Physics-Informed Directed Spatio-Temporal Graph Convolutional Networks for Traffic Flow Prediction

Anonymous authors
Paper under double-blind review

Abstract

Accurate urban traffic flow prediction is crucial for Intelligent Transportation Systems (ITS). Existing spatio-temporal neural networks typically adopt symmetric adjacency matrices, failing to capture the asymmetric spatial dependencies of traffic flow, and lack macroscopic physical constraints, leading to non-conservative predictions in long-term forecasting. To address these issues, we propose PI-Dir-STGCN. Unlike prior methods that loosely couple physical penalties with graph networks, our framework achieves a tight structural coupling between directed graph structures and macroscopic traffic dynamics. Specifically, we construct an asymmetric directed weighted adjacency matrix to capture these asymmetric traffic dependencies, and develop a graph-topology generalized upwind difference scheme to embed the LWR model into the graph neural network. This innovation transforms the directed graph into a numerical grid for fluid dynamics and encourages mass-conserving predictions through a computable physics residual. Experiments on PeMS04 and PeMS08 show that our model effectively suppresses cumulative errors, integrates data-driven learning with physical evolution, and provides strong physical interpretability.

1 Introduction

With the acceleration of global urbanization and the continuous growth of vehicle ownership, urban traffic congestion has become a global problem restricting social and economic development. Intelligent Transportation System (ITS) relies on accurate traffic flow prediction to support traffic control, route guidance, and congestion warning. However, urban traffic flow is a typical spatio-temporal coupled complex system, whose evolution is simultaneously constrained by the spatial topology and temporal dynamics. In the spatial dimension, vehicle movements strictly follow directed road connectivity, whereas traffic-state disturbances (e.g., congestion) can propagate both downstream and upstream. In the temporal dimension, traffic flow exhibits daily/weekly periodicity and non-linear dynamics such as peaks. These complex dependencies make traditional methods unable to meet medium- and long-term accuracy requirements. How to effectively capture spatio-temporal coupling while ensuring physical consistency remains a key challenge.

Existing traffic flow forecasting methods fall into three categories. First, data-driven spatio-temporal graph neural networks (STGNNs) have become the mainstream paradigm: STGCN (Yu et al., 2018) first stacks spatio-temporal convolution blocks in a "sandwich" structure to jointly extract spatial and temporal features, and subsequent works including ASTGCN (Guo et al., 2019), Graph WaveNet (Wu et al., 2019) and MTGNN (Wu et al., 2020) further introduce attention mechanisms and adaptive graphs to boost performance. However, although these models can learn asymmetric adjacency matrices, they still lack explicit physical causal constraints, failing to fully match the inherent asymmetric causality of macroscopic traffic flow. Second, pure physics-informed methods built on PINNs (Raissi et al., 2019) solve LWR-derived traffic PDEs (Lighthill and Whitham, 1955; Richards, 1956) with strong interpretability, yet they focus on state reconstruction rather than future prediction and are limited to one-dimension road segments. Third, physics-guided graph learning incorporates LWR priors (Pan et al., 2025) as regularization or adjacency guidance, but treats graph topology and physical dynamics as loosely coupled modules; standard symmet-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ric graph convolutions also conflict with fluid dynamic directionality, violating the upwind scheme required for numerical PDE solving. Deep integration of graph topology and macroscopic traffic dynamics thus remains an open problem.

Despite significant advances across these three lines of research, fundamental limitations persist and three critical gaps remain unaddressed. First, the spatial propagation mechanisms in most existing STGNNs lack strict alignment with the physical directionality of traffic flow; even with adaptive asymmetric graphs, the absence of macroscopic dynamic constraints leaves them prone to learning spurious correlations rather than true causal dependencies, resulting in over-smoothed features and phase lag in prediction results. Second, existing physics-informed methods rely on one-dimensional road discretization schemes that cannot be generalized to arbitrary graph topologies. Third, neither data-driven nor physics-informed approaches enforce effective physical constraints to suppress error accumulation in long-horizon multi-step forecasting, leading to non-physical phenomena. To address these gaps, this paper proposes PI-Dir-STGCN, a physics-informed directed spatio-temporal graph convolutional network, with three main contributions:

1. Asymmetric directed graph convolution operator. We construct an asymmetric directed weighted adjacency matrix with a minimal attenuation factor, and design a random-walk normalized directed graph convolution operator. It accurately captures the one-way causal dependencies between upstream and downstream road segments, effectively solving the structural defect of traditional undirected graph convolution that violates traffic flow directionality.

2. Graph-topology generalized upwind discretization. Unlike methods that naively combine graph networks with physical penalties, we propose a theoretically principled coupling. We design a graph-topology generalized upwind scheme that directly aligns the message-passing on our directed graph with the advection operator of the LWR PDE. This innovation transforms the directed graph from a feature extractor into a computational grid for fluid dynamics.

