within the variable speed limit control system is a crucial factor that impacts the control effect. The introduction of CAV changed this situation. Grumert and Tapani [29] indicated that the variable speed limit system based on the CAV environment has smoother speed changes and reduces exhaust emissions by 1.5-2.5% compared to the existing VSL system included in Stockholm. Yu and Fan [30] proposed an optimal VSL strategy for a CAV environment on a freeway corridor with multiple bottlenecks. The simulation results showed that as the penetration rate of CAV increases, better control performance can be achieved.

Relying on intelligent decision-making control technology and Vehicle-to-Everything (V2X) communication technology, CAVs possess certain autonomous driving and connected communication capabilities. They can travel not only with a smaller expected headway but also have a high degree of obedience and timely response to control commands [31]. Through Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) communication, various participants in the transportation system can collaborate to enhance road safety, regulate traffic flow, and reduce energy consumption.

Considering that the full coverage of CAVs will take a considerable amount of time, a mixed traffic flow environment composed of Human Driven Vehicles (HDVs) and CAVs will likely exist in the near future. In the experiments of this paper, unless otherwise stated, the CAV penetration rate is set to 60% CAV and 40% HDV, consistent with mixed-traffic findings reported in [32]. CAVs are assumed to comply with commanded VSL settings (100% compliance via V2I), whereas HDV compliance is modeled probabilistically with a default rate. 70% of HDVs use a compliant speed configuration, and the other 30% of HDVs will be speeding by 10%. To simulate the mixed traffic flow environment, it is necessary to construct a mixed traffic flow model, which consists of two parts: a car-following model and a lane-changing model. For HDVs, the longitudinal following and lateral lane-changing behaviors of HDVs are modeled using the Intelligent Driving Model (IDM) [33] and the LC2013 model. The paper adopts an existing realistic CACC car-following model calibrated and validated based on field test data to model the longitudinal movement of CAVs [34]. The parameters of the LC2013 model for CAVs were selected according to the TransAID project to realistically model lane-changing behavior [35].

C. MULTI AGENT REINFORCEMENT LEARNING STRATEGY

Reinforcement learning is a learning mechanism that makes decisions through the dynamic interaction between the agent and the environment. The basic principle of reinforcement learning is that the agent performs a certain action in the environment, changes the state of the environment, and receives a reward signal from the environment to strengthen the mapping relationship between a specific state and the optimal action strategy. By repeatedly performing this process, the agent can develop the ability to provide the optimal action strategy in any environmental state.

The reinforcement learning methodology is derived formally from the Markov decision process (MDP) [36]. MDP is a mathematical framework for solving sequential decision-making problems. To address multi-location coordination, we formulate VSL control as a cooperative multi-agent Markov game and apply the Multi-Agent Proximal Policy Optimization (MAPPO) algorithm. Consider N VSL controllers (agents) deployed in an urban network (e.g., approaching intersections). Each agent i generates an action a_t^i based on its local observations o_t^i at each decision step t and a shared policy to maximize the cumulative reward with discounts. It also learns a central value function $V_\psi(s_t)$ based on the global state s_t. The joint observations is $o_t = \{o_t^1, \ldots, o_t^N\}$ and the joint action $a_t = \{a_t^1, \ldots, a_t^N\}$. The environment (traffic dynamics) then transitions to a new state s_{t+1} according to a stochastic model $\mathcal{T}(s_{t+1} \mid s_t, a_t)$, and all agents receive a global reward r_t. Formally, this is a Markov decision process $\langle \mathcal{S}, \mathcal{A}_i, \mathcal{T}, \mathcal{R}, \gamma, \mathcal{N} \rangle$, where $\mathcal{S}$ represents the state space; $\mathcal{A}_i$ represents the action space; $\mathcal{R}$ represents the reward function, it calculates the reward obtained after taking action a_t and transitioning from state s_t to the next state s_{t+1} based on an artificially defined function. $\gamma \in [0, 1]$ is the discount factor used to calculate the expected cumulative reward.

Proximal Policy Optimization (PPO) [37] is an improved method based on the policy gradient algorithm (PG) [38] and trust region policy optimization (TRPO) [39]. The PG algorithm is susceptible to problems such as high variance and gradient vanishing, while the TRPO algorithm is complex to implement and difficult to apply to practical problems. The PPO algorithm introduces a truncation function to limit the amplitude of the policy update. This truncation function

is called Proximal Clip Objective (PPO-Clip), which limits the amplitude of policy updates to a smaller range, thereby avoiding instability caused by excessive policy updates. The PPO algorithm improves training efficiency and stability by using multiple parallel actors to train the strategy. The PPO algorithm offers significant improvements over the PG and TRPO algorithms in terms of ensuring stability and convergence speed, making it widely used in practice.

The Centralized-Training Decentralized-Execution (CTDE) paradigm is leveraged, extending the PPO algorithm to MAPPO. During training, a centralized critic network has access to the global state to estimate a joint value function, while each agent's actor network sees only that agent's local observation. This setup allows the critic to guide learning with global information, mitigating non-stationarity, while the actors remain decentralizable for deployment. In our VSL control case, the centralized critic might observe all traffic sensor data, whereas each agent's actor only takes its own sensor readings as input. CTDE has been shown to stabilize multi-agent learning and improve scalability [40].

The actor-critic (AC) architecture was utilized for MAPPO. The AC algorithm [41] combines the low-variance, single-step bootstrapping of Q-learning with the direct optimization of policy-gradient (PG) methods. Let $\pi_\theta(a_t|s_t)$ be the stochastic policy (actor) for agents parameterized by neural network θ, mapping its observation s_t to an action probability, and $Q_t(a_t, s_t)$ the action-value function (critic) for π_θ, yielding a value estimate $V_\psi(s_t)$. Typically, both actor and critic are multilayer perceptrons; in our control framework, each had three hidden layers of size 64. The MAPPO algorithm uses the importance sampling method. MAPPO adopts the new and old strategy $\pi_\theta(a_t \mid s_t)$ and $\pi_{\theta old}(a_t \mid s_t)$ to allow the network to interact with the traffic environment and use the collected data to train the policy π_θ so that the agent can perform multiple parameter updates in one interaction with the environment, which improves the update speed.

$$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta\,\mathrm{old}}(a_t \mid s_t)} \tag{1}$$

where $r_t(\theta)$ is the probability ratio. To improve sample efficiency and generalization, agents share network weights. In reinforcement learning, the advantage function A_t measures how much better an action a_t is compared to the average action at state s_t:

$$A_t = Q_t(a_t, s_t) - V_\psi(s_t) \tag{2}$$

However, computing $Q_t(a_t, s_t)$ exactly is often impractical. The Generalized Advantage Estimation (GAE) method, proposed by [42], provides a way to estimate advantages with a balance between bias and variance control, using a parameter $\lambda \in [0, 1]$. The temporal-difference residual at time step is defined as:

$$\delta_t = r_t + \gamma V_\psi(s_{t+1}) - V_\psi(s_t) \tag{3}$$

Then, the generalized advantage estimate is computed recursively as:

$$A_t = \delta_t + (\gamma\lambda)\delta_{t+1} + (\gamma\lambda)^2\delta_{t+2} + \cdots = \sum_{t=0}^{T}(\gamma\lambda)^l\delta_{t+l}^V \tag{4}$$

where γ is the discount factor, l represents the time step offset between the current step t and a future step, λ controls the trade-off between bias and variance:

- $\lambda = 1$: high variance but low bias
- $\lambda = 0$: low variance but high bias

To avoid the problem of policies converging to the local optimum, PopArt [43] is used to normalize the advantage function at all time steps, increasing the diversity among agents and making the policies of different agents more diverse. The normalized advantage function is as follows:

$$\hat{A}_t = \sigma(A_t - \mu)/\sigma \tag{5}$$

where μ and σ represent the mean and standard deviation, respectively. Based on the magnitude of gradient updates, PopArt adaptively adjusts the learning rate during training, significantly improving training efficiency and performance.

