applied sciences

MDPI

Article

Multi-Agent Reinforcement Learning for Traffic State Estimation on Highways Using Fundamental Diagram and LWR Theory

Xulei Zhang and Yin Han *

Department of Transportation System Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China; 232301111@st.usst.edu.cn
* Correspondence: hanyin2000@126.com

Abstract

Traffic state estimation (TSE) is a core task in intelligent transportation systems (ITSs) that seeks to infer key operational parameters—such as speed, flow, and density—from limited observational data. Existing methods often face challenges in practical deployment, including limited estimation accuracy, insufficient physical consistency, and weak generalization capability. To address these issues, this paper proposes a hybrid estimation framework that integrates multi-agent reinforcement learning (MARL) with the Lighthill–Whitham–Richards (LWR) traffic flow model. In this framework, each roadside detector is modeled as an agent that adaptively learns fundamental diagram (FD) parameters—the free-flow speed and jam density—by fusing local detector measurements with global CAV trajectory sequences via an interactive attention mechanism. The learned parameters are then passed to an LWR solver to perform sequential (rolling) prediction of traffic states across the entire road segment. We design a reward function that jointly penalizes estimation error and violations of physical constraints, enabling the agents to learn accurate and physically consistent dynamic traffic state estimates through interaction with the physics-based LWR environment. Experiments on simulated and real-world datasets demonstrate that the proposed method outperforms existing models in estimation accuracy, real-time performance, and cross-scenario generalization. It faithfully reproduces dynamic traffic phenomena, such as shockwaves and queue waves, demonstrating robustness and practical potential for deployment in complex traffic environments.

Keywords: traffic state estimation; multi-agent reinforcement learning; LWR model; physics-informed; intelligent transportation systems

Check for updates

Academic Editor: Thomas Lindner

Received: 26 December 2025
Revised: 19 January 2026
Accepted: 21 January 2026
Published: 24 January 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Traffic state estimation (TSE) is a core component of intelligent transportation systems (ITSs) [1–3]. TSE aims to infer the operational state of a road network—characterized by flow, speed, and density—using limited observational information [4]. Accurate TSE underpins traffic management and control, signal optimization, route guidance, and congestion mitigation and is critical for enabling smart mobility and smart cities [5].

From an engineering and policy perspective, traffic congestion on highways has been widely recognized as a major source of economic loss, energy waste, and environmental impact. Large-scale transportation studies have shown that congestion-related delays and inefficiencies impose substantial societal costs and that the effectiveness of many ITS strategies critically depends on the availability of accurate and timely traffic state estimation.

For TSE methods to be viable in practice, they must satisfy several fundamental requirements. First, estimates must be highly accurate, because low precision directly undermines the effectiveness of traffic management and control strategies [6]. In practice, inaccurate real-time traffic state estimation may further propagate errors into downstream control and guidance systems, leading to inefficient or even counterproductive operational decisions. Second, real-time performance is essential given the rapidly evolving nature of traffic flow; delayed estimates cannot support online decision-making [7,8]. Third, estimates must remain physically plausible and comply with governing constraints. Because traffic flow evolution is governed by continuum conservation laws, estimates that violate these principles undermine interpretability and reliability [9]. Finally, strong generalization is required for real-world deployment: the method should maintain stable performance across detector configurations, network sizes, and topologies [1].

Recent research on TSE has largely fallen into three methodological categories. Supervised learning approaches fit models to historical observations and can achieve reasonable performance when data coverage is sufficient. Deep learning methods employ expressive neural architectures to capture spatiotemporal traffic patterns and have become a mainstream direction in recent years [10]. Optimization-based methods formulate TSE as a constrained mathematical program, using traffic flow models to infer states from detector data [11]. Despite their respective merits, all three approaches face notable limitations in practical deployment.

First, supervised learning methods depend heavily on accurately labeled data. However, sensor deployment and maintenance are costly, and spatiotemporally sparse measurements often prevent models from capturing the full network state [12]. Second, although deep learning methods can learn complex traffic patterns, their generalization is often limited: models are sensitive to detector layout and may fail when sensor locations change, and models trained on small road sections transfer poorly to larger networks. Thirdly, supervised learning inherently aims to minimize single-step prediction error, which can lead to accumulated errors during long sequential rolling prediction, weakening the representation of traffic flow evolution dynamics. Finally, while optimization-based methods can embed physical constraints [13], they often require solving computationally intensive programs, which hampers real-time estimation at second or sub-second timescales.

These limitations stem from a trade-off between data dependence and physical modeling: supervised and deep learning methods are primarily data-driven and often do not explicitly represent governing evolution laws, weakening generalization and temporal robustness. In contrast, optimization-based methods offer stronger physical consistency but are often computationally expensive and therefore less suitable for real-time applications. Accordingly, a method that jointly balances data dependence, generalization, physical consistency, and real-time performance remains needed.

To address these challenges, we propose a hybrid framework that integrates multi-agent reinforcement learning (MARL) with macroscopic traffic flow theory. Specifically, we cast TSE as a sequential decision-making problem in which each roadside detector acts as an agent. Using local observations and global trajectory data from Connected and Automated Vehicles (CAVs), agents learn to output fundamental diagram parameters (e.g., free-flow speed and jam density) for a macroscopic Lighthill–Whitham–Richards (LWR) model [14]. The learned parameters are then passed to an efficient LWR solver to generate rolling, high-resolution estimates of density, speed, and flow over the entire road segment. A reward function that accounts for both estimation error and physical-consistency violations encourages cooperative learning, enabling accurate, real-time inference while respecting constraints such as flow conservation.

The main contributions of this paper are threefold:

First, we develop a physics-guided multi-agent reinforcement learning architecture that couples data-driven policies with an LWR model solver. This architecture leverages reinforcement learning for long-horizon optimization while using the macroscopic model to ensure physically plausible state reconstruction. This coupling mitigates the limited physical interpretability of purely data-driven methods and the high computational burden of purely optimization-based approaches.

Second, we design a multi-source fusion strategy based on interactive attention. This strategy adaptively integrates local detector measurements with global CAV trajectory sequences, improving perception under partial observability and enhancing robustness to changes in sensor layout.

Third, we construct a comprehensive simulation and evaluation suite to validate the proposed method across diverse traffic scenarios (e.g., fixed and variable highway lengths and different fluctuation patterns). Results from simulation and real-world data demonstrate that the proposed method outperforms representative baselines in estimation accuracy, real-time performance, physical consistency, and generalization.

The remainder of this paper is structured as follows: Section 2 reviews related work. Section 3 formally defines the problem. Section 4 details the proposed methodological framework. Section 5 describes the experimental setup, baseline models, and evaluation metrics. Section 6 presents and discusses the experimental results. Section 7 concludes the paper and outlines future research directions.

2. Literature Review

Traffic state estimation (TSE), a core component of intelligent transportation systems (ITSs), has evolved significantly, driven by advancements in data acquisition capabilities and modeling paradigms. This study categorizes existing TSE literature into three primary classes: physics-based methods, purely data-driven methods, and physics–data hybrid methods. Based on this categorization, we summarize key limitations in the current literature.

2.1. Physics Model-Based Traffic State Estimation Method

Early research primarily relied on macroscopic traffic flow theory to construct state estimators, explicitly encoding governing laws using physical principles (e.g., vehicle conservation). Li et al. [15] pioneered a multimodel ensemble framework that dynamically fused various macroscopic models, such as the Cell Transmission Model (CTM), using machine learning algorithms (e.g., Lasso), effectively enhancing estimation robustness under model parameter uncertainty. Fountoulakis et al. [16] eliminated the need for complex calibration of fundamental diagram (FD) models by relying on the vehicle-conservation law and integrating speed data from low-penetration Connected Vehicles (CVs) with flow data from sparse fixed detectors to estimate density. The advantage of such methods lies in their strong physical interpretability, but their estimation accuracy heavily depends on the validity of model assumptions and high-quality parameter calibration. To further enhance model realism, Zheng et al. [17] introduced a stochastic kinetic model that represents driver heterogeneity via random parameters, thereby generating smoother and more realistic trajectories. Takenouchi et al. [18] innovatively utilized vehicles in the opposing lane as moving sensors to obtain reverse flow data, combining it with Variational Theory to solve the LWR model, demonstrating the value of integrating novel data sources with physical models. Lee and Lee [19] proposed a congestion boundary framework that captures traffic phase transitions by identifying critical thresholds from bimodal traffic state distributions, offering an alternative macroscopic description beyond traditional fundamental diagram assumptions. Ni [20] demonstrated that the fundamental diagram is a reduced repre-

sentation of a more fundamental phase diagram derived from first principles of physics, reformulating traffic flow dynamics through a Hamiltonian framework that explains the physical origin and limitations of equilibrium-based macroscopic models.

2.2. Purely Data-Driven Traffic State Estimation Method

With the rise of artificial intelligence—particularly deep learning—numerous approaches have been applied to TSE that eschew explicit physical models and rely primarily on data-driven learning. These methods aim to learn complex nonlinear spatiotemporal mapping relationships within traffic states directly from data. Xu et al. [21] proposed the GE-GAN framework, which uses graph-embedding techniques to learn road-network structure and employs generative adversarial networks (GANs) to impute missing detector data, enabling fully data-driven state reconstruction. The work by Lei et al. [22] advanced this line of research by combining a conditional diffusion model with spatiotemporal networks to produce probabilistic traffic state estimates at sensor-free locations, thereby providing uncertainty quantification. Rowan et al. [23] presented a systematic review of machine learning-based microscopic traffic flow models, critically analyzing car-following and lane-changing approaches and identifying key challenges related to model transferability, interpretability, physical consistency, and scalability toward mesoscopic and macroscopic traffic applications. However, purely data-driven methods often exhibit limited physical interpretability and may generalize poorly beyond the training distribution. As a result, their predictions can violate fundamental physical constraints (e.g., conservation relationships).

2.3. Physics–Data Dual-Driven Traffic State Estimation Method

To balance physical consistency with data-driven flexibility, physics–data hybrid approaches have become a dominant research direction in TSE. Kawasaki et al. [24] extended the CTM state-space model to two-dimensional networks and used particle filtering to jointly estimate traffic states, OD matrices, and route-choice parameters, representing a typical integration of filtering techniques with physical models. In recent years, physics-informed neural networks (PINNs) have emerged as a powerful tool for realizing dual-driven modeling. Lu et al. [25] constructed a hierarchical computational graph, embedding continuum flow models and queueing fluid approximations as constraints into the network, achieving joint estimation of traffic states and queue profiles. Huang et al. [26] further integrated Bayesian neural networks with the LWR model, proposing a physics-informed Bayesian deep learning (PIBDL) framework that provides uncertainty quantification while ensuring physical consistency. The PIDL-MSTE framework proposed by Luo et al. [27] innovatively introduced ramp flow as an explicit source term into the model, learning road network topology and conservation constraints through neural networks, thereby addressing the estimation challenge in multi-ramp scenarios. Zhao et al. [28] focused on boundary sensing, combining deep learning with closed-loop PDE boundary observer theory to propose an Observer-Informed Deep Learning (OIDL) framework, which significantly improved estimation accuracy and stability under extreme conditions with only boundary data.

2.4. Limitations of Current Research

Despite the significant progress achieved by the aforementioned studies, existing TSE methods still face two key limitations. First, current methods often struggle to learn and predict the spatiotemporal evolution of traffic flow under long-horizon forecasting. Most deep learning-based models are trained with the objective of minimizing single-step prediction error. During multi-step rolling prediction, errors can accumulate and amplify over time, hindering accurate characterization of congestion formation, propagation, and dissipation.

This limitation partly arises because many models are not explicitly guided during training to learn the underlying physical mechanisms that govern traffic flow evolution.

Second, the generalization capability of existing models is often insufficient, limiting their ability to adapt to common configuration changes in real-world networks. When deployment scenarios deviate from training settings—e.g., changes in highway length or detector number/layout—performance often degrades, particularly for models that rely on fixed network structures. The time and economic costs of collecting large-scale data for retraining in new scenarios can be prohibitive, thereby hindering large-scale deployment of TSE in practice. Therefore, developing a TSE framework capable of adapting to different road network configurations, possessing strong generalization ability, and effectively learning the spatiotemporal evolution laws of traffic flow has become an urgent and important research task.

3. Problem Statement

Traffic state estimation on highways under mixed traffic flows involves inferring the traffic state across the entire highway segment using data from roadside detectors and trajectory data from Connected and Automated Vehicles (CAVs). Specifically, roadside detectors collect data from visible portions of the entire highway, such as the number of vehicles and average vehicle speed within a certain range upstream and downstream of the detector location, providing static local observations. CAV data, on the other hand, transmit trajectory data to the cloud in real time via vehicle–infrastructure cooperation systems, providing dynamic global observations. The estimation targets are traffic state variables, including flow, average speed, and density. This study selects these three states as the objectives for traffic state estimation. In previous research, some studies primarily used speed as the key estimation feature, while others used density; this study utilizes average speed and density, among other information.

Formally, this study categorizes the observable information into two types: roadside detector data and CAV trajectory data. The definitions of these two data types are as follows:

Definition 1. *Roadside detector data can be formally expressed as* $D_{fixed} = \{(x_i, t_j, v_{i,j}, \rho_{i,j}, q_{i,j}) | i \in [1, N], j \in [1, M]\}$, *where i and j are the indices for the roadside detector and sampling time point, respectively.* $(x_i, t_j, v_{i,j}, \rho_{i,j}, q_{i,j})$ *is a quintuple, representing the spatial coordinates of the ith fixed detector at the jth sampling time point,* $v_{i,j}$ *denotes the observed average speed at location* x_i *and time* t_j, $\rho_{i,j}$ *denotes the observed density at location* x_i *and time* t_j, *and*$q_{i,j}$ *denotes the observed flow at location* x_i *and time* t_j.