3. Physical regularization for long-horizon stability. We introduce the physics residual as a physics-guided regularization term into the loss function, which constrains the model's evolution boundary from the underlying level. Extensive experiments show that the proposed model effectively suppresses error accumulation in long-horizon prediction.

2 Preliminaries

2.1 Problem Formulation

Urban traffic flow is a typical spatio-temporal coupled complex system. To capture the complex spatial dependencies and temporal dynamics, we abstract the traffic network as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, A_{\text{dir}})$, where $\mathcal{V}$ is the set of n fixed roadside sensors (nodes), $\mathcal{E}$ is the set of directed edges aligned with vehicle travel direction , and $A_{\text{dir}} \in \mathbb{R}^{n\times n}$ is the weighted adjacency matrix quantifying the propagation strength and direction of traffic flow between road segments.

At time step t, let q_i^t, O_i^t, and v_i^t denote the traffic flow, occupancy, and space-mean speed at node i. In macroscopic traffic flow theory, density ρ is the core control variable for constructing conservation equations. but it is not directly collected in real engineering datasets such as the Caltrans PeMS system; loop detectors instead output three basic parameters: flow, occupancy and speed. Since occupancy is linearly proportional to density, we adopt O_i^t as a proxy variable for macroscopic density. To ensure the dimensional self-consistency of physical equations in the discrete data field, we introduce a node-specific scaling scalar c_i to formulate the calibrated density proxy as $\tilde{\rho}_i^t = c_i O_i^t$.

Thus, the traffic prediction problem is defined as: given the historical traffic state observation matrix sequence $\mathcal{X} = \{q_t, O_t, v_t \mid t = T-M+1, \ldots, T\}$ of M consecutive time steps, predict the future spatio-temporal evolution sequence $\hat{Y} = \{\hat{q}_t, \hat{O}_t, \hat{v}_t \mid t = T+1, \ldots, T+H\}$ of H

time steps:

$$\hat{Y} = \arg\max \log P(\hat{q}_{T+1:T+H}, \hat{O}_{T+1:T+H}, \hat{v}_{T+1:T+H} \mid q_{T-M+1:T}, O_{T-M+1:T}, v_{T-M+1:T}). \tag{1}$$

Unlike traditional pure data-driven methods, we further require that the predictions satisfy the fundamental physical law of traffic flow, the mass conservation law, to enhance physical interpretability and long-term prediction stability.

2.2 Review of Spatio-Temporal Graph Convolutional Networks (STGCN)

STGCN is the first deep learning framework applying a pure convolutional architecture to graph-structured traffic time series prediction. Its core is stacked spatio-temporal convolution blocks (ST-Conv blocks) for joint spatial-temporal feature extraction, each adopting a "sandwich" structure with a spatial graph convolution layer sandwiched between two temporal gated convolution layers.

The temporal convolution layer contains a 1D causal convolution of width K_t, followed by a Gated Linear Unit (GLU) as non-linearity. Formally, given input $Y \in \mathbb{R}^{M \times C_i}$ and kernel $\Gamma \in \mathbb{R}^{K_t \times C_i \times 2C_o}$, the output is split into P and Q along the channel dimension, and the temporal gated convolution is defined as:

$$\Gamma *_{\mathcal{T}} Y = P \odot \sigma(Q) \in \mathbb{R}^{(M-K_t+1) \times C_o}, \tag{2}$$

where $\odot$ is element-wise Hadamard product and σ is the sigmoid activation.

Graph convolution extracts spatial patterns directly from graph-structured data. With stacked localized layers based on first-order approximation of the graph Laplacian (Kipf and Welling, 2017), the layer-wise formulation is:

$$\Theta *_{\mathcal{G}} X = \theta(\tilde{D}^{-1/2}\tilde{W}\tilde{D}^{-1/2})X, \tag{3}$$

where $\tilde{W} = W + I_n$ and $\tilde{D}_{ii} = \sum_j \tilde{W}_{ij}$. However, this symmetric normalized Laplacian assumes equal bidirectional information propagation, violating the directionality of traffic flow.

Despite the progress made by attention-based models like ASTGCN (Guo et al., 2019), pure data-driven approaches face two bottlenecks: (1) parameter redundancy and computational latency; (2) lack of physical interpretability.

2.3 LWR Equations

To endow the model with physical awareness, we introduce the PDE prior from macroscopic traffic flow theory. The Lighthill-Whitham-Richards (LWR) model (Lighthill and Whitham, 1955; Richards, 1956) describes traffic flow evolution through the continuity equation (mass conservation):

$$\partial_t \rho + \partial_x(\rho v) = 0, \tag{4}$$

where ρ and v are macroscopic density and speed, and $q = \rho v$ is the traffic flow. We adopt the mass conservation law as a fundamental physical constraint, which is sufficient to promote consistency with the conservation law. By discretizing the continuity equation into a computable physics residual and incorporating it as a regularization term into the loss function, we guide the network to learn conservation-compliant spatio-temporal mappings, thus enhancing long-term prediction stability.