The PPO algorithm introduces a truncation function to limit the magnitude of policy updates, shown in Fig. 1. This truncation function, called the Proximal Clip Objective (Clip), restricts the probability ratio $r_t(\theta)$ updates to a small range $(1-\epsilon, 1+\epsilon)$, where ϵ is a hyperparameter, thus avoiding instability caused by excessively large policy updates.

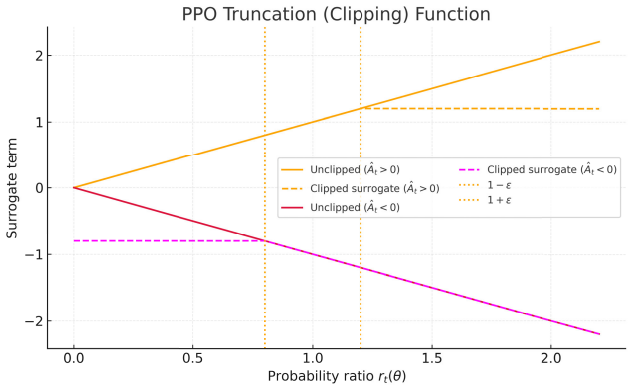

FIGURE 1. PPO truncation (clipping) function.

- for $\hat{A}_t > 0$, the curve saturates at $(1+\epsilon)\hat{A}_t$ when $r_t(\theta) \geq (1+\epsilon)$ (right-side truncation).
- for $\hat{A}_t < 0$, the curve saturates at $(1-\epsilon)\hat{A}_t$ when $r_t(\theta) \leq (1+\epsilon)$ (left-side truncation).
- inside $(1-\epsilon, 1+\epsilon)$, the clipped and unclipped terms coincide.

To ensure coordination, each agent's state can include its local traffic measurements, as well as the previous action of a downstream agent. MAPPO is a variant of PPO applied to the multi-agent setting. Each agent's actor is updated via a clipped surrogate objective to ensure stable policy

improvements. Concretely, the collective objective can be written as:

$$max\, J(\theta) = \mathbb{E}_t \left[r_t(\theta)\hat{A}_t, \text{clip}\big(r_t(\theta), 1-\epsilon, 1+\epsilon\big)\hat{A}_t \right] \quad (6)$$

where $\mathbb{E}_t$ is the expectation over time steps.

The MAPPO training loop operates as follows: at each time step, all agents observe the current traffic state (from the simulation environment) and each agent sequentially selects an action according to its policy. We adopt a sequential decision scheme to enforce step-down constraints: agents choose speed limits one after another in a fixed order, passing information through their state. Once every agent has taken the corresponding action for that step, the simulator advances and returns new observations and rewards. All observed transitions (s_t, a_t, r_t, s_{t+1}) from all agents are stored in a shared experience buffer D.

After collecting enough experiences, we perform MAPPO updates. Using the collected trajectories, we compute advantage estimates using the centralized critic value and then update each agent's policy by maximizing the PPO objective over several epochs. The centralized critic is updated to fit the observed cumulative rewards. The updated network weights are applied to all agents for the next training due to the parameter sharing. The training will continue for many episodes until convergence is achieved.

In decentralized execution, each agent uses only its local actor network. At each time step, it reads its local sensors and outputs a speed limit without needing global information. The training process, however, has ensured coordination via the shared critic and sequential decision scheme. This MAPPO-based MARL framework thus enables scalable VSL control, agents learn to coordinate posting speed limits across the network while reacting to local congestion, all without a central controller at execution time.

The optimization framework of the MAPPO algorithm for VSL control is shown in Fig. 2. The entire process framework consists of two parts: the SUMO simulation environment and the MAPPO-VSL agents.

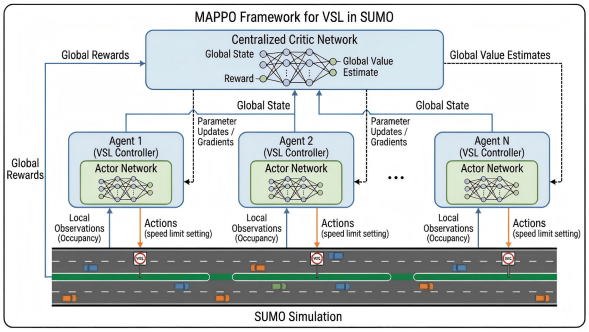

FIGURE 2. MAPPO framework for VSL control.

III. SIMULATION ENVIRONMENT

A. NETWORK SCENARIO

The implementation of MAPPO algorithms for Variable Speed Limit (VSL) control requires an environment that accurately captures the stochastic and dynamic characteristics of real-world traffic flow. Analytical or macroscopic simulation models, which represent traffic through aggregate variables such as average flow and density, are insufficient for RL-based control due to their limited temporal and spatial resolution. By contrast, microscopic traffic simulation provides a vehicle-level representation of the traffic system, in which the acceleration, deceleration, and lane-changing behaviors of vehicles are explicitly modeled according to driver and vehicle dynamics.

In MAPPO-based VSL control, agents learn optimal control policies through continuous interaction with the environment via a trial-and-error process. This process involves observing traffic states, applying control actions, and receiving feedback in the form of rewards. The microscopic simulation platform SUMO enables such interactions in a safe, controllable, and repeatable virtual environment. It allows agents to explore a wide range of traffic conditions, including congestion, stop-and-go waves, and mixed autonomy scenarios, without the risks and costs associated with real-world experiments.

Furthermore, microscopic simulators support multi-agent environments, which are crucial for implementing distributed or cooperative VSL control strategies. Each VSL controller can be represented as an independent learning agent interacting with a shared environment, enabling the evaluation of coordination among multiple control sections. Microscopic simulation also allows the modeling of heterogeneous traffic flows, including CAVs and HDVs, thereby reflecting the transitional characteristics of modern urban networks.

Another key advantage is the microscopic traffic simulator's ability to capture second-by-second variations in speed, headway, and vehicle interactions, which are essential for defining precise state and reward functions in MAPPO. The feedback generated by these microscopic interactions enables the RL agents to learn nuanced control strategies, such as dynamic speed harmonization and congestion mitigation near intersections.

Therefore, microscopic traffic simulation serves as an indispensable tool in the design, training, and evaluation of MAPPO-based VSL control systems. It bridges the gap between theoretical algorithm development and real-world applicability, providing a scalable, reproducible, and high-fidelity testing environment for intelligent urban traffic control.