Definition 2. *CAV data can be formally expressed as* $D_{CAV} = \{(x_k(t_j), v_k(t_j), t_j) | k \in [1, K], j \in [1, M]\}$, *where k and j are the indices for the CAV and sampling time point, respectively.* $(x_k(t_j), v_k(t_j), t_j)$ *is a triple, representing the position* $x_k(t_j)$ *and speed* $v_k(t_j)$ *of the kth CAV at time* t_j.

In this study, for the traffic state estimation problem, the objective is to estimate the traffic state over the entire road segment using both roadside detector data and CAV trajectory data. The problem can be defined as follows:

Definition 3. *Given the roadside detector* D_{fixed} *and the CAV trajectory data* D_{CAV}, *the traffic state estimation problem can be formalized as finding a mapping function:*

$$f : \left(D_{fixed}, D_{CAV}\right) \rightarrow \{v(x,t), \rho(x,t), q(x,t)\}_{x \in [0,L], t \in [0,T]} \quad (1)$$

where $\{v(x,t), \rho(x,t), q(x,t)\}$ *represents the field of average speed, flow, and density within the spatiotemporal domain, where L means highway section index and T means the detection horizon. This mapping function must satisfy the mathematical expression*

$$\begin{aligned} &\min \sum_{i,j} \left\| v(x,t) - v_f\left(x,t \middle| D_{fixed}, D_{CAV}\right) \right\| + \left\| \rho(x,t) - \rho_f\left(x,t \middle| D_{fixed}, D_{CAV}\right) \right\| \\ &subject\ to\ \frac{\partial \rho(x,t)}{\partial t} + \frac{\partial q(\rho(x,t))}{\partial x} = 0 \\ &\mathrm{q} = \rho \cdot v \end{aligned} \tag{2}$$

That is, the estimated traffic state within the spatiotemporal domain should match the true traffic state, subject to the constraints of the flow conservation equation and the fundamental diagram relationship.

4. Methodology

4.1. Model Framework

The model framework is illustrated in Figure 1. Agents deployed at roadside detectors infer the traffic state in the next time step based on current local observations and global CAV trajectory information. Unlike conventional reinforcement learning frameworks, the proposed architecture explicitly integrates an LWR model-based solver into the loop: agents output fundamental diagram parameters, which are first fed into the LWR solver, and the solver then produces the traffic state estimation for the next time step. In this way, reinforcement learning is used to optimize long-horizon estimation performance through repeated interaction with a physics-based estimator rather than serving as an offline data-fitting module.

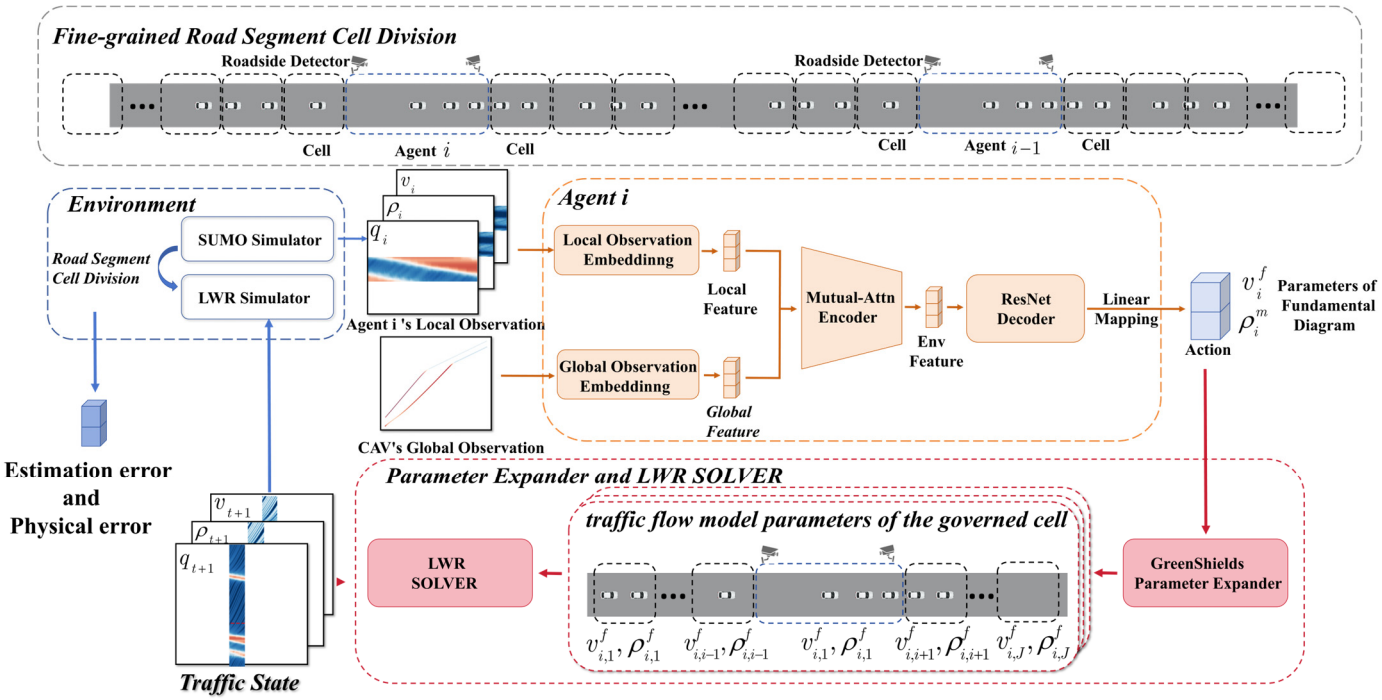

Figure 1. Method framework.

It is important to clarify that the above formulation does not imply control or intervention on the physical traffic system. Instead, the reinforcement learning framework is used to learn a sequential estimation policy that adaptively updates the internal parameters of the physics-based estimator (i.e., the fundamental diagram parameters governing the LWR solver). The environment interaction refers exclusively to the evolution of the estimator's

internal state through the LWR model, while the true traffic dynamics remain entirely exogenous. This formulation follows a growing line of research that casts physics-constrained inference and model calibration problems as sequential decision-making processes in order to explicitly account for long-horizon error accumulation.

In our formulation, the SUMO simulator provides the ground truth traffic dynamics and the corresponding detector/CAV observations, while the LWR-based simulator acts as the RL environment that propagates macroscopic states forward in time given the agents' actions. This design avoids a prohibitively high-dimensional action space that would arise if agents directly output full spatiotemporal traffic states and instead enables stable learning by letting agents adjust only low-dimensional physical parameters.

The remainder of this section is organized as follows: Section 4.2 describes the MDP, Section 4.3 details the agent design, Section 4.4 introduces the LWR solver, and Section 4.5 presents the MARL training algorithm.

4.2. Multi-Agent MDP Modeling for TSE

This study employs multi-agent reinforcement learning (MARL) for traffic state estimation on highways, transforming the temporal process of state estimation into a sequential decision-making problem. As described in Section 4.1, this MDP is represented by a 6-tuple $\left\langle s, o, a, P, r^{estimate}, r^{physic} \right\rangle$. The following sections detail the formal definitions of each element within the MDP from time step t to $t+1$.

Global State, s: This represents the environmental state obtained from the SUMO simulator, encompassing local observations from roadside detectors. Formally, it includes the roadside detector data as defined in Definition 1, $D_{fixed} = \left\{ \left(x_i, t_j, v_{i,j}, \rho_{i,j}, q_{i,j} \right) | i \in [1, N], j \in [1, M] \right\}$. Here, $v_{i,j}$, $\rho_{i,j}$, and $q_{i,j}$ represent the average speed, density, and flow, respectively, within the monitoring range of the i-th detector at time j.

Additionally, s includes Connected and Automated Vehicle (CAV) data. Within the detection time window from t to $t+1$, CAV transmission data is assumed to be uploaded to the cloud at 10 Hz. Therefore, the CAV data is treated as a time series comprising data from all CAVs present on the highway during this current time window.

In our MDP setting, one agent–environment interaction corresponds to a 1 s decision interval; thus, the CAV sequence within each step contains 10 samples (10 Hz) and is fed to the LSTM encoder as the global observation for that step.

Local Observation, $\overset{t}{o}_i$: In MARL, individual agents typically cannot directly access the global state and only receive partial observations. Thus, $\cup D^t_{CAV} \underset{i \in N}{\cup} o_i = s$, where o_i denotes the local observation of a single agent. Specifically, $\overset{t}{o}_i$ corresponds to the component of D^t_{fixed} associated with the agent's index i.

Action Space, $\overset{t}{a}_i$: In this study, each agent outputs the fundamental diagram parameters for the cell where its corresponding roadside detector is located in the current time step. These parameters include the free-flow speed $v^f_{i,t}$ and the jam density $\rho^m_{i,t}$. It is important to note that these two parameters are subject to specific upper- and lower-bound constraints. Although traffic flow exhibits strong spatial coupling due to vehicle interactions, the proposed agent partition does not assume independence between adjacent regions. Instead, each agent represents a local parameter adaptation unit anchored at a detector location, while the strong inter-cell coupling is explicitly preserved through the LWR conservation law, the shared reward design, and the parameter diffusion mechanism across agent boundaries.

To ensure physical plausibility and training stability, the continuous actions predicted by each agent are constrained to bounded intervals, as shown in Table 1. Specifically, the policy network outputs an unconstrained action vector, which is first squashed by a

Hyperbolic tangent function and then linearly re-scaled to $\left[v_f^{\min}, v_f^{\max}\right]$ and $\left[k_j^{\min}, k_j^{\max}\right]$, respectively. This design prevents nonphysical parameter values that may distort wave propagation in the LWR solver and destabilize MAPPO.

Table 1. Feasible range of fundamental diagram parameters used in the action space.

Parameter	Physical Meaning	Lower Bound	Upper Bound	Unit
v_f	Free-flow speed	10	40	m/s
k_j	Jam density	0	1	veh/m

Transition Dynamics, P: This is a state transition function describing how the global state changes upon taking actions. In this study, the state is directly obtained from the SUMO simulator. Concurrently, as mentioned in Section 4.1, the state of the LWR simulator from the previous time step also influences the traffic state estimation for the current time step. Therefore, the transition dynamics in this study are primarily represented by the LWR simulator.

Reward Function, $r^{estimate}, r^{physic}$: The reward function in this study consists of two components. The first, $r^{estimate}$, is the negative of the difference between the estimated traffic state and the true state. Since all agents interact through the LWR solver, a fully cooperative relationship among agents is assumed. This reward component can be expressed by Equation (3):

$$r_t^{estimate} = -\frac{1}{N}\sum_{i=1}^{N} w_v \|v(x_i,t) - v(x_i,t)\|^2 + w_\rho \|\rho(x_i,t) - \rho(x_i,t)\|^2 \tag{3}$$

Furthermore, since speed and density differ in both physical units and numerical scales, we introduce a scaling factor, w_v, w_ρ, to make the two error terms comparable. Specifically, w_v, w_ρ is selected by range-based normalization using the empirical bounds of the state variables in the training simulations:

$$\frac{\omega_v}{\omega_\rho} = \frac{v_f^{\max} - v_f^{\min}}{\rho_f^{\max} - \rho_f^{\min}} \tag{4}$$

According to the bounded intervals in Table 1, we use $\omega_v = 30$ and $\omega_\rho = 1$ for all experiments.

Note that flow is not explicitly considered here because it can be derived from speed and density; thus, estimating speed and density suffices to obtain flow. Furthermore, since speed and density have different units, a weight, w_v, w_ρ, is introduced, the value of which is determined through simulation statistics to balance the different terms in the reward function.

Additionally, while a single agent estimates the traffic state within its jurisdiction, the results from two adjacent agents at their shared cell boundaries might violate flow conservation constraints. Therefore, a physics-based reward function is defined for pairs of agents corresponding to adjacent boundary cells, representing the negative of the degree of physical constraint violation, as shown in Equation (5):

$$r_{t,i,i+1}^{physic} = -\left|q_{j,t,i} - q_{j,t,i+1}\right| \tag{5}$$

The total constraint violation for all agents is given by Equation (5):

$$r_t^{physic} = \frac{1}{|\Omega|}\sum_{(i,i+1)\in\Omega} r_{t,i,i+1}^{physic} \tag{6}$$

The total reward function is the weighted sum of these two components, as defined in Equation (6):

$$r_t = \lambda^1 \cdot r_t^{estimate} + \lambda^2 \cdot r_t^{physic} \tag{7}$$

For simplicity, we set the trade-off weights between the estimation error term and the constraint violation term to 1, i.e., $\lambda^1 = 1$ and $\lambda^2 = 1$ in Equation (7).

4.3. Agent Design

In this study, the agent is designed to implement a complete mapping process from observations to actions, aiming to effectively fuse multi-source traffic data and generate fundamental diagram model parameters that conform to traffic flow principles while ensuring physical consistency. Specifically, an agent located at a roadside detector receives observations from two distinct sources: one is the local static observation provided by the fixed detector, including features such as average speed, flow, and density in the current time step; the other is the trajectory information uploaded by Connected and Automated Vehicles (CAVs), which constitutes a continuous time series capable of characterizing the global dynamic features of vehicles during spatiotemporal evolution.

To extract the core representations of these two types of data separately, the former is transformed into a low-dimensional local embedding vector via linear mapping, while the latter is processed by a Long Short-Term Memory (LSTM) network to model temporal dependencies, yielding a global feature vector. Subsequently, these two types of features are fed simultaneously into an interactive attention mechanism encoder to obtain a unified environmental representation. This representation captures the complementary relationship between local and global observations, enabling the agent to fully leverage the global temporal information from CAVs even when local observations are limited. The environmental representation is then input into a multi-layer residual network for non-linear transformation, outputting a vector of fundamental diagram parameters. Finally, this vector is compressed within a physically constrained range via a linear mapping function and passed as the agent's action in the current time step to the LWR solver.

Formally, let the roadside detector observation be denoted as a vector, $o_{i,t} \in \mathbb{R}^3$, and the trajectory sequence uploaded by CAVs within the time window $[t, t + \Delta t]$ be denoted as $D^{CAV}_{i,t:t+\Delta t} \in \mathbb{R}^{K \times 2 \times \Delta t}$, where K represents the number of CAVs within this time window. After processing by linear mapping and LSTM extraction, we obtain the local and global feature representations, respectively, as shown in Equations (8)–(10):

$$h^{loc}_{i,t} = \mathrm{W}_{loc} o_{i,t} + b_{loc} \tag{8}$$

$$h^{glob}_{i,t,k} = \mathrm{LSTM}\left(D^{k}_{i,t:t+\Delta t}\right),\ for\ \forall k \tag{9}$$

$$h^{glob}_{i,t} = \mathrm{avePooling}\left(\left\{h^{glob}_{i,t,k}\right\}_k\right) \tag{10}$$

Special attention is required for the processing of CAV embeddings. During the traffic state estimation period, CAVs may exit the estimated road section, or the total number may change. Therefore, this study uses an LSTM to embed the trajectory data of each CAV, followed by average pooling as in Equation (10) to obtain a global representation. This design ensures permutation invariance and robustness to time-varying CAV counts while maintaining a fixed-dimensional representation. Although average pooling may suppress higher-order distributional information across vehicles, its impact is alleviated by the prior LSTM-based temporal encoding of each CAV and the subsequent interactive attention mechanism, where the aggregated CAV feature serves as complementary global context rather than a dominant decision input.

Here, $h_{i,t}^{loc} \in \mathbb{R}^{k1}$ and $h_{i,t}^{\text{glob}} \in \mathbb{R}^{k1}$ represent the local and global feature representations, respectively. These two features are input into the interactive attention mechanism to obtain the environmental representation. The main computational flow of the interactive attention mechanism encoder is illustrated in Figure 2.

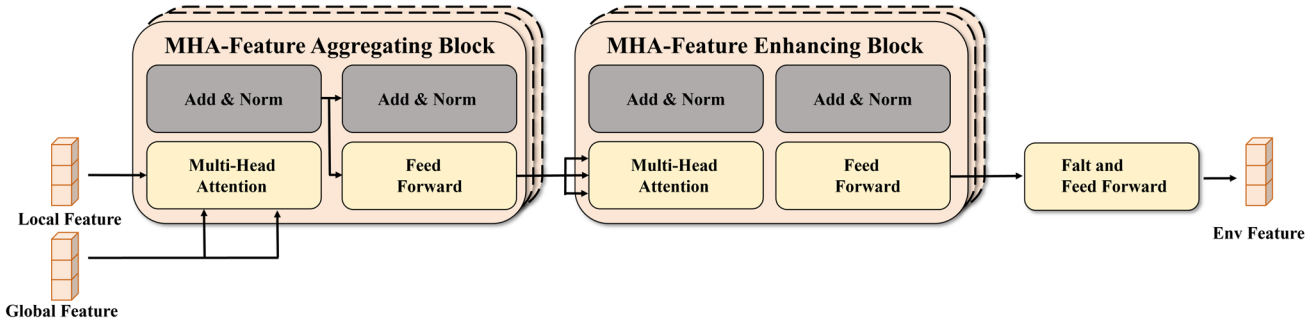

Figure 2. Description of the interactive attention mechanism-based encoder.

The workflow of the interactive attention mechanism is to enhance the connection between local and global observations by computing an attention matrix. Specifically, the local observation features are primary and are input into the attention mechanism as the Key (K) and Value (V), while the global observation features are input as the Query (Q). This design is based on the rationale that estimating the local state is the agent's most critical task, and global state observations play a supplementary role. The computational steps for this process are shown in Equations (11)–(13):

$$V_i = h_{i,t}^{loc} \cdot \mathrm{W}_V \tag{11}$$

$$K_i = h_{i,t}^{loc} \cdot \mathrm{W}_K \tag{12}$$

$$Q = h_{i,t}^{glob} \cdot \mathrm{W}_Q \tag{13}$$

Subsequently, the dot product between the Query (Q) and the Key (K) is first computed to quantify the correlation between them. Then, the Softmax function is applied to normalize the dot product results, yielding the attention weights (i.e., the degree of attention each Key pays to the Query). These weights determine the influence of each feature in the subsequent computation. Following this, the weighted Values (V) are aggregated to form a new feature representation. The core implication of this process is that by dynamically adjusting the weights, the model can focus on the features most critical to the current estimation task, thereby modeling the traffic state dependencies more accurately. The main calculation process is shown in Equation (14):

$$Att(Q_{h,l}, K_{h,l}, V_{h,l}) = \text{softmax}\left(\frac{Q_{h,l}K_{h,l}^{\mathrm{T}}}{\sqrt{C}}\right)V_{h,l} \tag{14}$$

Finally, the results from multiple computation heads are concatenated and passed through a linear mapping to obtain the fused output of this interactive attention layer, as shown in Equation (15):

$$MHAtt = \left[Att(Q_{1,l}, K_{1,l}, V_{1,l}), \ldots, Att\left(Q_{|h|,l}, K_{|h|,l}, V_{|h|,l}\right)\right] W_l^{MHATT} \tag{15}$$

The feature obtained in this step serves as the input for the Key and Value of the next multi-head interactive attention layer, until the computation of the final interactive attention layer is completed.

Subsequently, to further enhance the representation, this study employs a self-attention mechanism to strengthen the model's global perception capability. The main difference between the self-attention mechanism and the interactive attention mechanism is that it does not rely on externally input interactive features when computing attention but instead utilizes the intrinsic relationships among the elements within the sequence. Each input feature is compared with other features in the sequence, forming a set of adaptive attention weights that determine the contribution of each feature to the output. This process is similar to Equations (10)–(14), except that the Query, Key, and Value computations in each step all input the same features, yielding the encoder's environmental representation output $h_{i,t}^{Enc}$.

The final environmental representation is used to generate the action vector through a residual network:

$$\overline{v}_{i,t}^{f}, \overline{\rho}_{i,t}^{m} = \mathrm{Tanh}\left(f^{\mathrm{ResNet}}\left(h_{i,t}^{Enc}\right)\right) \tag{16}$$

The detailed computation process of the residual network f^{ResNet} can be referred to in relevant papers [29]. The output of the residual network is mapped to the range $[-1, 1]$ using the tanh activation function. The action is then linearly mapped to obtain the actual values of the free-flow speed parameter and the jam density parameter, as shown in Equations (17) and (18):

$$v_{i,t}^{f} = \mathrm{LB}^{v} + \overline{v}_{i,t}^{f} \cdot (\mathrm{UB}^{v} - \mathrm{LB}^{v}) \tag{17}$$

$$\rho_{i,t}^{m} = \mathrm{LB}^{\rho} + \overline{\rho}_{i,t}^{m} \cdot (\mathrm{UB}^{\rho} - \mathrm{LB}^{\rho}) \tag{18}$$

In this architecture, the interactive attention mechanism is introduced due to the multimodal nature of the observation data. Roadside detector data and CAV trajectory data differ significantly in temporal resolution, spatial coverage, and statistical characteristics; the former is static and sparse, while the latter is dynamic and continuous. Simply concatenating these two types of features makes it difficult for the model to effectively capture their complementarity. The interactive attention mechanism establishes adaptive weight allocation relationships between local and global features, highlighting the observation modality more critical under the current traffic state, thereby enhancing the expressiveness and robustness of the overall environmental representation.

On the other hand, the fundamental diagram parameters output by the model must satisfy strict physical constraints. For example, both free-flow speed and jam density should fall within reasonable numerical intervals. A common practice is to use a saturated activation function (like Tanh) to compress the network output into a finite interval, followed by linear mapping to obtain the physical quantity. However, in deep neural networks, non-saturated activation functions can easily cause gradient vanishing, leading to unstable model training. To address this, this study introduces a multi-layer residual network in the action generation stage. Its skip connections can directly propagate low-level features during forward propagation and alleviate gradient vanishing during backpropagation, thus ensuring the trainability and stability of the network. Combined with the final linear mapping function, the output both conforms to physical constraints and maintains the learning capability of the deep model.

4.4. LWR Solver and Parameter Diffuser

The traffic state estimation (TSE) problem inherently involves strict physical constraints, as traffic dynamics are governed by conservation laws. Purely data-driven methods often fail to capture these constraints and may produce physically inconsistent results. Moreover, directly allowing each agent to output the traffic states of all cells within its control region would lead to an extremely high-dimensional action space, thereby hindering reinforcement learning (RL) convergence and stability.

To address these issues, this study integrates a physics-based LWR solver into the multi-agent reinforcement learning (MARL) framework, accompanied by a parameter diffuser to ensure spatial continuity of the macroscopic model parameters. Together, these two components form the physical computation core that bridges agent decisions with spatiotemporal traffic evolution. The LWR solver guarantees physical consistency through conservation laws, while the diffuser ensures smooth parameter transitions between adjacent regions.

4.4.1. Parameter Diffuser

In macroscopic traffic flow theory, the LWR model assumes that the evolution of the traffic state at each spatial position is determined by two fundamental diagram (FD) parameters: the free-flow speed u_f and the maximum density $\rho_{\max}$. In the proposed MARL framework, each agent is located at a fixed detector and outputs the FD parameters corresponding to its local control region. However, directly assigning these discrete outputs to all cells governed by the agent would introduce abrupt parameter discontinuities at the boundaries between adjacent agents, which may distort the propagation of traffic waves.

To maintain spatial smoothness and physical realism, an Inverse Distance Weighting (IDW) interpolation method is employed to diffuse the FD parameters across the entire highway segment. IDW is adopted here because it is lightweight and parameter-free, producing a continuous detector-anchored parameter field with negligible per-step overhead, which fits the real-time MARL–LWR inference loop. Let the FD parameters generated by agent i be denoted as $\theta_i = \left[u_{f,i}, \rho_{\max,i}\right]$ and those of cell j as θ_j. The spatial interpolation is then defined as

$$\theta_j = \frac{\sum_i w_{i,j}\theta_i}{\sum_i w_{i,j}}, \qquad w_{i,j} = \frac{1}{d_{i,j}^2} \tag{19}$$

where $d_{i,j}^2$ represents the Euclidean distance between cell j and agent i. This weighted formulation ensures that cells closer to a detector inherit stronger influence from its output, resulting in a continuous and physically meaningful spatial distribution of FD parameters. Because IDW relies purely on geometric proximity, its effective smoothing strength depends on detector spacing and distribution; sparse or uneven detector layouts may over-smooth the diffused parameter field and attenuate abrupt local heterogeneity.

Consequently, the parameter diffuser converts discrete agent actions into a smooth field of u_f and $\rho_{\max}$ across space, which provides a coherent foundation for the subsequent LWR-based reconstruction and temporal propagation of traffic states. This diffusion step further prevents numerical instability that could arise from abrupt parameter changes between adjacent agents.

4.4.2. Spatial Reconstruction in the LWR Numerical Solver

In each time step t_k, it is necessary to reconstruct the instantaneous spatial distribution of traffic states—density, speed, and flow—throughout the highway based on detector observations and the diffused FD parameters. To achieve this, a spatial reconstruction module is incorporated into the LWR framework using the Godunov numerical flux scheme, which effectively captures nonlinear wave propagation and guarantees conservation.

The continuous form of the LWR model is expressed as

$$\frac{\partial \rho(x,t)}{\partial t} + \frac{\partial q(\rho(x,t))}{\partial x} = 0 \tag{20}$$

where the flow–density relationship follows the Greenshields fundamental diagram [30]:

$$q(\rho) = u_f \cdot \rho\left(1 - \frac{\rho}{\rho_{\max}}\right) \tag{21}$$

We adopt the Greenshields fundamental diagram primarily for its parsimony and computational convenience: it yields a smooth unimodal flux with only two interpretable parameters, which aligns with our agent action design and enables a stable Godunov-based LWR solver within the MARL loop. Although the linear speed–density assumption is a simplification, it is widely used as a representative FD in traffic state estimation and physics-informed learning settings, especially when the goal is robust reconstruction under sparse observations. Nevertheless, this assumption may become less accurate under strongly congested, non-equilibrium conditions, e.g., those pronounced hysteresis/capacity drop or abrupt stop-and-go dynamics, where more flexible multi-regime diagrams could be preferable.

To provide a more intuitive understanding of the Greenshields model, a diagram of the model under the parameters of maximum density and free-flow speed is shown in Figure 3.