To evaluate the proposed MAPPO-based Variable Speed Limit (VSL) control strategy, a large-scale urban traffic network was constructed in the microscopic traffic simulator SUMO. The Traffic Control Interface (TraCI) enables real-time interaction between the MAPPO-VSL controller and the traffic simulation environment. The generation of the urban traffic network is based on the open data provided by OpenITS [44]. The research area is located in Hefei Economic and Technological Development Zone, Anhui Province, China. The main road

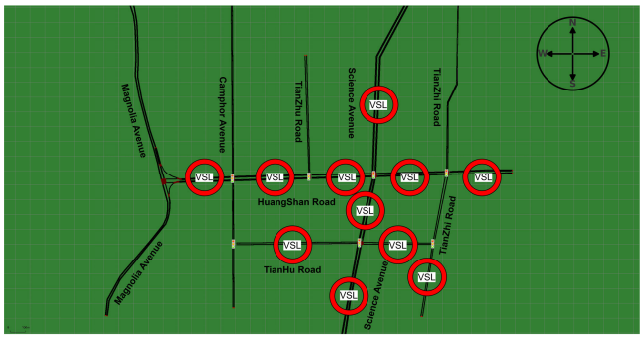

FIGURE 3. The real-world study area (31°50′24.8″N 117°12′03.1″E).

in the study area is `HuangShan` Road, which includes seven intersections and their adjacent road sections. Fig. 3 illustrates the geometric structure of the network in SUMO. The approaching road segments of intersections are controlled by the MAPPO-VSL framework. Including the first two sections of HuangShan Road from east to west (`HuangShanW1`,`HuangShanW2`) and the first three sections from west to east (`HuangShanE1`,`HuangShanE2`,`HuangShanE3`); Two sections of KeXue Avenue from north to south (`KeXueS1`,`KeXueS2`) and two sections from south to north (`KeXueN1`,`KeXueN2`); The first section of TianZhi Road from north to south (`TianZhiN1`); The sections of TianHu Road from east to west (`TianHuW1`) and from west to east (`TianHuE1`). A total of 12 MAPPO-VSL controllers were deployed in the study area. The signal control schemes for intersections in the study area fixed schedule. The phase duration settings match the real-world implementation. It should be noted that, unlike most traffic control schemes in Europe, the right-turn traffic lights at all intersections are constantly green.

To test the universality of the proposed MAPPO-VSL control scheme under different traffic demand distributions according to historical data from the detectors in this area, three different traffic load conditions were modeled and simulated as follows:

- Free-flow traffic:
  The traffic flow operates below the intersection's capacity. Vehicles arriving during the green phase can pass the stop line without delay. The available green time is not fully utilized.
- Capacity flow:
  The intersection operates at its effective capacity. The green phase is fully utilized, and vehicles depart continuously during the green interval. No residual queue remains.
- Extreme, perturbed traffic flow:
  The incoming demand exceeds the discharge capacity of the intersection, resulting in a residual queue that persists into the next signal cycle.

The traffic demand distribution for different load conditions is shown in Fig. 4.

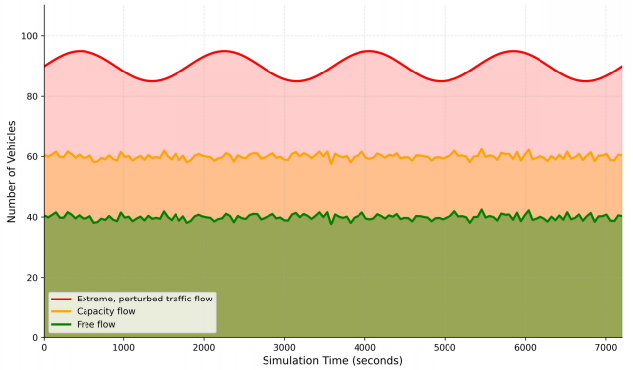

FIGURE 4. The applied different traffic demands.

B. TRAINING DETAILS OF MAPPO-BASED VARIABLE SPEED LIMIT STRATEGY

To integrate the dynamic traffic simulation environment generated in SUMO with the MAPPO algorithm, a customized reinforcement learning environment class `SUMO_Env()` was embedded in PyCharm. This environment allows reinforcement-learning agents to observe traffic states, apply actions, and receive feedback as rewards. This environment encapsulates all essential functional components required for multi-agent training. Specifically, it includes functions for initializing the simulation, state representation, action space definition, reward formulation, episode termination, and the fundamental `reset()` and `step()` functions. The environment enables each agent to interact with the traffic network in real-time, exchange state and reward information, and iteratively update its policy within the MAPPO training loop. The detailed definition is as follows:

- The initialization function configures the essential parameters for the simulation environment, setting the simulation duration to 2 hours (7200 seconds). Consistent with local traffic regulations, the initial speed limit for urban road sections is defined as 19.44 m/s.
- The state-space representation in this research is based on vehicle occupancy data collected from designated control road segments. Occupancy refers to time-occupancy (detector occupancy rate), for each controlled segment, loop detectors report the fraction of an aggregation interval δt during which the detector is occupied by a vehicle. In SUMO, this corresponds to loop detector occupancy outputs, aggregated per segment and concatenated into the agents' observation vector.
- The action space, corresponding to the actions available to PPO agents, comprises six discrete speed limit options: [22.22, 19.44, 16.67, 13.89, 11.11, 8.33] m/s. The agents collaboratively optimize these speed limits across different road sections.

- A critical component of reinforcement learning is the reward function, which encodes the optimization objectives of the training process. We design a centralized reward for the cooperative multi-agent VSL controller. Let $\mathcal{E}$ denote the set of controlled edges ($|\mathcal{E}| = 12$). At each simulation step t, we compute three traffic-related costs: carbon dioxide emissions, travel time, and a safety surrogate based on the time-to-collision (TTC).
  - Emissions and travel time: SUMO provides edge-level CO_2 emission and travel time measurement. We aggregate them over $\mathcal{E}$:

$$E_t = \sum_{e\in\mathcal{E}} \frac{\mathrm{CO_2Emission}_e(t)}{10^6}\ [kg], \tag{7}$$

$$T_t = \sum_{e\in\mathcal{E}} \frac{\mathrm{Traveltime}_e(t)}{3600}\ [h]. \tag{8}$$

  - TTC-based safety penalty: For each edge e, we compute the minimum TTC among vehicles on that edge:

$$\mathrm{TTC}_{e,t} = \min_{j\in\mathcal{V}_e(t)} \begin{cases} \dfrac{d_{j,t}}{v_{j,t} - v_{\ell(j),t}}, & v_{j,t} > v_{\ell(j),t}, \\ +\infty, & \text{otherwise}, \end{cases} \tag{9}$$

    where $d_{j,t}$ is the gap to the leader vehicle $\ell(j)$ and $v_{j,t}$ and $v_{\ell(j),t}$ are the speeds of the vehicle and its leader. Using a TTC threshold $\tau = 3$ s, we define an edge-level violation and a smooth penalty:

$$P_t = \sum_{e\in\mathcal{E}} \begin{cases} \dfrac{\tau - \mathrm{TTC}_{e,t}}{\tau}, & \mathrm{TTC}_{e,t} < \tau, \\ 0, & \mathrm{TTC}_{e,t} \geq \tau. \end{cases} \tag{10}$$

  The final reward is:

$$r_t = -\left(w_E \bar{E}_t + w_T \bar{T}_t + w_P \bar{P}_t\right), \tag{11}$$

  with weights $w_E = 0.5$, $w_T = 0.3$, and $w_P = 0.2$. Maximizing r_t therefore corresponds to minimizing a weighted combination of emissions, travel time, and TTC risk.
- The training concludes once a predefined iteration threshold is reached, at which point the interface between the MAPPO framework and the SUMO simulation is terminated. Before initiating a new episode, the environment is restored to its initial conditions through the `reset()` function.
- The `step()` function plays a pivotal role by executing the agents' chosen actions within the environment, subsequently collecting new observations, calculating rewards, and implementing speed limits. Following this, the function gathers environmental state information used for policy updates, calculates the global reward based on the predetermined reward function, and assesses whether the current environment state constitutes a terminal condition. The resulting speed limits represent the operational control outputs generated by the agents.