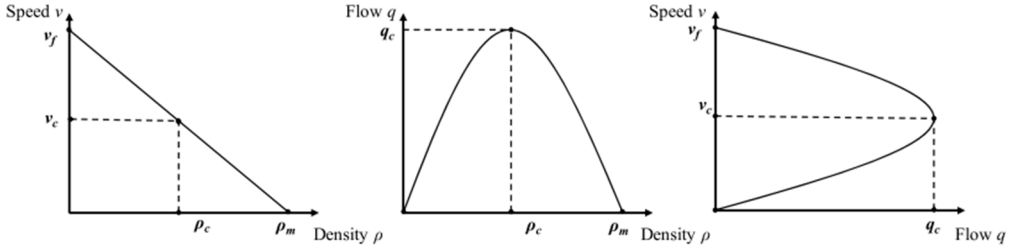

Figure 3. Fundamental diagram model based on Greenshields.

The maximum flow $q_{\max} = u_f \frac{\rho_{\max}}{4}$ occurs at the critical density $\rho_c = \frac{\rho_{\max}}{2}$ [31].

Discretizing the spatial domain $[x_1, x_2]$ into NNN cells of length Δx, the conservation equation becomes

$$\rho_i^{t+\Delta t} = \rho_i^t - \frac{\Delta t}{\Delta x}\left(F_{i+\frac{1}{2}}^t - F_{i-\frac{1}{2}}^t\right) \tag{22}$$

where $F_{i+\frac{1}{2}}^t$ denotes the numerical flux between cells i and $i+1$.

The Godunov flux solves a local Riemann problem between adjacent cells to determine the physically correct propagation direction of traffic waves. It is defined as

$$F_{i+\frac{1}{2}} = \begin{cases} \min_{\rho\in[\rho_i,\rho_{i+1}]} q(\rho), & \text{if } \rho_i \le \rho_{i+1} \\ \max_{\rho\in[\rho_{i+1},\rho_i]} q(\rho), & \text{if } \rho_i > \rho_{i+1} \end{cases} \tag{23}$$

Since $q(\rho)$ is unimodal with respect to ρ, the flux can be further simplified as

$$F_{i+\frac{1}{2}} = \begin{cases} q(\rho_i), & \text{if } \rho_i, \rho_{i+1} \le \rho_c \\ q(\rho_{i+1}), & \text{if } \rho_i, \rho_{i+1} \ge \rho_c \\ q(\rho_c), & \text{if } \rho_i < \rho_c < \rho_{i+1} \end{cases} \tag{24}$$

This scheme inherently preserves the correct wave propagation direction and automatically captures both shockwaves and rarefaction waves. Given the upstream and downstream boundary conditions provided by the detectors,

$$\rho_1^{t_k} = \rho_{up}^{t_k}, \quad \mathrm{q}_1^{t_k} = q_{up}^{t_k}, \quad \rho_N^{t_k} = \rho_{down}^{t_k}, \quad \mathrm{q}_N^{t_k} = q_{down}^{t_k} \tag{25}$$

the density field at time t_k can be reconstructed iteratively in space:

$$\rho_{i+1}^{t_k} = \rho_i^{t_k} - \frac{\Delta x}{u_{f,i} \cdot \rho_{mzx,i}} \left(F_{i+\frac{1}{2}}^{t_k} - F_{i-\frac{1}{2}}^{t_k} \right) \tag{26}$$

And the corresponding velocity and flow can be obtained from

$$v_i^{t_k} = u_{f,i} \cdot \left(1 - \frac{\rho_i^{t_k}}{\rho_{\max,i}} \right), \; \mathrm{q}_i^{t_k} = \rho_i^{t_k} v_i^{t_k} \tag{27}$$

Through this process, the solver reconstructs the complete traffic state field in the current time step, ensuring spatial continuity and physical plausibility even under non-stationary traffic conditions.

4.4.3. Temporal Inference in the LWR Solver

After reconstructing the spatial distribution at time t_k, the temporal evolution of the traffic state is simulated using the finite-difference method. Given the density and flux fields $\left(\rho_i^{t_k}, \mathrm{q}_i^{t_k} \right)$ obtained from spatial reconstruction, the LWR equation is discretized in time as

$$\rho_i^{t_{k+1}} = \rho_i^{t_k} - \frac{\Delta t}{\Delta x} \left(F_{i+\frac{1}{2}}^{t_k} - F_{i-\frac{1}{2}}^{t_k} \right) \tag{28}$$

The updated velocity and flow are then computed as

$$v_i^{t_{k+1}} = u_{f,i} \cdot \left(1 - \frac{\rho_i^{t_{k+1}}}{\rho_{\max,i}} \right), \quad \mathrm{q}_i^{t_{k+1}} = \rho_i^{t_{k+1}} v_i^{t_{k+1}} \tag{29}$$

4.5. Multi-Agent Reinforcement Learning Training Algorithm Based on MAPPO and LWR Solver

Within the proposed MARL–LWR framework, the agents aim to optimize the fundamental diagram parameters through interaction with the LWR solver so that the estimated traffic states match the ground truth generated by SUMO under long-horizon rolling prediction. We adopt a centralized training and decentralized execution (CTDE) paradigm, where each roadside detector is modeled as an agent and outputs continuous FD parameters. These parameters are propagated across the highway segment by the LWR solver to produce the next-step traffic state estimation.

Given the continuous action space and the need for stable learning under physical constraints, we employ Multi-Agent Proximal Policy Optimization (MAPPO). During training, a centralized critic is used to evaluate the global return using aggregated information from all agents (including local observations, global CAV features, and the LWR-provided traffic state summary from the previous step). During execution, each agent selects actions independently based on its own observations and the encoded environmental representation, enabling distributed deployment.

During each training episode, agents repeatedly interact with the LWR solver to generate rolling state estimates. The reward is computed by comparing the LWR-estimated states with the SUMO ground truth, including both estimation error and physics-consistency penalties. The policy is updated using the PPO clipping objective with advantages computed by GAE, and the value network is updated by minimizing the value regression loss. The overall procedure is summarized in Algorithm 1.

Algorithm 1. Multi-agent PPO for traffic state estimation

Initialize policy network π_θ and value network V_φ
for each training iteration do
  for each episode do
    Reset SUMO and LWR simulators
    for t = 1 to T do
      for each agent i do
        Obtain local $o_{i,t}$ and global $D^{CAV}_{i,t:t+\Delta t}$
        Encode features $\rightarrow h^{Enc}_{i,t}$
        Sample action $a_{i,t} \sim \pi_\theta\left(a_{i,t} \middle| h^{Enc}_{i,t}\right)$
      end for
      Apply actions to LWR solver $\rightarrow$ inference next step traffic state for all cell
      Compute reward r_t from estimation error and physics error
      Store observation, action and reward
    end for
  end for
  Compute advantage $\hat{A}_t$ using GAE
  Update θ by maximizing PPO-Clip objective
  Update φ by minimizing value loss
end for

5. Experiment

5.1. Dataset Description

This study adopts a strategy combining simulation and real-world data to train and validate the proposed multi-agent reinforcement learning framework. The use of simulated data for training is primarily motivated by the fact that traffic state estimation aims to reconstruct macroscopic spatiotemporal traffic evolution (e.g., the propagation of density and speed fields), rather than predicting individual vehicle trajectories. The faithful reproduction of macroscopic traffic flow phenomena (e.g., shockwave formation, congestion propagation) in simulation has been widely validated and accepted in research [1,2]. SUMO (Simulation of Urban Mobility), an open-source microscopic traffic simulator, can simulate the generation, propagation, and dissipation of congestion with high fidelity by specifying car-following (e.g., the IDM) and lane-changing models and is therefore well suited for generating large-scale, diverse datasets with known ground truth for training. In this study, SUMO is used to generate diverse training scenarios, and the trained model is subsequently evaluated on real-world data to assess cross-domain generalization.

The training dataset is designed to promote generalization rather than scenario-specific memorization. Because all agents share a single parameterized policy, randomly generating a large number of scenarios with diverse configurations encourages the policy to learn transferable traffic flow dynamics, rather than overfitting to specific detector layouts or road geometries. Each scenario instance is generated by randomly sampling key parameters, including the number and spatial layout of roadside detectors (agents), the CAV penetration rate, the highway-segment length, and the upstream inflow demand at the mainline entrance. To ensure that the dataset contains rich fluctuation patterns for learning spatiotemporal dynamics, we introduce stochastic disturbance events: at random times and locations, randomly selected vehicles are subjected to sustained forced deceleration. This operation effectively simulates disturbances triggered by driver behavior or incidents in reality, thereby exciting backward-propagating congestion waves and forming complex spatiotemporal congestion evolution patterns. Ultimately, this study randomly generated

12,600 unique training instances. In each training episode, 30 instances are sampled for parallel simulation and data collection, which increases batch diversity and improves training efficiency. To control computational overhead, the maximum simulation horizon for each instance is set to 3600 s.

To evaluate the trained model in real traffic environments, this study uses one publicly available real-world dataset for validation: the NGSIM US-101 dataset. The NGSIM dataset records high-precision vehicle trajectories on the US-101 highway. We selected two time periods (morning and afternoon) and extracted continuous trajectory segments to cover conditions under different demand levels. The dataset provides real vehicle trajectories, but the raw data inevitably contains noise. Therefore, a moving-average filter with a window length of 5 frames is applied to smooth the trajectories, providing a more stable basis for macroscopic-state computation and more reliable ground truth for evaluation.

5.2. Model Architecture and Training Hyperparameter Specification

To balance physical consistency and data-driven flexibility for highway traffic state estimation, we propose a hybrid framework that integrates multi-agent reinforcement learning with macroscopic traffic flow theory. This framework consists of a multi-agent policy network, an interactive attention mechanism, an LWR solver, and a reward function, aiming to achieve high precision, high real-time performance, and physically interpretable traffic state estimation through agent collaboration and the embedding of physical constraints.

The core architectural parameters of the model are summarized in Table 2. Each agent receives two types of inputs: one is the local static observation from roadside detectors, and the other is the global dynamic trajectory sequence from CAVs. The local observation is mapped to a 128-dimensional feature vector via a linear embedding layer, while the CAV trajectories have their temporal features extracted by a Long Short-Term Memory (LSTM) network, also with an output dimension of 128. To fuse these heterogeneous features, the model employs an interactive attention module in which the Keys and Values are derived from local features, whereas the Queries are derived from global features, enabling adaptive local–global fusion. The interactive attention module comprises two layers with eight heads per layer, and each head outputs a 64-dimensional representation. The fused features then pass through two layers of the self-attention mechanism to further strengthen the modeling of temporal dependencies, with a structure consistent with the interactive attention. Finally, the environmental representation undergoes non-linear transformation through a residual network containing three hidden layers, outputting an action vector, which is then constrained to a reasonable physical range via a Tanh activation function and linear mapping.

Table 2. Model structural parameter configuration.

Parameter Category	Specific Parameter	Value Range
Local Observation Embedding Dim	Linear Embedding Output Dimension	128
Global Trajectory LSTM Hidden Dim	LSTM Output Dimension	128
Interactive Attention Heads	Number of Attention Heads per Layer	8
Interactive Attention Output Dim	Output Dimension per Head	64
Interactive Attention Layers	Total Number of Layers	2
Self-Attention Heads	Number of Attention Heads per Layer	8
Self-Attention Output Dim	Output Dimension per Head	64
Self-Attention Layers	Total Number of Layers	2
Residual Network Hidden Layers	Number of Hidden Layers	3
Residual Network Hidden Dim	Number of Neurons per Layer	256
Activation Function	Action Output Activation Function	Tanh

In terms of training strategy, this paper adopts the Multi-Agent Proximal Policy Optimization (MAPPO) algorithm for centralized training and decentralized execution. Both the policy network and the value network take the aforementioned fused features as input, with the value network additionally receiving a global state summary to more accurately assess the state value. During training, the model optimizes the policy by minimizing the sum of the estimation error and the physical violation loss, and it uses Generalized Advantage Estimation (GAE) to compute the advantage function. The detailed hyperparameter configuration is shown in Table 3. We use the Adam optimizer with an initial learning rate of 3×10^{-4} and a linear decay schedule. The batch size is 256, and 4096 environment steps are collected per update. The discount factor is 0.99 and the GAE parameter is 0.95. Policy updates adopt PPO clipping with $\epsilon = 0.2$, while the value function is trained by minimizing the mean squared error. All experiments were conducted with a fixed random seed to ensure result reproducibility.

Table 3. Training and algorithm hyperparameter configuration.

Parameter Category	Specific Parameter	Value Range
Optimizer	Gradient Optimization Algorithm	Adam
Learning Rate	Parameter Update Step Size	3×10^{-4}
Batch Size	Samples per Update	256
Discount Factor	Future Reward Discount Rate	0.99
GAE Parameter	Advantage Estimation Coefficient	0.95
PPO Clip Range	Policy Update Clip Threshold	0.2
Value Function Update Epochs	Value Network Updates per Round	5
Gradient Clipping Threshold	Gradient Clipping Upper Limit	0.5
Random Seed	Experiment Reproducibility Seed	42

To systematically evaluate the generalization capability and robustness of the proposed framework, this study constructs a large-scale set of randomly generated traffic scenario instances during the training phase. As shown in Table 4, each instance is randomly configured across multiple dimensions to cover various structural and dynamic conditions that may exist in real highway environments. Specifically, each training instance is generated by randomly sampling a set of key parameters, including road geometry parameters, CAV penetration rate, main inflow traffic demand, and the attributes of randomly triggered traffic events. This design aims to force the agents to learn generalized traffic flow physical laws rather than overfitting to specific scenario configurations.

Unless otherwise specified, the simulation and the MARL–LWR interaction operate in a 1 s time step, while CAV trajectories are sampled at 10 Hz and aggregated within each 1 s window for global feature extraction.

It is particularly noteworthy that the introduction of random traffic events is the core mechanism for simulating real-world traffic disturbances and congestion propagation processes. At random time points during the simulation, at random spatial locations, randomly selected vehicles are subjected to mandatory deceleration for a certain duration, thereby exciting traffic waves and inducing the formation of complex spatiotemporal congestion evolution patterns. This mechanism significantly enhances the dynamic diversity and physical authenticity of the training samples, providing rich and challenging learning signals for the multi-agent policy to learn collaborative estimation in a partially observable environment.

Table 4. Training instance random generation parameter configuration.

Parameter Category	Specific Parameter	Value Range
Road Geometry	Highway Length	2.0–5.0 km
CAV Parameters	CAV Penetration Rate	2–25%
Demand Parameters	Main Inflow Base Density	15–45 vehicles/km
Random Event Parameters	Event Trigger Time	[300 s, 2700 s]
	Event Location	Random within segment
	Event Duration	60–180 s
	Mandatory Deceleration	−3.0–−1.5 m/s
Temporal Resolution	Agent–Environment Interaction Interval	1 s
	CAV Trajectory Sampling Rate	0.1 s
Simulation Setup	Total Instances	12,600
	Instances per Batch	30
	Max Simulation Time per Instance	3600 s

The design of this model architecture and training strategy enables the agents to effectively fuse multi-source traffic data in a partially observable environment, output fundamental diagram parameters that conform to physical laws, and subsequently achieve accurate, real-time estimation of the traffic state across the entire roadway segment through the LWR solver.

5.3. Evaluation Metrics and Baseline Models

To evaluate the proposed MARL–LWR framework for highway traffic state estimation, we assess performance in terms of estimation accuracy and generalization. Estimation accuracy is quantified using the widely adopted Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), computed against the ground truth state derived from high-fidelity traffic flow data provided by the SUMO microscopic simulator. Specifically, for the estimation results of all cells in the spatiotemporal domain, the RMSE and MAE values for average speed, density, and flow are calculated, respectively. Among them, the RMSE focuses on penalizing larger estimation errors and is more sensitive to outliers, whereas the MAE measures the average absolute deviation, better reflecting the overall bias level of the estimated values. These two metrics together form the core basis for evaluating the accuracy of the estimation results, and their mathematical expressions are as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{M}\sum_{j=1}^{M}\left(y_j - \hat{y}_j\right)^2} \tag{30}$$

$$\mathrm{MAE} = \frac{1}{M}\sum_{j=1}^{M}\left|y_j - \hat{y}_j\right| \tag{31}$$

In the selection of baseline models, this study aims to compare the proposed model with advanced methods under existing mainstream paradigms to highlight the superiority of the MARL–LWR framework. The selected baselines cover representative methods from both purely data-driven and physics–data dual-driven categories.

MME: The multimodel ensemble (MME) framework proposed by Li and Li [15] is a widely cited approach for freeway traffic state estimation. This method uses macroscopic traffic flow models, such as the Cell Transmission Model (CTM) and second-order models, as a candidate set and dynamically fuses the outputs of multiple models within a rolling time window through linear regression and regularization (LASSO, Ridge).

PCF: The Physics-Constrained Filter (PCF) estimation method proposed by Fountoulakis et al. [16] is a representative physics-embedded filtering framework. It fuses flow information from fixed detectors with speed data from Connected Vehicles (CVs), achieving dynamic fusion estimation of heterogeneous observation data through the continuity conservation equation. This model does not rely on explicit fundamental diagram calibration but directly uses vehicle conservation as the core constraint, avoiding complex parameter calibration.

PIDL: The physics-informed Bayesian deep learning (PIBDL) model proposed by Zhang et al. [32] combines LWR theory with Bayesian neural networks, providing uncertainty quantification while enforcing physical consistency.

These works represent successful applications of physics-guided deep learning in the TSE domain.

Furthermore, to examine the impact of decision granularity, we report MARL–LWR results under different cell lengths (10 m, 20 m, 50 m, and 100 m) as self-baselines. Finer cell division (e.g., 10 m) implies higher state space dimensionality and the potential for finer estimation but also poses greater challenges to the learning and coordination capabilities of the agents, whereas coarser division (e.g., 100 m) reduces complexity but may fail to effectively capture the fine propagation process of traffic waves. Through this comparison, the model's performance and trade-offs under different computational accuracy and complexity requirements can be assessed.

6. Results

6.1. Model Tuning

In this section, we compare the performance of the proposed model against two widely used fundamental diagram formulations: the triangular FD and the Underwood model. Figure 4 illustrates the convergence curves for each model, indicating that the Greenshields FD-based model achieves more stable convergence and lower estimation error than both alternatives. Specifically, the triangular FD model exhibits slower convergence, while the Underwood model struggles with error accumulation in long-term prediction, which could be attributed to its simplified traffic dynamics. Overall, the Greenshields FD provides a favorable trade-off between physical fidelity and computational efficiency and was therefore adopted in the subsequent experiments.

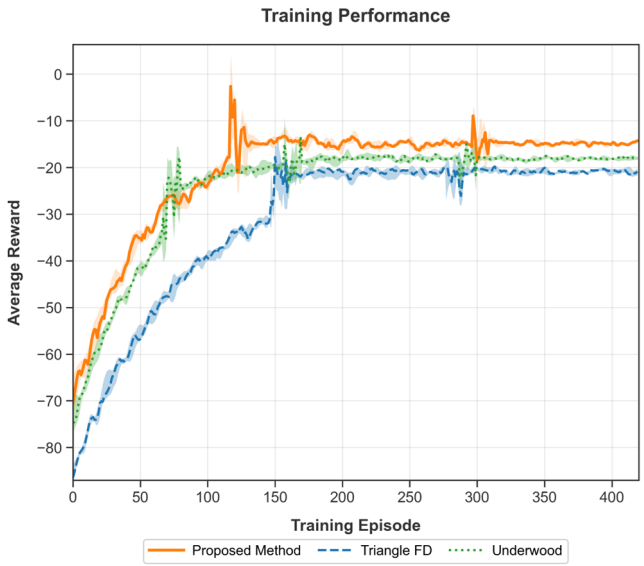

Figure 4. Reinforcement learning algorithm convergence curve.

In further comparative experiments, we investigated the impact of different cell discretization granularities on model performance. Specifically, the highway was partitioned into cells of different lengths (10 m, 20 m, 50 m, and 100 m), and we computed the velocity MAE of the validation set together with the LWR-solver inference time. The specific results are shown in Figure 5. The results indicate that a cell length of 50 m yields lower errors in speed, density, and flow than the other partitioning schemes. This phenomenon may occur because excessively fine discretization, while capturing more detailed fluctuations, increases the effective state dimensionality, thereby amplifying learning difficulty and noise sensitivity. Taken together, a cell length of 50 m provides a reasonable balance between computational complexity and estimation accuracy.

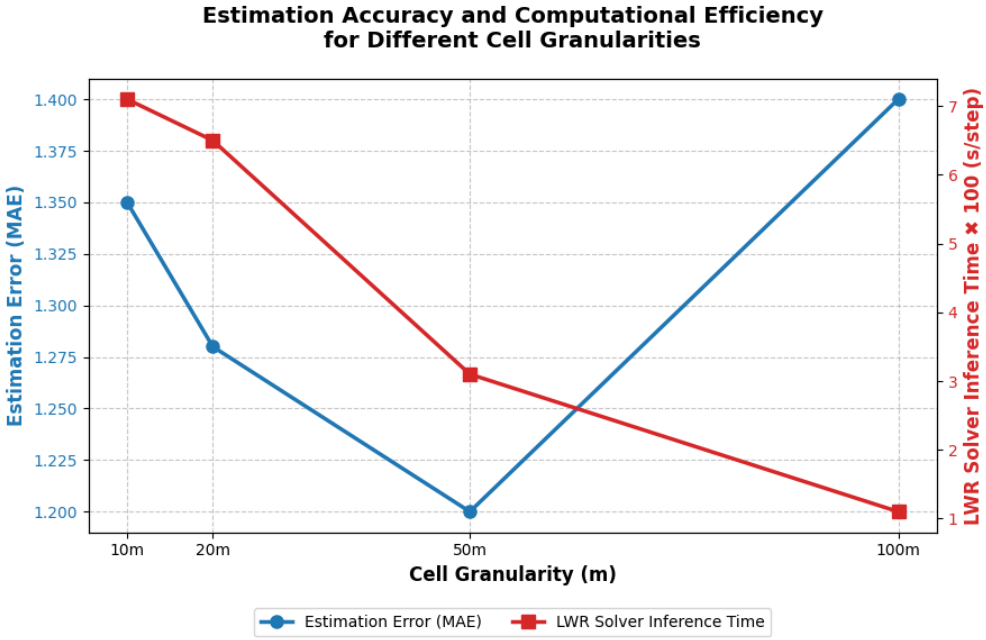

Figure 5. Accuracy and efficiency curves for different cell granularities.

Based on the above analysis, all subsequent experiments adopted a cell length of 50 m as the standard configuration. Performance comparisons and generalization tests were further conducted based on this setting.

To quantitatively validate the claimed real-time capability, this study benchmarked the runtime of the proposed MARL–LWR framework and representative baselines on the same hardware platform (NVIDIA RTX 3090, AMD Ryzen 9 5950X, and 128 GB of RAM). The results are summarized in Table 5. The inference step time denotes the computation required to produce a full traffic state estimate for the entire road segment given new observations.

Table 5. Inference latency and real-time capability benchmark.

Method	Stage	Performance Metric	Value
Proposed	Training	Total Time	18.5 h
	Inference	Step Time	15.2 ± 0.13 ms
MME	Inference	Step Time	45.3 ms
PCF	Inference	Step Time	21.8 ms
PIBDL	Training	Total Time	22.0 h
	Inference	Step Time	14.9 ms

In MDP formulation, the agent–environment interaction interval is 1 s, while CAV trajectories are sampled at 10 Hz and aggregated within each 1 s window. Therefore, the framework satisfies the real-time requirement as long as the end-to-end inference latency per step is well below 1 s.

As shown in Table 5, the proposed method achieves an inference latency of 15.2 ± 0.13 ms per step, which is faster than the 1 s operational interval, leaving an ample margin for practical overheads such as CAV data transmission, feature extraction, and multi-agent action generation.

In comparison, MME, PCF, and PIBDL require 45.3 ms, 21.8 ms, and 14.9 ms per step, respectively, indicating a higher computational cost under the same hardware conditions.

Because the conservation constraint is central to the proposed MARL–LWR framework, we further provide a quantitative assessment of mass-conservation violation at the interfaces between adjacent detector-controlled regions. Specifically, at each internal boundary, we compute the absolute mismatch between the boundary flow reconstructed from the upstream side and that from the downstream side in each decision step, and we report the averaged mismatch over all internal boundaries and time steps as the flow-conservation violation metric. This metric is consistent with the physics-based penalty term used in the reward design, where the total reward balances the estimation error component and the conservation violation component via the trade-off coefficients λ^1 and λ^2.

To justify the choice of the fixed weights and to demonstrate robustness, this study conducted a short sensitivity study by fixing $\lambda^1 = 1.0$ and varying $\lambda^2 \in \{0, 0.2, 0.5, 1.0, 2.0, 5.0\}$, while reporting both estimation accuracy (RMSE and MAE of speed and density) and the quantitative conservation violation metric. As summarized in Table 6, removing the physics term, e.g., $\lambda^2 = 0$, yields a slightly lower estimation error, e.g., the speed RMSE = 1.18, but leads to a pronounced increase in conservation violation, indicating that the policy can reduce numerical error while producing boundary-inconsistent states. In contrast, increasing λ^2 monotonically reduces the conservation violation, but overly large λ^2 noticeably degrades estimation accuracy. Importantly, in a moderate range around the default setting, the estimation errors change only marginally, e.g., the speed RMSE from 1.21 to 1.28 and density RMSE from 4.13 to 4.34, while conservation violation is substantially improved, suggesting that the performance is not overly sensitive to small perturbations of λ^2. Therefore, $\lambda^1 = \lambda^2 = 1.0$ was adopted as a principled and balanced configuration that achieves low conservation violation without sacrificing estimation accuracy, and it was used in all subsequent experiments.

Table 6. Quantitative mass-conservation violation and sensitivity analysis of reward weights.

λ^1	λ^2	λ^1/λ^2	Velocity (m/s)		Density (veh/m)		Constraint Violation of Flow (veh/s)
			RMSE	MAE	RMSE	MAE	
1.0	0	/	1.18	1.55	4.05	4.75	5.22
1.0	0.2	5	1.19	1.57	4.10	4.79	3.84
1.0	0.5	2	1.21	1.60	4.13	4.82	2.15
1.0	1.0	1.0	1.23	1.63	4.18	4.88	1.12
1.0	2.0	0.5	1.28	1.70	4.34	5.02	1.05
1.0	5.0	0.2	1.45	1.95	4.75	5.56	0.81

6.2. Model Performance Under Fixed Highway Length

Under the experimental setting of a fixed highway length, we selected a fixed observation period and lane segment from the NGSIM-US101 dataset. Vehicles were randomly designated as Connected and Automated Vehicles (CAVs), and experiments were conducted under CAV penetration rates of 5%, 10%, 15%, 20%, and 25%. Each experiment was repeated 10 times, and the average results were taken to mitigate the effects of randomness. Table 7 summarizes the performance comparison results between the proposed method and the three mainstream baseline models.

Table 7. Proposed model and baseline model performance under fixed highway length.