The complete training process is as follows:

1) Set simulation parameters: simulation duration (7200s), original speed limit (19.44 m/s), training iterations I, and number of PPO epochs per update; Initialize training parameter: discount factor γ, learning rates α.
2) Initialize the experience replay buffer D, actor network parameter θ, and critic network parameter ϕ.
3) The iteration begins based on the traffic scenario and obtains the initial traffic environment state s_1 from the detectors deployed on the simulated road segment.
4) Use the policy network π_θ to select actions.
5) The calculated speed limit actions are applied to the designated road segments via TraCI, regulating the maximum allowed speed for CAVs and HDVs traveling on those edges. The subsequent rewards r_t obtained and the next state s_{t+1} are observed from the SUMO simulation.
6) Calculate the advantage function at each time step, store the quadruple (s_t, a_t, r_t, s_{t+1}) as an empirical sample in D.
7) Update parameters θ using gradient descent and clip the gradient norm.
8) Update the old policy network π_θ, if the number of iterations does not reach the threshold, proceed to Step 4; otherwise, proceed to Step 9.
9) Training process complete.

Table 1 presents the hyperparameter settings used for the MAPPO algorithms during the training process. Due to the centralized learning setting, the actor and critic networks share parameters across all agents.

TABLE 1. Multi-agent proximal policy optimization hyperparameters for training.

Hyperparameter	Value
Number of agents (N)	12
Number of training iterations (I)	600
Episode length	300
N_rollout_threads	1
Learning rate (α)	3×10^{-4}
Discount factor (γ)	0.99
GAE parameter (λ)	0.95
PPO clip parameter (ϵ)	0.2
Number of PPO epochs per update	10
Number of hidden layers	3
Hidden layer size	64
Hidden layer activation function	ReLU

The detailed pseudo-code of MAPPO-VSL can be referred to Algorithm 1.

IV. EVALUATION

A. TRAINING RESULTS

The training dynamics of the proposed MAPPO-VSL framework under the three demand scenarios are summarized in Fig. 5. To address the significant difference in scales between the three traffic scenarios, the min-max normalization was applied to the rewards. This scales the reward for each scenario to a range of [0, 1], independent of their

Algorithm 1 MAPPO-VSL Training Process

Initialize policy network π_θ and centralized value network V_ϕ, discount factor γ, and learning rates α.
Initialize experience replay buffer D
for episode = 1 to I **do**
  Collect trajectories (s_t, a_t, r_t, s_{t+1}) from initial state s_1 for T time steps using policy π_θ
  Calculate reward r_t and discount factor γ for each time step
  Estimate value v_t of each state using value network V_ϕ
  Compute Generalized Advantage Estimation (GAE) for each time step using PopArt normalization
  Store trajectories (s_t, a_t, r_t, s_{t+1}) into experience replay buffer D
  for t = 1 to T **do**
    Sample a minibatch of transitions $\{s_t, a_t, \hat{A}_t\}$
    Compute policy loss:

$$L_\pi(\theta) = -\mathbb{E}\left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}\hat{A}_t\right]$$

    Compute value loss:

$$L_V(\phi) = \mathbb{E}\left[(V_\phi(s_t) - R_t)^2\right]$$

    where $R_t = r_t + \gamma v_{t+1} + \cdots + \gamma^{T-t} v_T$
    Update parameters θ using gradient descent:

$$\theta \leftarrow \theta - \alpha\nabla_\theta L_\pi(\theta), \quad \phi \leftarrow \phi - \alpha\nabla_\phi L_V(\phi)$$

    Update old policy network: $\pi_{\theta_{old}} \leftarrow \pi_\theta$
  end for
end for

absolute reward magnitudes. The bold lines represent a moving average to visualize the trend, while the faint lines show the raw normalized data variability. The experiments were conducted on a computer equipped with 16 GB of RAM, a 16-core CPU, and an NVIDIA GeForce GTX 1660 Ti GPU. Training tasks take between 138 and 187 minutes, depending on the different traffic demands.

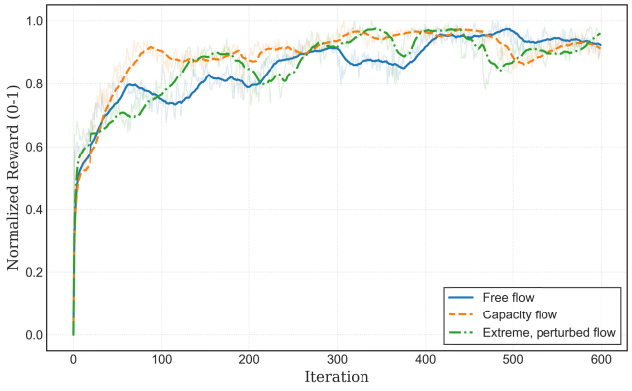

FIGURE 5. The learning curve under three traffic loads (see Fig. 4).

As shown in Fig. 5, the reward increases sharply during the early iterations, indicating that the emission-efficiency-safety reward provides an informative learning signal and allows the policy to rapidly escape weak initial behaviors. The reward then enters a slower improvement regime and gradually converges to a high-performance plateau. After approximately 250-350 iterations, all scenarios maintain normalized reward values above 0.9, suggesting that the learned policies are stable and near-converged under diverse traffic loads. Among the three cases, the capacity-flow scenario demonstrates the fastest early improvement and reaches a high reward level within the first 100 iterations, whereas the free-flow and extreme scenarios show slightly slower progress and more pronounced oscillations. Notably, a temporary reward degradation is observed around the late training stage (roughly iterations 470–510) for the capacity-flow and extreme-flow curves, after which the performance recovers, reflecting the inherent stochasticity of mixed traffic interactions and perturbation sensitivity in microscopic simulation.

To further interpret the learned behaviors beyond scalar rewards, Fig. 6 visualizes the spatiotemporal distributions of the speed-limit actions generated by the trained MAPPO-VSL policy. Each row corresponds to one controlled road segment, and each column denotes a control step over the evaluation horizon. The policy outputs one of six discrete speed limits in [22.22, 19.44, 16.67, 13.89, 11.11, 8.33] m/s, where warmer colors indicate more permissive limits and cooler colors indicate more restrictive limits.

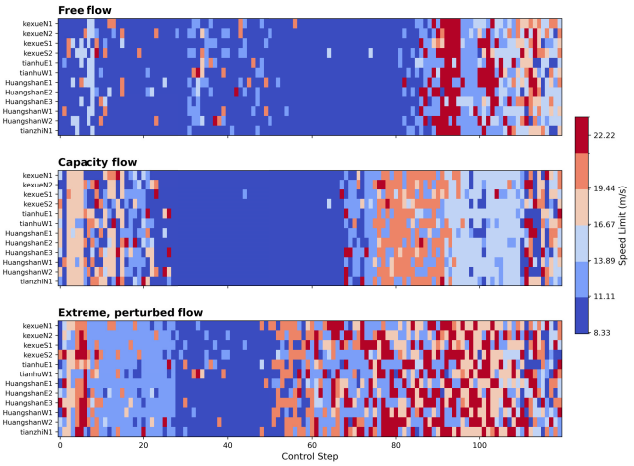

FIGURE 6. Spatial distribution of speed limit control policies for three distinct traffic demands.