Metrics RMSE/MAE	Proposed Model		MME		PCF		PIBDL	
	Velocity (m/s)	Density (veh/m)	Velocity (m/s)	Density (veh/m)	Velocity (m/s)	Density (veh/m)	Velocity (m/s)	Density (veh/m)
5%	1.8/2.5	5.4/6.8	2.8/3.6	9.2/11.5	2.3/2.7	6.8/8.4	1.9/2.4	6.0/7.5
10%	1.5/2.2	4.8/5.8	2.4/3.1	7.8/9.6	1.8/2.3	5.9/7.3	1.6/2.1	5.2/6.5
15%	1.3/1.9	4.5/4.9	2.1/2.8	6.9/8.5	1.6/2.1	5.2/6.4	1.4/1.9	4.6/5.7
20%	1.2/1.6	4.0/4.8	1.9/2.5	6.2/7.6	1.4/1.9	4.7/5.8	1.3/1.7	4.1/5.1
25%	1.0/1.4	3.6/4.5	1.7/2.3	5.6/6.9	1.3/1.8	4.3/5.3	1.1/1.6	3.8/4.7

Note: For presentation clarity, the reported density RMSE and MAE values are multiplied by 100, while the underlying computations are based on the original physical unit (veh/m).

The results indicate that, across all CAV penetration rates, the proposed method achieves lower RMSEs and MAEs for speed, density, and flow compared with the baseline models. For instance, at a CAV penetration rate of 20%, the RMSE/MAE for speed estimation are 1.2/1.6 for the proposed method, compared with 1.9/2.5 (MME), 1.4/1.9 (PCF), and 1.3/1.7 (PIBDL). Similar advantages are observed for density and flow metrics. It is particularly noteworthy that as the CAV penetration rate increases, our method consistently maintains lower errors and demonstrates a more pronounced performance improvement trend. This suggests that the proposed multi-agent framework can effectively fuse heterogeneous observations from fixed detectors and CAVs, thereby capturing traffic fluctuations under limited observability.

To further assess whether the performance gains reported in Table 7 are statistically significant, Figure 6 presents violin plots of the estimation errors together with paired t-test results under different CAV penetration rates. The proposed method achieves statistically significant improvements over the MME and PCF across all penetration levels and all metrics (mostly $p \ll 0.001$). Compared with the strongest baseline, PIBDL, the proposed method remains significantly better in most cases. Only a few configurations exhibit no significant difference, such as the velocity MAE at 15% penetration and density RMSE at 20% penetration, indicating comparable performance under these specific settings. Overall, the statistical analysis confirms that the observed accuracy improvements are robust and not attributable to random fluctuations.

More importantly, compared with supervised or deep learning-based baselines, the proposed approach mitigates error accumulation in long-horizon rolling prediction. This advantage stems from reinforcement learning explicitly optimizing long-horizon spatiotemporal dynamics, rather than only short-term prediction accuracy. Therefore, under the fixed-highway-length scenario, the proposed method not only surpasses the compared models in accuracy but also demonstrates stronger stability and robustness in long-term prediction and under complex traffic fluctuation patterns.

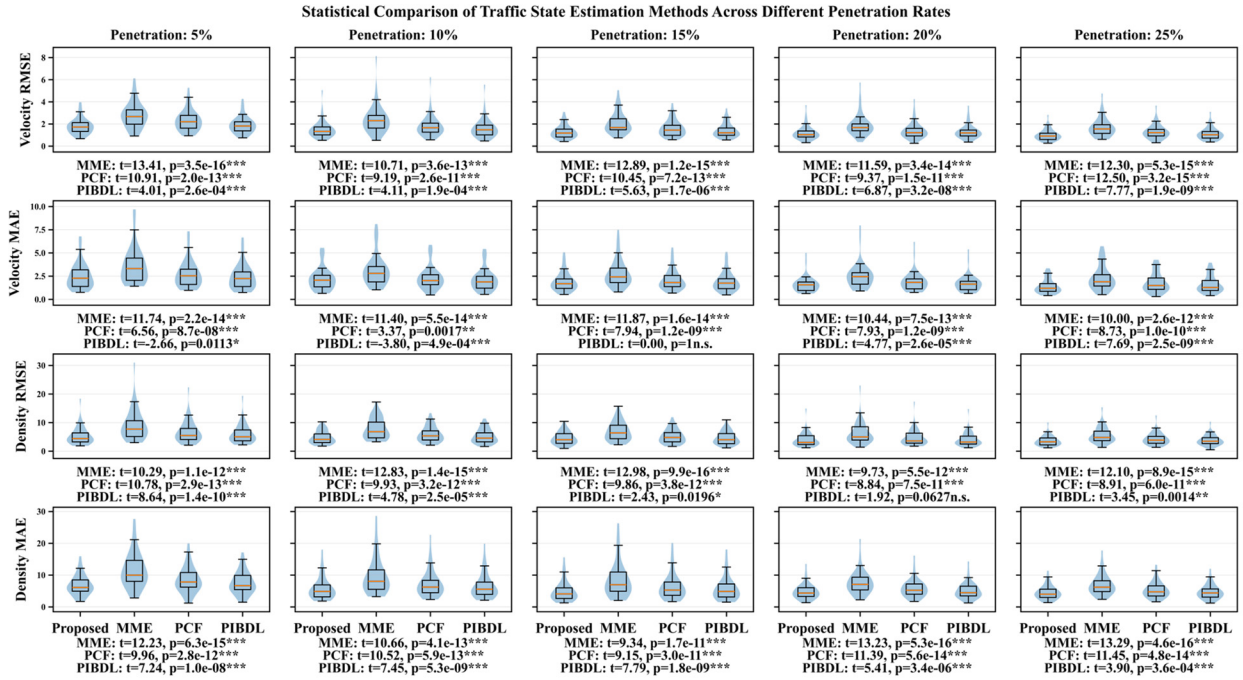

Figure 6. Statistical comparison of traffic state estimation errors under different CAV penetration rates using violin plots and paired *t*-tests. Statistical significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, n.s. = not significant.

6.3. Model Performance Under Variable Highway Length

Beyond temporal rolling prediction, the ability of traffic state estimation methods to generalize across spatial scales is equally crucial. To this end, we further investigated the model's performance under different highway lengths. In the experiments, fixed lane data from the NGSIM-US101 dataset were selected and tested at four different lengths: 2000 m, 3000 m, 4000 m, and 5000 m. For fairness, all baseline methods were retrained for each length, whereas the proposed method required no retraining and was adapted by setting the agents' average perception range to approximately 1000 m (covering the segment boundaries). Table 8 presents the comparative results of each model under different segment lengths.

Table 8. Proposed model and baseline model performance under variable highway length.

Metrics RMSE/MAE	Proposed Model		MME		PCF		PIBDL	
	Velocity (m/s)	Density (veh/m)	Velocity (m/s)	Density (veh/m)	Velocity (m/s)	Density (veh/m)	Velocity (m/s)	Density (veh/m)
2000 m	1.2/1.8	4.4/5.7	2.2/2.9	7.2/8.8	1.7/2.2	5.5/6.7	1.5/2.0	4.9/6.0
3000 m	1.3/1.6	4.7/5.8	2.4/3.1	7.8/9.6	1.8/2.3	5.9/7.3	1.7/2.1	5.2/6.5
4000 m	1.4/1.8	4.8/6.0	2.5/3.3	8.0/9.8	1.9/2.5	6.1/7.5	1.8/2.2	5.2/6.6
5000 m	1.6/2.5	5.2/6.3	2.7/3.5	8.4/9.9	2.2/2.8	6.4/7.8	2.1/2.4	5.7/6.9

Note: For presentation clarity, the reported density RMSE and MAE values are multiplied by 100, while the underlying computations are based on the original physical unit (veh/m).

The results show that estimation errors increase with segment length for all models; however, the proposed method exhibits the smallest degradation across speed, density, and flow. For example, for the 5000 m segment, the RMSE/MAE for speed of our method are 1.2/2.8, while the errors for the MME, PCF, and PIBDL reach 2.7/3.5, 2.2/2.8, and 1.8/2.4, respectively. A similar trend is observed in the estimation of density and flow. This

indicates that the proposed method can better adapt to changes in spatial scale, possessing strong generalization capability.

Figure 7 further reports the statistical comparison results under varying highway lengths. Despite the increasing estimation difficulty as the segment length grows, the proposed method consistently outperforms the baseline models with strong statistical significance across most metrics and lengths. In particular, the improvements over the MME and PCF remain highly significant ($p \ll 0.001$), while comparisons with PIBDL are also significant in the majority of cases, with only marginal reductions in significance for the velocity MAE under the longest segment. These results demonstrate that the proposed MARL–LWR framework preserves its performance advantage across spatial scales, providing strong evidence of its generalization capability without retraining.

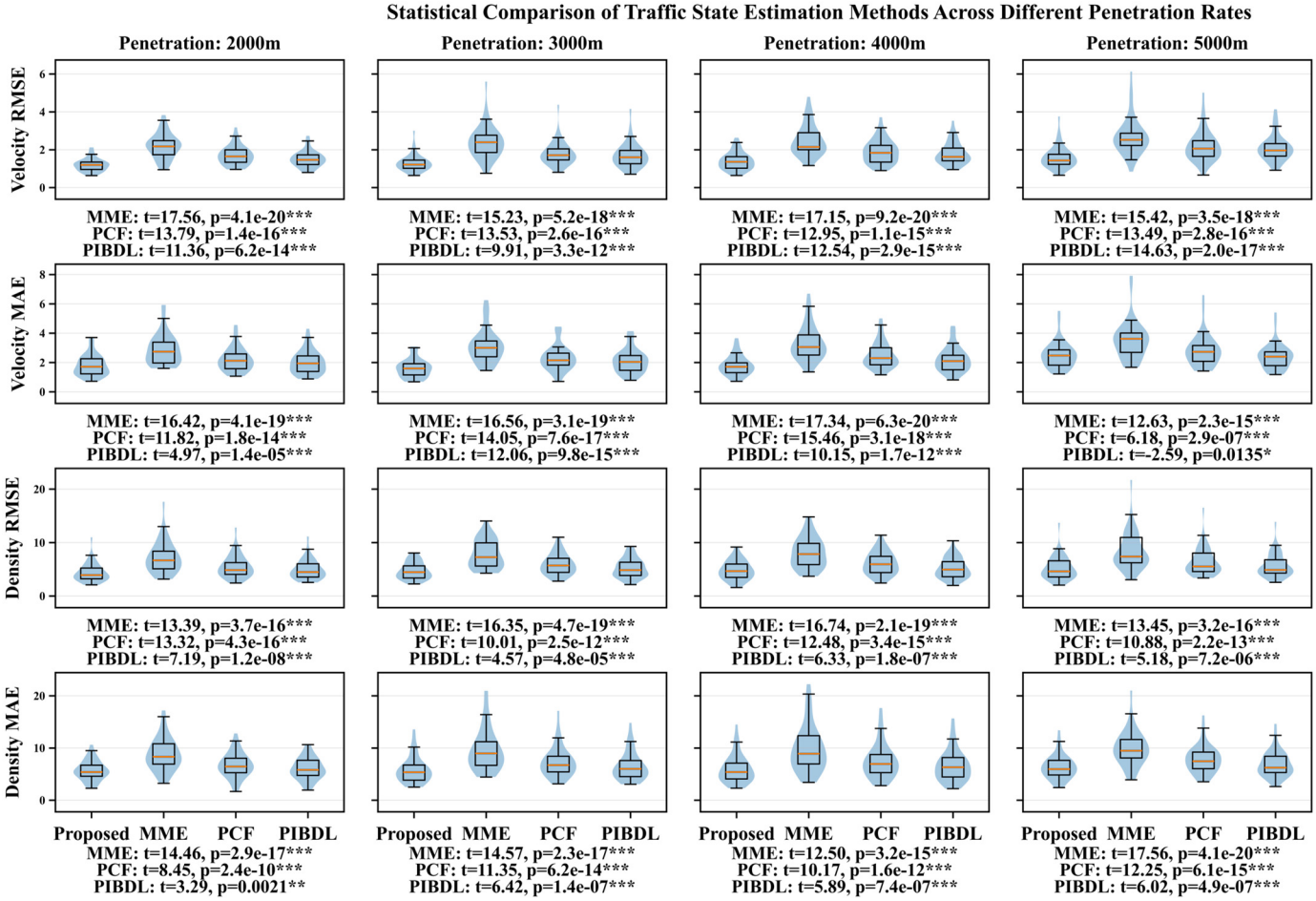

Figure 7. Statistical comparison of traffic state estimation errors under different highway lengths using violin plots and paired *t*-tests. Statistical significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

This advantage is attributed to the introduction of the multi-agent and cell partitioning strategy, which enables the model to flexibly adjust its perception and estimation range across different spatial scales, thereby maintaining stable performance. More importantly, the proposed method offers practical advantages for deployment: it requires no retraining across road lengths and achieves satisfactory estimation given appropriate detector placement. This not only enhances the model's flexibility in practical deployment but also reduces the construction and maintenance costs of traffic monitoring systems, holding great significance for the widespread adoption of large-scale intelligent transportation systems.

6.4. Model Performance Under Different Traffic Fluctuations

In the preceding sections, we have systematically validated the generalization capability and stability of the proposed multi-agent reinforcement learning framework (hereinafter referred to as MARL–LWR) under varying road lengths and vehicle penetration rates. However, the operational characteristics of real-world traffic systems are often non-stationary. They are jointly influenced by external disturbances and heterogeneous driving behaviors, exhibiting complex nonlinear fluctuation patterns. These fluctuations can be amplified through vehicle interactions, forming propagating patterns with spatiotemporal regularity, such as shockwaves or queue waves. Accurately characterizing such phenomena is crucial for assessing physical consistency and for evaluating whether a model captures the governing spatiotemporal evolution of traffic flow. This section aims to verify whether the proposed MARL–LWR framework can correctly learn and reproduce the spatiotemporal evolution mechanisms of traffic waves under complex dynamic conditions.

To this end, two typical and representative traffic fluctuation patterns were selected as experimental subjects:

(1) Shockwave-type fluctuations induced by local forced deceleration;
(2) Queue-type fluctuations caused by bottleneck constraints.

To avoid interference from overlapping multiple disturbances and ensure the identifiability and controllability of the two fluctuation patterns, the experiments were conducted in an idealized highway simulation environment constructed based on the two-dimensional Intelligent Driver Model (2D-IDM). The trained MARL–LWR model was deployed for state estimation. By comparing the true states from the simulation with the estimated results, the model's spatiotemporal response characteristics and estimation accuracy under different wave mechanisms are analyzed.

6.4.1. Shockwave-Type Fluctuation

This pattern simulates traffic shockwaves triggered by local driving disturbances. The experiment constructs a 10 km long open highway where vehicles can enter and exit freely. Within the section from 5500 m to 6000 m, a certain number of vehicles are randomly selected with a given probability to undergo forced deceleration for a short period before resuming normal driving. This setup is equivalent to scenarios where drivers suddenly decelerate on the mainline, effectively inducing the upstream propagation of speed disturbances.

From a traffic dynamics perspective, such disturbances typically manifest as a sudden drop in speed downstream, forming a congestion nucleus, which subsequently propagates upstream, forming a typical shockwave structure. Simulation results show that irregular deceleration bands emerge within the disturbance zone, and the impact of speed decay gradually diffuses upstream, forming distinct low-speed strips.

Figure 8 shows a comparison between the true traffic state and the model's estimated results under this scenario. It can be observed that the proposed MARL–LWR model can accurately reproduce the entire process of shockwave formation and propagation. Through long-term interaction with the LWR solver, the agents learn the dynamic adjustment patterns of free-flow speed and jam density in the region where fluctuations occur, resulting in estimated results that are highly consistent with the true state in terms of spatiotemporal morphology. Minor temporal offsets are observed at the boundaries of the low-speed bands, corresponding to a slightly underestimated shockwave propagation speed relative to the ground truth. This discrepancy likely arises from time-step mismatch and state synchronization errors between the LWR solver and SUMO, resulting in a small systematic bias. Overall, the model successfully captures the shockwave propagation rule under this

endogenously perturbed scenario, validating its ability to learn the nonlinear dynamics of traffic flow.

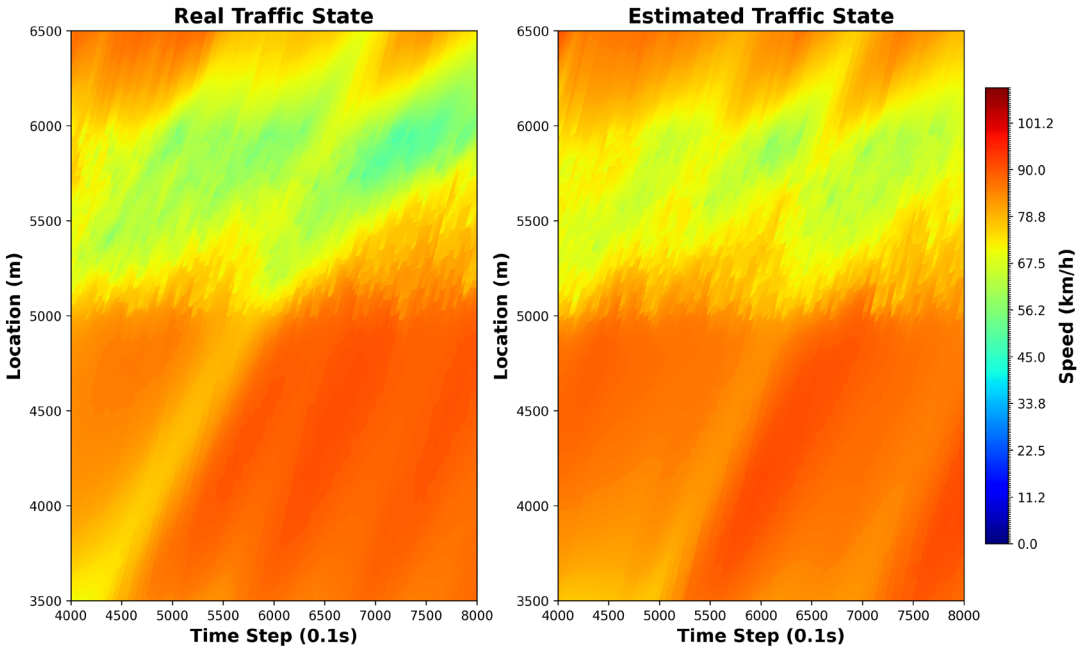

Figure 8. Real and estimated traffic states under shockwave-type fluctuations.

Figure 9 presents the estimated fundamental diagram parameters for the upstream and downstream agents following disturbance onset. We observe that as the disturbance develops, the downstream free-flow speed parameter gradually decreases, while the jam density parameter increases. Conversely, the upstream free-flow speed remains relatively stable but gradually decreases as congestion propagates upstream. The traffic density increases in both directions, with the downstream area seeing a more rapid increase in density due to the congestion formation.

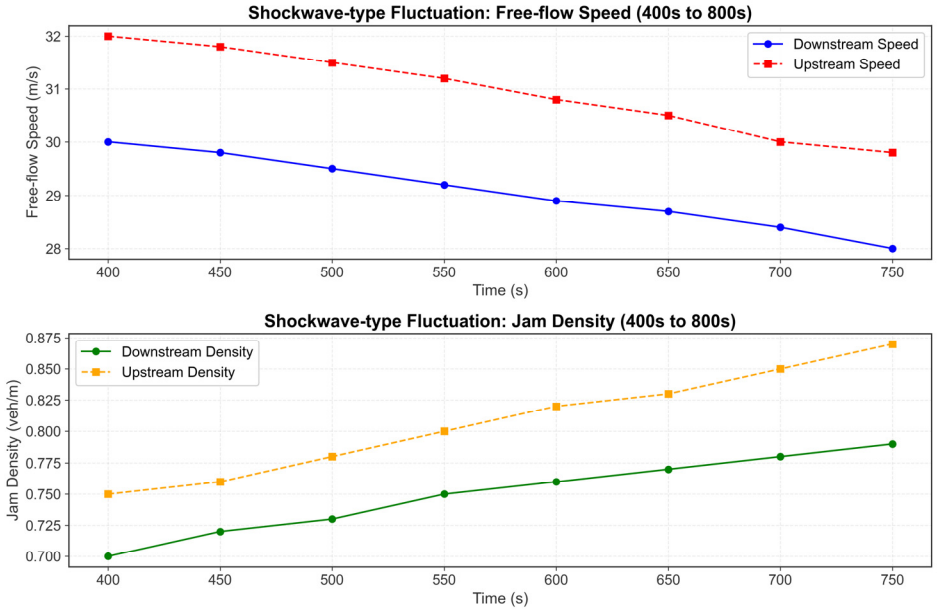

Figure 9. Time series of free-flow speed and jam density for downstream and upstream agents under shockwave-type fluctuations.

Figure 9 illustrates the time series of the free-flow speed and jam density for both the upstream and downstream agents. The results clearly demonstrate the effects of the Shockwave-type fluctuation as the disturbance propagates through the traffic system. Specifically, downstream speed decreases significantly after the formation of the bottleneck, which is consistent with the shockwave phenomenon in traffic dynamics. The downstream density increases steadily as the congestion builds up, reflecting the typical behavior observed when a shockwave is triggered in a traffic flow system.

In contrast, the upstream speed fluctuates within a range of 60 to 80 km/h, showing a gradual decline, while the upstream density increases more slowly over time as vehicles accumulate upstream. The downstream speed quickly recovers to a higher free-flow state, stabilizing around 100 km/h, as congestion starts to dissipate.

These results underscore the ability of the proposed framework to capture the dynamic evolution of traffic states and to reproduce key patterns of shockwave propagation and congestion dissipation in a physically consistent manner. Through continuous interaction with the environment, the agents learn to predict both the spatial and temporal evolution of traffic conditions, maintaining physical consistency in the process. This ability to model the shockwave-type fluctuation validates the effectiveness of the framework in real-time traffic state estimation, demonstrating its robustness in handling dynamic, non-equilibrium traffic flows.

6.4.2. Queue-Type Fluctuation

Unlike shockwaves, queue-type fluctuations are typically triggered by external constraints (such as road bottlenecks or speed limit controls). They manifest as the formation of a stable low-speed area downstream, with vehicles gradually accumulating upstream, producing persistent queuing. To simulate this pattern, a 10 km open highway scenario consistent with the previous one is constructed, and a speed-limited section is set between 7000 m and 7500 m, forcing all vehicles to pass through this area at a low speed (approximately 10–15 km/h). Vehicles resume free-flow conditions after passing this zone. This setup is equivalent to introducing a geometric bottleneck or external constraint in the downstream section of the road, which stably induces the formation and expansion of upstream queues.

Simulation results show that a long-lasting, stable low-speed zone forms downstream of the speed-limited section. Upstream vehicles accumulate due to restricted flow capacity, and the queue wave gradually extends upstream. Figure 10 shows the true state and the model's estimated results under this scenario. It can be observed that the MARL–LWR framework successfully reproduces the formation, expansion, and stable propagation processes of the queue wave and can satisfactorily predict the spatial extent and duration of the low-speed zone. Compared to the simulation results, the estimated low-speed band is slightly higher in value, manifesting as speeds slightly greater than the true state. This discrepancy arises because the LWR model, under an equilibrium flow assumption, may not fully capture the strongly constrained downstream state induced by the bottleneck. Moreover, during reinforcement learning, a relatively high weight on the physical constraint term encourages flow conservation, which can lead to slightly higher estimated speeds under extreme low-speed conditions. Despite this bias, the overall fluctuation structure and propagation characteristics remain consistent with the ground truth, validating the model's robustness and physical interpretability under exogenous disturbances.

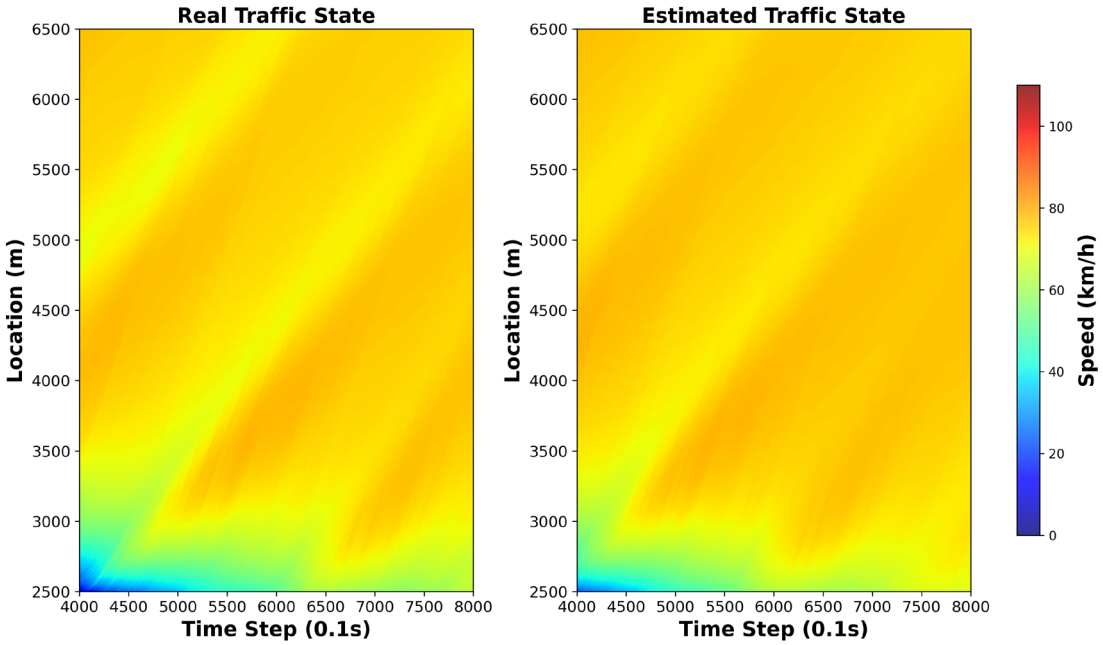

Figure 10. Real and estimated traffic states under queue-type fluctuations.

Figure 11 illustrates the time series of the free-flow speed and jam density for both upstream and downstream agents. The results show that, as congestion forms downstream, the estimated downstream free-flow speed parameter recovers toward approximately 100 km/h, while the upstream free-flow speed parameter fluctuates between 60 and 80 km/h, consistent with upstream propagation of bottleneck-induced congestion. This is consistent with real-world traffic dynamics, where congestion from a bottleneck moves upstream, slowing down vehicles.



Figure 11. Time series of free-flow speed and jam density for downstream and upstream agents under queue-type fluctuations.

In this case, the downstream speed gradually increases as the congestion dissipates, while the upstream speed exhibits more substantial fluctuations due to the vehicles' interactions within the bottleneck area. The downstream density increases as vehicles accumulate near the bottleneck and decreases as the flow resumes after the bottleneck, reflecting

the typical recovery of free-flow conditions. Conversely, the upstream density gradually builds up as more vehicles arrive in the area, signaling the formation of a queue that extends upstream.