In the free-flow scenario, the controller exhibits a largely consistent control regime over most of the horizon with only sporadic local adjustments, and it relaxes speed limits more noticeably toward the later part of the episode.

In the capacity-flow scenario, the policy quickly forms a coordinated, network-wide restrictive phase, characterized by almost uniform low-speed limits on various road sections for an extended period; subsequently, it entered a phase of partial relaxation, selectively restoring medium-to-high speed

limits on some peripheral road sections. This behavior is consistent with speed harmonization: proactively damping flow instabilities during critical operating conditions while avoiding unnecessary restrictions when the system becomes more stable.

In extreme, perturbed flow scenarios, the learned strategies become markedly more unstable, frequently switching between restrictive and permissive speed limits in time and space, indicating an adaptive response to time-varying congestion patterns and queuing dynamics caused by disturbances.

Overall, the learning curves and the control policy heatmaps confirm that the proposed MAPPO-VSL framework can reliably converge to high reward levels under different traffic demands, and learn temporally and spatially differentiated VSL strategies rather than applying a uniform limit across the entire network.

B. ENVIRONMENTAL AND TRAFFIC PERFORMANCE MEASURE

For a comprehensive assessment, the proposed MAPPO-based VSL controller is compared with three benchmark strategies: a no-control case with the default fixed speed limit, a rule-based VSL controller, and an IPPO-based VSL controller.

To benchmark the proposed MAPPO-VSL controller against a representative non-learning heuristic, we implement a rule-based variable speed limit (VSL) strategy driven by local observation. For each controlled road segment e at control step t, the controller reads its local occupancy $o_{t,e} \in [0, 1]$ and selects a discrete speed-limit action index $a_{e,t} \in \{0, 1, 2, 3, 4, 5\}$ via pre-defined thresholds:

$$a_{e,t} = \begin{cases} 0, & o_{t,e} < 0.15, \\ 1, & 0.15 \le o_{t,e} < 0.30, \\ 2, & 0.30 \le o_{t,e} < 0.45, \\ 3, & 0.45 \le o_{t,e} < 0.60, \\ 4, & 0.60 \le o_{t,e} < 0.75, \\ 5, & o_{t,e} \ge 0.75. \end{cases} \tag{12}$$

The published speed limit is then set to the corresponding discrete level $v(a_{e,t}) \in \{22.22, 19.44, 16.67, 13.89, 11.11, 8.33\}$ m/s, where a larger occupancy leads to a more restrictive speed limit. For a fair comparison, it operates on the same set of controlled segments and uses the same discrete action set and update frequency as the learned MAPPO-VSL policy during evaluation.

To study the effect of centralized training, we implement an IPPO ablation by replacing the centralized critic $V_\psi(s_t)$ with decentralized critics $V^i_\psi(o^i_t)$. All other components (actor architecture, observation/action spaces, reward, and PPO hyperparameters) are kept identical to MAPPO. Therefore, any performance gap can be attributed to the availability of global information in value estimation and the CTDE training paradigm.

TABLE 2. CO_2 emissions (kg) under different traffic demands.

Demand	No control	Rule-based	IPPO VSL	MAPPO VSL
Free-Flow	2,220 (0.0%)	2,253 (+1.5%)	2,157 (-2.8%)	2,141 (-3.6%)
Capacity-Flow	6,858 (0.0%)	7,103 (+3.6%)	6,638 (-3.2%)	6,536 (-4.7%)
Extreme-Flow	16,330 (0.0%)	16,378 (+0.3%)	16,307 (-0.1%)	16,300 (-0.2%)

The performance evaluation considers three traffic-related indicators: total carbon dioxide emissions, cumulative travel time, and TTC-based safety index. Carbon dioxide emissions and travel time are aggregated over all controlled edges during the 7200-s simulation horizon. Safety index computed from the number of time steps at which the minimum TTC on a controlled edge falls below 3 s:

$$safety\ index = 1 - \frac{ttc_{violations}}{episode\_length \times \mathcal{E}} \tag{13}$$

where $ttc_{violations}$ is count of edges whose $minTTC < 3s$ at that step. $\mathcal{E}$ is number of edges. The range of the safety index is [0, 1]. According to this definition, a larger safety index value indicates better safety performance; a value of 1 corresponds to zero TTC violations, whereas a value of 0 indicates a violation on every controlled edge at every time step.

The environmental results are summarized in Fig. 7 and Table 2. For CO_2 emission and surrogate safety index, we adopt a baseline-referenced visualization to highlight the marginal effect of each control strategy: for any metric $m(t)$, the plotted quantity is defined as $\Delta m(t) = m(t) - m_{\mathrm{NC}}(t)$, where the subscript NC denotes the no-control case. Accordingly, $\Delta CO_2(t) < 0$ indicates emission reduction relative to no control, while ΔSafety index$(t) > 0$ indicates improved safety relative to no control.

From the $\Delta CO_2(t)$ curves, the MAPPO-based VSL consistently achieves the largest emission reduction under free-flow and capacity-flow demands, with its trajectory remaining predominantly below zero and below the IPPO-based VSL curve, whereas the rule-based controller yields positive $\Delta CO_2(t)$ over extended intervals, implying higher emissions than the baseline. Under extreme flow, the emission advantage becomes noticeably smaller and more time-varying, yet MAPPO-VSL still provides the best result. These findings indicate that coordinated VSL is most effective for emission mitigation when the network operates near, but not far beyond, its effective capacity; once the system becomes heavily congested, the emission benefit remains positive but is naturally constrained by the high demand level.

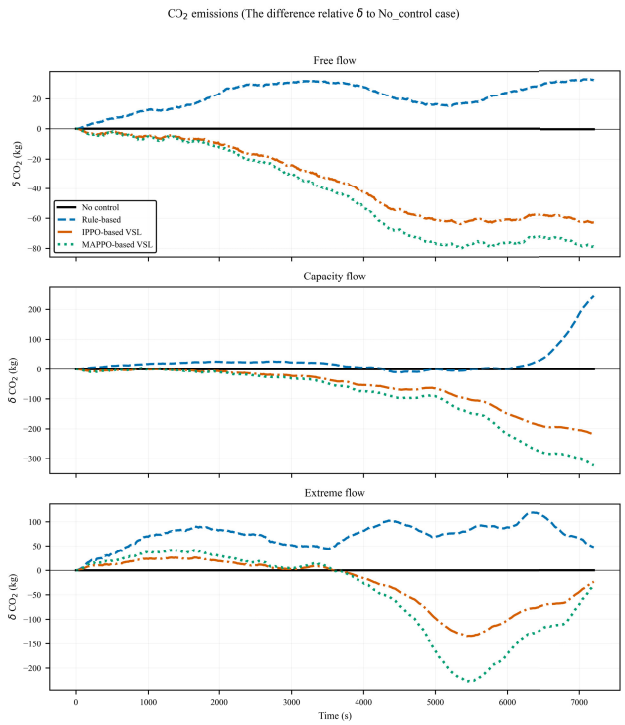

FIGURE 7. Comparison of carbon dioxide emissions under different traffic flow conditions.

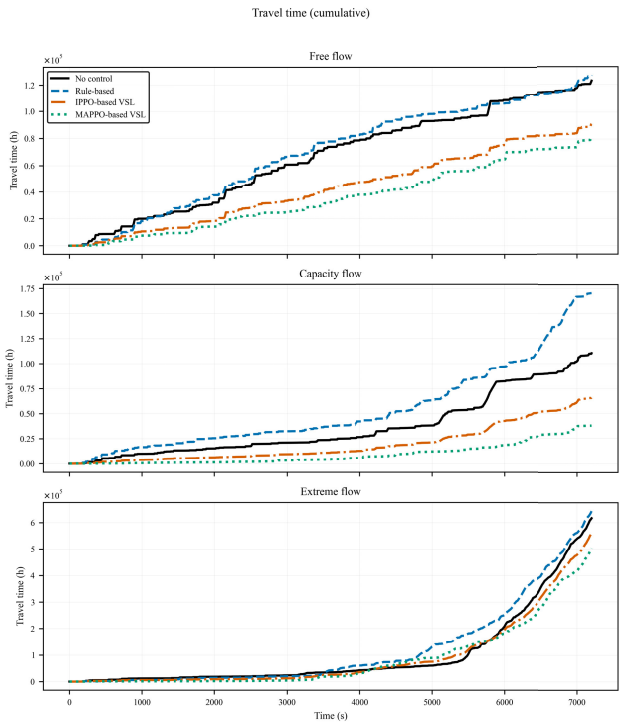

FIGURE 8. Comparison of cumulative travel time under different traffic flow conditions.

TABLE 3. Travel time (h) under different traffic demands.

Demand	No control	Rule-based	IPPO VSL	MAPPO VSL
Free-flow	122,781 (0.0%)	126,813 (+3.3%)	91,021 (-25.9%)	81,011 (-34.0%)
Capacity-flow	110,493 (0.0%)	170,509 (+54.3%)	65,654 (-40.6%)	37,752 (-65.8%)
Extreme-flow	616,693 (0.0%)	643,567 (+4.4%)	559,893 (-9.2%)	503,191 (-18.4%)

TABLE 4. Total TTC violations under different traffic demands.

Demand	No control	Rule-based	IPPO VSL	MAPPO VSL
Free-flow	23,413 (0.0%)	23,209 (-0.9%)	22,462 (-4.1%)	22,145 (-5.4%)
Capacity-flow	41,539 (0.0%)	41,908 (+0.9%)	40,151 (-3.3%)	39,364 (-5.2%)
Extreme-flow	57,536 (0.0%)	57,958 (+0.7%)	57,103 (-0.8%)	56,820 (-1.2%)

The efficiency gains are even more pronounced in Fig. 8 and Table 3. The cumulative travel time results further corroborate the efficiency benefit of MARL-based coordination. Across all three demands, MAPPO-VSL produces the lowest cumulative travel time, followed by IPPO-VSL, while the rule-based VSL strategy significantly increases travel time under capacity flow, indicating that purely local thresholding may induce adverse speed-limit patterns and exacerbate congestion spillback. The cumulative curves further show that MAPPO-VSL slows the growth of travel time throughout the episode, which suggests better queue containment and stronger suppression of stop-and-go propagation across the urban corridor.

The safety results in Fig. 9 and Table 4 further support the effectiveness of the proposed controller on improving traffic safety. Both learning-based methods yield persistent safety improvements relative to the no control case, with MAPPO-VSL exhibiting the most stable and largest positive margin, especially during the transient period after demand ramp-up. The rule-based VSL strategy does not provide a stable safety advantage and is even slightly worse than no control under the two heavier demand conditions. This result is important because it shows that locally reactive speed-limit rules can be too myopic in a signalized urban network. By contrast, MAPPO-VSL uses network-level value estimation during training to coordinate upstream and downstream actions, thereby achieving a better balance among mobility, emissions, and surrogate safety.

Overall, the evaluation demonstrates a consistent dominance of the MAPPO-based VSL controller over all benchmark methods across the three demand scenarios. In a fixed-time urban network, local occupancy is an imperfect

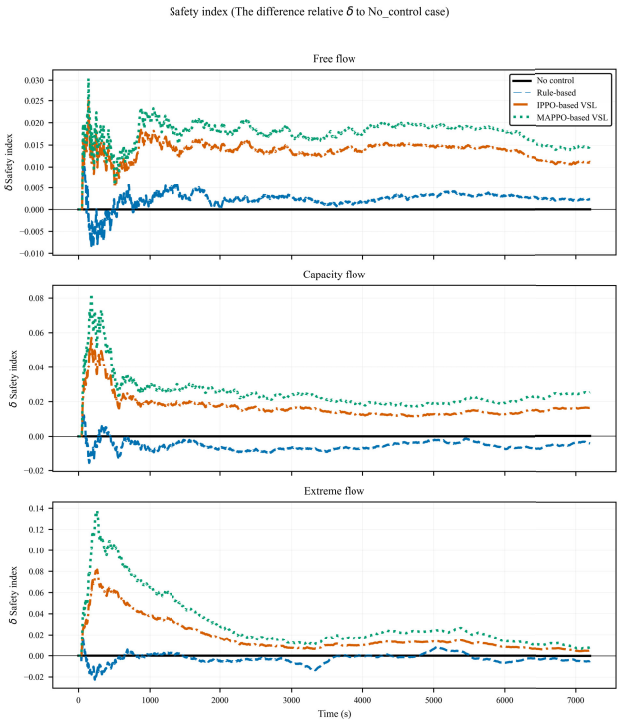

FIGURE 9. **Comparison of safety index under different traffic flow conditions.**

proxy for the true traffic state, because a high occupancy level may represent either efficient platoon discharge during a fully utilized green phase or the onset of downstream spillback. As a result, the heuristic controller (e.g. Rule-based VSL) can overreact to normal signal-induced compression, impose unnecessarily restrictive limits, and reduce progression efficiency rather than stabilizing the corridor. By contrast, MAPPO-VSL is trained under the CTDE paradigm with a centralized critic and a sequential decision mechanism, enabling agents to learn cooperative speed harmonization policies across the 12 controlled sections rather than isolated threshold reactions. The demand-dependent differences among the cumulative curves are also consistent with classical traffic-flow theory. In the free-flow condition, traffic demand remains below intersection capacity, queues are discharged within each signal cycle, and the available green time is not fully utilized. Under such conditions, the system is already close to an efficient operating point, so the benefit of VSL is naturally limited, and the main requirement is that the controller remain non-intrusive. This is exactly what the learned MAPPO policy exhibits: the control pattern is largely uniform, with only sporadic local interventions and a tendency to relax limits later in the episode. In the capacity-flow, the network operates near a critical state in which the green phase is fully utilized, and small disturbances can trigger residual queues and shockwaves. Accordingly, this scenario shows the largest divergence among the curves. The learned MAPPO policy rapidly forms a coordinated, network-wide restrictive phase and then selectively relaxes peripheral links, which is consistent with proactive speed harmonization; as a result, it substantially flattens the growth of cumulative travel time and suppresses emission accumulation. Under extreme flow, incoming demand exceeds discharge capacity and residual queues persist across cycles, so no strategy can fully restore uncongested operation. In this condition, the practical role of VSL is no longer to prevent congestion altogether, but to delay breakdown, contain queue propagation, and reduce the severity of stop-and-go oscillations. This explains why all travel time curves rise sharply in the later stage, while MAPPO-VSL still ends with the lowest cumulative travel time and the fewest TTC violations. The performance gap relative to IPPO-VSL confirms that the centralized critic is not merely an implementation detail, but a key contributor to network-level coordination, especially under capacity-flow conditions where interactions among adjacent links are strongest.