These results confirm that the proposed multi-agent reinforcement learning framework can accurately model traffic wave phenomena, including the dissipation of congestion downstream and the gradual accumulation of traffic upstream. By capturing these dynamics, the model offers valuable insights into how congestion propagates and dissipates in real-world scenarios, supporting the development of more effective traffic management strategies.

Integrating the results from the two typical fluctuation experiments, it is evident that the proposed multi-agent reinforcement learning framework not only achieves high-precision traffic state estimation under steady-state conditions but also learns and reproduces the spatiotemporal evolution mechanisms of traffic flow in dynamic, non-equilibrium traffic environments. Through interaction with the LWR solver, the agents learn to adjust fundamental diagram parameters in a way that is consistent with flow conservation under coupled speed–density disturbances. This physics-consistent learning mechanism helps maintain stable estimation performance under complex disturbances.

7. Discussion

From a methodological standpoint, existing studies on TSE can generally be categorized into three main paradigms. The first paradigm comprises physics-driven approaches, which rely on macroscopic traffic flow theory (e.g., the LWR model and variational principles) to reconstruct traffic states by enforcing conservation laws and fundamental diagram (FD) assumptions [18]. These methods offer strong physical interpretability and robustness under sparse sensing conditions; however, their performance is often constrained by stringent model assumptions and reduced flexibility when traffic conditions deviate from stationarity. The second paradigm is purely data-driven approaches, including recent deep learning and generative modeling techniques, which directly learn complex spatiotemporal mappings from historical data and can capture nonlinear dependencies and uncertainty [22]. However, such methods typically lack explicit physical constraints, suffer from limited interpretability, and demonstrate poor generalization when sensor layouts or traffic patterns change. The third paradigm is physics–data hybrid approaches, such as physics-informed neural networks, which embed traffic flow equations as soft constraints during model training to balance physical consistency and data-driven expressiveness [32]. While these approaches alleviate data sparsity to some extent, they often involve complex training procedures and still require retraining or recalibration when applied to new scenarios. Collectively, these studies highlight a fundamental challenge in TSE: achieving a balance between physical consistency, adaptability to dynamic environments, and robust generalization under incomplete observations, which motivates exploring alternative formulations beyond direct state regression.

Against this background, the proposed multi-agent reinforcement learning (MARL) framework recasts traffic state estimation as a sequential decision-making problem rather than a static prediction task. Instead of directly estimating traffic states, agents learn to adaptively adjust fundamental diagram parameters based on environmental feedback, with the LWR model serving as a physics-based environment that governs state evolution. This design yields several key advantages. First, when ultra-high single-step estimation accuracy is not the primary objective, the proposed method exhibits superior generalization because the learned policy can be directly transferred to new highway configurations, detector layouts, and traffic demand levels without retraining. Second, the agents' behaviors are environment-adaptive: their actions dynamically change in response to evolving traffic

conditions, leading to regime-specific parameter adjustment strategies under free-flow, congested, or transitional regimes. This behavior suggests that agents do not merely memorize data patterns but instead learn underlying traffic dynamics and wave-propagation mechanisms through interaction with the physical model. Third, the multi-agent structure enables implicit spatial coordination via conservation laws and traffic flow dynamics, ensuring global consistency while preserving local adaptability. Overall, the proposed MARL-based framework shifts the focus of traffic state estimation from minimizing instantaneous prediction error to learning physically grounded, behaviorally adaptive policies, offering a promising direction for scalable and robust deployment in real-world intelligent transportation systems.

8. Conclusions

This study proposes a hybrid estimation framework that integrates multi-agent reinforcement learning (MARL) with the LWR traffic flow model, thereby combining the flexibility of data-driven learning and the interpretability of physics-based modeling for highway traffic state estimation. The experimental results show that the proposed method consistently outperforms representative baseline approaches in estimation accuracy, real-time performance, physical consistency, and generalization. Moreover, the proposed framework captures the spatiotemporal evolution of traffic states under complex fluctuation patterns.

A central contribution of this work is the development of an estimation framework that balances physical interpretability and learning flexibility. The framework facilitates the integration of traffic flow modeling and intelligent decision-making within intelligent transportation systems (ITSs) and offers practical guidance for deploying real-world traffic monitoring systems.

Although the proposed approach yields encouraging results, several limitations warrant further investigation. Future work may proceed along two complementary directions. First, incorporating additional data sources and evaluating the method under a wider range of traffic scenarios may further improve adaptability and robustness. Moreover, the current study focuses on single-lane macroscopic traffic flow without lane-changing effects and adopts an equilibrium LWR–FD formulation with free-flow speed and jam density as learnable parameters; therefore, non-equilibrium mechanisms such as capacity drop, congestion hysteresis, and multi-lane interactions are not explicitly modeled, and due to cross-corridor or cross-country heterogeneity in infrastructure and driving behavior, re-training or fine-tuning with local data is recommended prior to deployment for higher accuracy. Second, extending the framework to non-equilibrium traffic flow models could enhance its ability to represent abrupt events and complex wave dynamics. Overall, the proposed framework offers an efficient, physics-consistent solution for real-time traffic state estimation in intelligent transportation systems and shows promise for practical deployment.

Author Contributions: Conceptualization, X.Z. and Y.H.; methodology, X.Z. and Y.H.; software, X.Z.; validation, X.Z. and Y.H.; formal analysis, X.Z.; investigation, X.Z.; resources, X.Z. and Y.H.; data curation, X.Z.; writing—original draft preparation, X.Z. and Y.H.; writing—review and editing, X.Z. and Y.H.; visualization, X.Z. and Y.H.; supervision, Y.H.; project administration, Y.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to privacy.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wang, Y.; Luo, R.; Yang, X. Urban traffic state sensing and analysis based on etc data: A survey. *Appl. Sci.* **2025**, *15*, 6863. [CrossRef]
2. Wang, N.; Zhang, B.; Gu, J.; Kong, H.; Hu, S.; Lu, S. Urban road traffic spatiotemporal state estimation based on multivariate phase space–lstm prediction. *Appl. Sci.* **2023**, *13*, 12079. [CrossRef]
3. Yamada, T.; Shi, S. Estimating infection-related human mobility networks based on time series data of COVID-19 infection in japan. *Appl. Sci.* **2022**, *12*, 9236. [CrossRef]
4. Wu, F.; Cheng, Z.; Chen, H.; Qiu, Z.; Sun, L. Traffic state estimation from vehicle trajectories with anisotropic gaussian processes. *Transp. Res. Part C Emerg. Technol.* **2024**, *163*, 104646. [CrossRef]
5. Zhao, W.; Wang, T.; Zou, G.; Wang, H.; Li, Y. Adaptive deployment of fixed traffic detectors based on attention mechanism. *Systems* **2025**, *13*, 887. [CrossRef]
6. Sattarzadeh, A.R.; Pathirana, P.N.; Huynh, V.T. Traffic state estimation with spatio-temporal autoencoding transformer (stat model). *IEEE Access* **2025**, *13*, 87048–87067. [CrossRef]
7. Wang, S.; Dong, C.; Shao, C.; Luo, S.; Zhang, J.; Meng, M. Traffic state estimation incorporating heterogeneous vehicle composition: A high-dimensional fuzzy model. *Front. Eng. Manag.* **2024**, *12*, 952–970. [CrossRef]
8. Wang, Y.; Papageorgiou, M. Real-time freeway traffic state estimation based on extended kalman filter: A general approach. *Transp. Res. Part B Methodol.* **2005**, *39*, 141–167. [CrossRef]
9. Xing, J.; Wu, W.; Cheng, Q.; Liu, R. Traffic state estimation of urban road networks by multi-source data fusion: Review and new insights. *Phys. A Stat. Mech. Its Appl.* **2022**, *595*, 127079. [CrossRef]
10. Wang, T.; Li, Y.; Cheng, R.; Zou, G.; Dantsuji, T.; Ngoduy, D. Knowledge-data fusion oriented traffic state estimation: A stochastic physics-informed deep learning approach. *Transp. Res. Part C Emerg. Technol.* **2026**, *182*, 105422. [CrossRef]
11. Saleem, M.M.; Jo, H. Identifying traffic-induced bridge excitations using an optimised state estimation method. *Proc. Inst. Civ. Eng. Struct. Build.* **2025**, *178*, 161–172. [CrossRef]
12. Kyriacou, V.; Englezou, Y.; Panayiotou, C.G.; Timotheou, S. Bayesian traffic state estimation using extended floating car data. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 1518–1532. [CrossRef]
13. Huang, D.; Zhang, J.; Liu, Z.; Liu, R. Prescriptive analytics for freeway traffic state estimation by multi-source data fusion. *Transp. Res. Part E Logist. Transp. Rev.* **2025**, *198*, 104105. [CrossRef]
14. Rasulmukhamedov, M.M.; Tashmetov, K.S.; Tashmetov, T.S. Modeling traffic flow using the lwr model and finite difference numerical method: Analysis of efficiency and applicability in urban transport infrastructure conditions. In *Sustainable Development of Transport: Economy, Transformation, Logistics and ESG Agenda. Volume 2*; Sultanova, S.M., Babakhanova, N.U., Shaumarov, S.S., Masharipov, M.S.N., Eds.; Springer Nature: Cham, Switzerland, 2025; pp. 77–84.
15. Li, L.; Chen, X.; Zhang, L. Multimodel ensemble for freeway traffic state estimations. *IEEE Trans. Intell. Transp. Syst.* **2014**, *15*, 1323–1336. [CrossRef]
16. Fountoulakis, M.; Bekiaris-Liberis, N.; Roncoli, C.; Papamichail, I.; Papageorgiou, M. Highway traffic state estimation with mixed connected and conventional vehicles: Microscopic simulation-based testing. *Transp. Res. Part C Emerg. Technol.* **2017**, *78*, 13–33. [CrossRef]
17. Zheng, F.; Jabari, S.E.; Liu, H.X.; Lin, D. Traffic state estimation using stochastic lagrangian dynamics. *Transp. Res. Part B Methodol.* **2018**, *115*, 143–165. [CrossRef]
18. Takenouchi, A.; Kawai, K.; Kuwahara, M. Traffic state estimation and its sensitivity utilizing measurements from the opposite lane. *Transp. Res. Part C Emerg. Technol.* **2019**, *104*, 95–109. [CrossRef]
19. Lee, E.H.; Lee, E. Congestion boundary approach for phase transitions in traffic flow. *Transp. B Transp. Dyn.* **2024**, *12*, 2379377. [CrossRef]
20. Ni, D. There is something more fundamental than fundamental diagram. *Transp. Res. Part B Methodol.* **2025**, *195*, 103206. [CrossRef]
21. Xu, D.; Wei, C.; Peng, P.; Xuan, Q.; Guo, H. Ge-gan: A novel deep learning framework for road traffic state estimation. *Transp. Res. Part C Emerg. Technol.* **2020**, *117*, 102635. [CrossRef]
22. Lei, D.; Xu, M.; Wang, S. A conditional diffusion model for probabilistic estimation of traffic states at sensor-free locations. *Transp. Res. Part C Emerg. Technol.* **2024**, *166*, 104798. [CrossRef]
23. Rowan, D.; He, H.; Hui, F.; Yasir, A.; Mohammed, Q. A systematic review of machine learning-based microscopic traffic flow models and simulations. *Commun. Transp. Res.* **2025**, *5*, 100164. [CrossRef]
24. Kawasaki, Y.; Hara, Y.; Kuwahara, M. Traffic state estimation on a two-dimensional network by a state-space model. *Transp. Res. Procedia* **2019**, *38*, 299–319. [CrossRef]

25. Lu, J.; Li, C.; Wu, X.B.; Zhou, X.S. Physics-informed neural networks for integrated traffic state and queue profile estimation: A differentiable programming approach on layered computational graphs. *Transp. Res. Part C Emerg. Technol.* **2023**, *153*, 104224. [CrossRef]
26. Huang, A.J.; Filipovska, M. Physics-informed bayesian deep learning for traffic state estimation and uncertainty quantification. In Proceedings of the 2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC), Edmonton, AB, Canada, 24–27 September 2024; pp. 4288–4293. [CrossRef]
27. Luo, X.; Yu, H.; Yuan, H.; Wang, Y. A model and data dual-driven approach to traffic state estimation on freeways. In Proceedings of the 2024 5th International Conference on Computer, Big Data and Artificial Intelligence (ICCBD+AI), Jingdezhen, China, 1–3 November 2024; pp. 114–119. [CrossRef]
28. Zhao, C.; Yu, H. Observer-informed deep learning for traffic state estimation with boundary sensing. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 1602–1611. [CrossRef]
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
30. Kachroo, P.; Agarwal, S.; Ozbay, K. Macroscopic Fundamental Diagram: Alternative theoretical analysis and implications for traffic control. *IEEE Open J. Intell. Transp. Syst.* **2024**, *5*, 826–841. [CrossRef]
31. Hoogendoorn, S.P.; Bovy, P.H.L. State-of-the-art of vehicular traffic flow modelling. *Proc. Inst. Mech. Eng. Part I J. Syst. Control Eng.* **2001**, *215*, 283–303.
32. Zhang, J.; Mao, S.; Yang, L.; Ma, W.; Li, S.; Gao, Z. Physics-informed deep learning for traffic state estimation based on the traffic flow model and computational graph method. *Inf. Fusion* **2024**, *101*, 101971. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.