V. CONCLUSION

This paper presented a coordinated variable speed limit control framework for mixed urban traffic based on the MAPPO algorithm and the centralized-training decentralized-execution paradigm. The proposed method was implemented in a microscopic SUMO environment representing a real urban arterial network with 12 controlled road sections, mixed CAVs/HDVs traffic, and realistic fixed-time signal control. By combining shared multi-agent policies with a centralized value function and a reward that jointly considers carbon emissions, travel time, and TTC-based surrogate safety, the framework is able to learn coordinated VSL strategies for complex urban operations.

A key contribution of this work lies in enabling coordinated, network-level VSL control for mixed CAVs and HDVs in urban traffic. Unlike prior VSL approaches that treat control points in isolation or rely heavily on accurate traffic models, the proposed MAPPO-based method leverages centralized training with distributed execution to achieve cooperation across multiple arterial segments. This method enables agents to learn globally informed yet locally applicable control policies, thereby improving the scalability and robustness of VSL operations in complex urban environments. Another significant contribution is the development of a training and evaluation environment based on the microscopic traffic simulator SUMO and open real-world traffic data. The framework models a mixed urban network under three traffic conditions corresponding to free flow, capacity flow, and extreme, perturbed flow. The environment exposes step-wise objective metrics, including carbon dioxide emissions, travel time, and TTC-based surrogate safety. These metrics are further integrated into a scalarized multi-objective reward.

Simulation experiments across free-flow, capacity-flow, and extreme, perturbed traffic conditions demonstrate that the MAPPO-VSL strategy exhibits strong adaptability to diverse traffic demands. The coordinated VSL yields modest but

non-disruptive benefits in free-flow, the greatest gains near capacity, where shockwave suppression is most valuable, and meaningful resilience benefits under extreme, perturbed flow. The conventional rule-based VSL persistently underperforms and even exacerbates traffic conditions. This degradation occurs because the predefined occupancy thresholds often trigger overly restrictive speed limits prematurely, creating artificial moving bottlenecks and localized congestion. In contrast, the RL-based controllers learn to proactively adjust speeds before critical congestion materializes. Spatial analysis of learned policies further confirms that the agents are capable of identifying congestion-prone bottlenecks and applying spatially differentiated VSL actions rather than uniform speed reductions, reflecting a nuanced understanding of network topology and flow dynamics.

Ablation analysis results that centralized training is crucial. Although the IPPO ablation performs better than the no control case and heuristic controller, it remains consistently inferior to MAPPO, especially under capacity-flow conditions where cooperation among adjacent VSL locations is most critical. This confirms that global information in the critic helps the agents learn cooperative behaviors that cannot be recovered from purely local value estimation.

In summary, the results verify that MAPPO-based VSL control is a promising and scalable control layer for mixed urban traffic networks and that it can complement conventional signal control by improving environmental, operational, and safety-related outcomes simultaneously. Future work will focus on exploring continuous action spaces for speed limits rather than discrete levels, potentially enabling smoother speed transitions and greater passenger comfort. Furthermore, assessing the impact of varying CAVs penetration rates on control effectiveness could provide deeper insights into transition strategies toward fully automated traffic systems.

REFERENCES

[1] M. Aftabuzzaman, "Measuring traffic congestion—A critical review," *30th Australas. Transp. Res. Forum*, vol. 1, pp. 1–13, Sep. 2007.
[2] M. G. R. Alam, T. M. Suma, S. M. Uddin, M. B. A. K. Siam, M. S. B. Mahbub, M. M. Hassan, and G. Fortino, "Queueing theory based vehicular traffic management system through Jackson network model and optimization," *IEEE Access*, vol. 9, pp. 136018–136031, 2021.
[3] G. Weisbrod, D. Vary, and G. Treyz, "Measuring economic costs of urban traffic congestion to business," *Transp. Res. Rec.*, vol. 1839, no. 1, pp. 98–106, Jan. 2003.
[4] M. Treiber, A. Kesting, and C. Thiemann, "How much does traffic congestion increase fuel consumption and emissions? Applying a fuel consumption model to the NGSIM trajectory data," in *Proc. 87th Annu. Meeting Transp. Res. Board*, vol. 71, 2008, pp. 1–18.
[5] V. R. Vuchic, *Urban Transit Systems and Technology*. Hoboken, NJ, USA: Wiley, 2007.
[6] Q. Guo, L. Li, and X. (Jeff) Ban, "Urban traffic signal control with connected and automated vehicles: A survey," *Transp. Res. C, Emerg. Technol.*, vol. 101, pp. 313–334, Apr. 2019.
[7] M. Tajalli and A. Hajbabaie, "Dynamic speed harmonization in connected urban street networks," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 33, no. 6, pp. 510–523, 2018.
[8] Q. Lu and T. Tettamanti, "Traffic control scheme for social optimum traffic assignment with dynamic route pricing for automated vehicles," *Periodica Polytechnica Transp. Eng.*, vol. 49, no. 3, pp. 301–307, Sep. 2021.
[9] A. Sysoev, S. Zhikhoreva, V. Klyavin, and A. Pogodaev, "Neural networks approach to remodel capacity on urban road and street network," *Periodica Polytechnica Transp. Eng.*, vol. 53, no. 4, pp. 357–363, Jun. 2025.
[10] W. Genders and S. Razavi, "Asynchronous n-step Q-learning adaptive traffic signal control," *J. Intell. Transp. Syst.*, vol. 23, no. 4, pp. 319–331, Jul. 2019.
[11] S. Araghi, A. Khosravi, M. Johnstone, and D. Creighton, "Q-learning method for controlling traffic signal phase time in a single intersection," in *Proc. 16th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2013, pp. 1261–1265.
[12] S. S. Mousavi, M. Schukat, and E. Howley, "Traffic light control using deep policy-gradient and value-function-based reinforcement learning," *IET Intell. Transp. Syst.*, vol. 11, no. 7, pp. 417–423, Sep. 2017.
[13] S. G. Rizzo, G. Vantini, and S. Chawla, "Time critic policy gradient methods for traffic signal control in complex and congested scenarios," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 1654–1664.
[14] T. Zhao, P. Wang, and S. Li, "Traffic signal control with deep reinforcement learning," in *Proc. Int. Conf. Intell. Comput., Autom. Syst. (ICICAS)*, Dec. 2019, pp. 763–767.
[15] B. Kővári, L. Szőke, T. Bécsi, S. Aradi, and P. Gáspár, "Traffic signal control via reinforcement learning for reducing global vehicle emission," *Sustainability*, vol. 13, no. 20, p. 11254, Oct. 2021.
[16] B. Khondaker and L. Kattan, "Variable speed limit: An overview," *Transp. Lett.*, vol. 7, no. 5, pp. 264–278, Oct. 2015.
[17] D. Li and P. Ranjitkar, "A fuzzy logic-based variable speed limit controller," *J. Adv. Transp.*, vol. 49, no. 8, pp. 913–927, Dec. 2015.
[18] A. Coppola, L. Di Costanzo, L. Pariota, and G. N. Bifulco, "Fuzzy-based variable speed limits system under connected vehicle environment: A simulation-based case study in the city of naples," *IEEE Open J. Intell. Transp. Syst.*, vol. 4, pp. 267–278, 2023.
[19] X. Yang, Y. Lin, Y. Lu, and N. Zou, "Optimal variable speed limit control for real-time freeway congestions," *Proc.-Social Behav. Sci.*, vol. 96, pp. 2362–2372, Nov. 2013.
[20] S. Li, T. Wang, H. Ren, B. Shi, and X. Kong, "Optimization model and method of variable speed limit for urban expressway," *J. Adv. Transp.*, vol. 2021, pp. 1–13, May 2021.
[21] X.-Y. Lu and S. Shladover, "MPC-based variable speed limit and its impact on traffic with V2I type ACC," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 3923–3928.
[22] L.-Y. Yang, M. Zhao, D.-H. Sun, and H. Liu, "Variable speed limit control scheme for vehicle in the vicinity of signalized intersection," *Mod. Phys. Lett. B*, vol. 32, no. 19, Jul. 2018, Art. no. 1850218.
[23] B. Othman, G. De Nunzio, D. Di Domenico, and C. Canudas-De-Wit, "Analysis of the impact of variable speed limits on environmental sustainability and traffic performance in urban networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 21766–21776, Nov. 2022.
[24] A. Berbar, A. Gastli, N. Meskin, M. A. Al-Hitmi, J. Ghommam, M. Mesbah, and F. Mnif, "Reinforcement learning-based control of signalized intersections having platoons," *IEEE Access*, vol. 10, pp. 17683–17696, 2022.
[25] X. Fang, H. Li, T. Tettamanti, A. Eichberger, and M. Fellendorf, "Effects of automated vehicle models at the mixed traffic situation on a motorway scenario," *Energies*, vol. 15, no. 6, p. 2008, Mar. 2022.
[26] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.-P. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner, and E. Wiessner, "Microscopic traffic simulation using SUMO," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 2575–2582. [Online]. Available: https://elib.dlr.de/124092/
[27] B. Hellinga and M. Mandelzys, "Impact of driver compliance on the safety and operational impacts of freeway variable speed limit systems," *J. Transp. Eng.*, vol. 137, no. 4, pp. 260–268, Apr. 2011.
[28] Y. Wu, M. Abdel-Aty, L. Wang, and M. S. Rahman, "Combined connected vehicles and variable speed limit strategies to reduce rear-end crash risk under fog conditions," *J. Intell. Transp. Syst.*, vol. 24, no. 5, pp. 494–513, Sep. 2020.
[29] E. Grumert and A. Tapani, "Impacts of a cooperative variable speed limit system," *Proc.-Social Behav. Sci.*, vol. 43, pp. 595–606, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877042812010142
[30] M. Yu and W. D. Fan, "Optimal variable speed limit control in connected autonomous vehicle environment for relieving freeway congestion," *J. Transp. Eng., A, Syst.*, vol. 145, no. 4, Apr. 2019, Art. no. 04019007.

[31] H. Yu, R. Jiang, Z. He, Z. Zheng, L. Li, R. Liu, and X. Chen, "Automated vehicle-involved traffic flow studies: A survey of assumptions, models, speculations, and perspectives," *Transp. Res. C, Emerg. Technol.*, vol. 127, Jun. 2021, Art. no. 103101.
[32] X. Fang, T. Péter, and T. Tettamanti, "Variable speed limit control for the motorway–urban merging bottlenecks using multi-agent reinforcement learning," *Sustainability*, vol. 15, no. 14, p. 11464, Jul. 2023.
[33] M. Treiber and D. Helbing, "Realistische mikrosimulation von strassenverkehr mit einem einfachen modell," in *Proc. 16th Symp. Simulation stechnik ASIM*, 2002, p. 80.
[34] M. Garg and M. Bouroche, "Can connected autonomous vehicles improve mixed traffic safety without compromising efficiency in realistic scenarios?" *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 6, pp. 6674–6689, Jun. 2023.
[35] E. Mitsakis, J. F. Erdmann, A. Yun-Pang, M. R. Robert, K. Carlier, R. Blokpoel, and S. Boerma, "Transaid deliverable 3.1-modelling, simulation and assessment of vehicle automations and automated vehicles' driver behaviour in mixed traffic-iteration 2," European Union"s Horizon 2020 Project, Tech. Rep. D3.1, 2019.
[36] R. S. Sutton, "Reinforcement learning: An introduction," in *Reinforcement Learning: An Introduction*. Newton Center, MA, USA: Branford, 2018.
[37] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.
[38] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 12, 1999, pp. 1057–1063.
[39] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel, "Trust region policy optimization," 2015, *arXiv:1502.05477*.
[40] A. Shojaeighadikolaei, Z. Talata, and M. Hashemi, "Centralized vs. Decentralized multi-agent reinforcement learning for enhanced control of electric vehicle charging networks," 2024, *arXiv:2404.12520*.
[41] V. R. Konda and J. N. Tsitsiklis, "Actor-critic algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002.
[42] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," 2015, *arXiv:1506.02438*.
[43] J. W. Leigh and D. Bryant, "Popart: Full-feature software for haplotype network construction," *Methods Ecol. Evol.*, vol. 6, no. 9, pp. 1110–1116, Sep. 2015.
[44] B. Zhang and Z. Sha. (2021). *Openits Org. Opendata V6.0—Introduction of Open Data of Hefei Demonstration Area*. [Online]. Available: http://www.openits.cn/openData2/602.jhtml

XUAN FANG received the B.Sc. degree in vehicle engineering from Chongqing University of Technology, China, in 2017, and the M.Sc. degree in vehicle engineering from Budapest University of Technology and Economics, Budapest, Hungary, in 2019, where he is currently pursuing the Ph.D. degree with the Department of Control for Transportation and Vehicle Systems.

His research interests include intelligent traffic system modeling and control, microscopic road traffic simulation, and reinforcement learning-based variable speed limits control.

ISTVÁN VARGA received the Ph.D. degree in traffic engineering from Budapest University of Technology and Economics, in 2006, and the D.Sc. degree in traffic engineering from the Hungarian Academy of Sciences, in 2020.

He was the Dean of the Faculty of Transportation Engineering and Vehicle Engineering, Budapest University of Technology and Economics, from 2012 to 2025. He is currently the Vice-Rector for Strategy at Budapest University of Technology and Economics, holding the position of a Full Professor. He participates in research and industrial projects both as a researcher and project coordinator. He is the co-author of more than 140 scientific articles, two patents, and several books. His main research interests include traffic flow theory, intelligent transportation systems, autonomous vehicles, application of control theory in road traffic management. He is a member of the Public Body of the Hungarian Academy of Sciences and a member of the Committee on Transport Science and Vehicle Science of the Hungarian Academy of Sciences.

TAMÁS TETTAMANTI received the Ph.D. degree in traffic engineering from Budapest University of Technology and Economics, in 2013, and the D.Sc. degree in traffic engineering from the Hungarian Academy of Sciences, in 2023.

He is currently a Full Professor with the Faculty of Transportation Engineering and Vehicle Engineering, Budapest University of Technology and Economics. He participates in research and industrial projects both as a researcher and project coordinator. He is the co-author of more than 180 scientific articles, two patents, and several books. His main research interests include road traffic modeling, estimation, and control in the cooperative, connected, and automated mobility (CCAM) field, and related co-simulation technology developments. He is a member of the Public Body of the Hungarian Academy of Sciences. He serves as an Editorial Board Member for IEEE Transactions on Intelligent Transportation Systems and *Communications in Transportation Research* journal. He is a member of the Technical Committee "TC 7.4 Transportation Systems" at the International Federation of Automatic Control (IFAC).