3 Methodology

3.1 Overall Design

To address the difficulties of pure data-driven models in capturing directionality, error accumulation, and violation of physical conservation laws in long-term forecasting, we propose PI-Dir-STGCN. Its overall architecture is shown in Fig. 1. The model uses three-channel

(flow, occupancy, speed) multi-variate input; in the spatial dimension, it constructs an asymmetric directed weighted adjacency matrix and designs a random-walk directed graph convolution operator; in the temporal dimension, it follows STGCN's gated causal convolution; it also introduces an LWR mass conservation based physics constraint layer, jointly optimizing data-driven loss and physics residual.

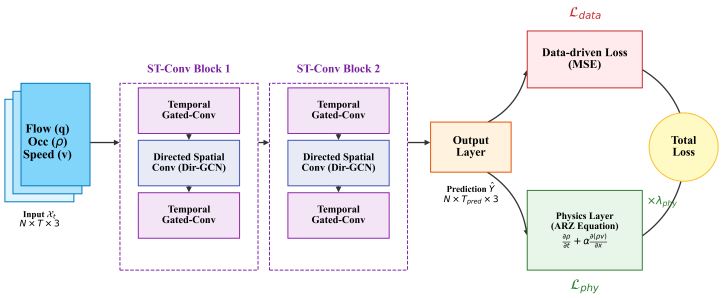

Figure 1: PI-Dir-STGCN architecture

3.2 Model Architecture

PI-Dir-STGCN consists of three core modules: (1) directed graph convolution layer based on spatial anisotropy, (2) gated temporal convolution layer, and (3) LWR physics-constrained loss function. The model stacks two directed ST-Conv blocks to jointly extract spatio-temporal features.

3.2.1 Directed Graph Convolution

While vehicle movements strictly follow directed road connectivity, traffic-state disturbances (e.g., congestion shockwaves) can propagate both downstream and upstream depending on the traffic regime. To accommodate this, we construct an asymmetric directed weighted adjacency matrix $A_{\text{dir}} \in \mathbb{R}^{n \times n}$:

$$A_{ij} = \begin{cases} \exp\left(-\dfrac{d_{ij}^2}{\sigma^2}\right), & \text{if node } j \text{ is upstream of node } i, \\ \epsilon \cdot \exp\left(-\dfrac{d_{ij}^2}{\sigma^2}\right), & \text{if node } j \text{ is downstream of node } i \quad (\epsilon \ll 1), \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

We assign strong weights to downstream information propagation while retaining a controlled reverse channel (via parameter ϵ which is empirically set to 10^{-3} based on a preliminary grid search) for upstream-moving congestion effects. This design balances flow directionality with physical realism, avoiding the over-simplified assumption of strict one-way causality. Because A is not symmetric, the symmetric normalized Laplacian used

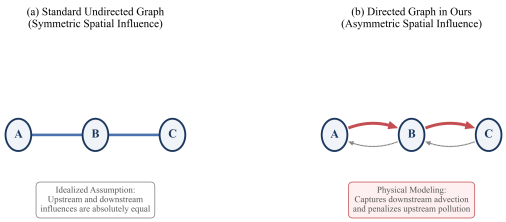

Figure 2: Directed vs. undirected graph concept

in standard STGCN is no longer valid. We perform random-walk normalization on the

directed matrix after adding self-loops:

$$\tilde{A} = A_{\text{dir}} + I_n, \quad \tilde{D}_{ii} = \sum_{j=1}^{n} \tilde{A}_{ij}. \tag{6}$$

The directed graph convolution operator $*_{\vec{\mathcal{G}}}$ is then defined as:

$$\Theta *_{\vec{\mathcal{G}}} X = \tilde{D}^{-1}\tilde{A}X\Theta, \tag{7}$$

where $X \in \mathbb{R}^{n\times C_i}$ is the spatial input feature matrix at a time step, and $\Theta \in \mathbb{R}^{C_i\times C_o}$ is the learnable mapping weight. The node-level aggregation then follows the standard random-walk normalized propagation.We employ the same gated causal convolution as STGCN (Eq. (2)) for temporal feature extraction, applied independently to each node.

3.2.2 Directed ST-Conv Block

To fuse directed spatial and temporal domains, we build a Directed ST-Conv block that maintains the "sandwich" structure. For block l with input $\mathcal{H}^{(l)} \in \mathbb{R}^{M\times n\times C^{(l)}}$, the output is:

$$\mathcal{H}^{(l+1)} = \Gamma_1^{(l)} *_{\mathcal{T}} \text{ReLU}\left(\Theta^{(l)} *_{\vec{\mathcal{G}}} \left(\Gamma_0^{(l)} *_{\mathcal{T}} \mathcal{H}^{(l)}\right)\right). \tag{8}$$

Our network stacks two such blocks, each with residual connections and bottleneck structures.

3.2.3 LWR Physics-Constrained Loss

To discretize $\partial_t\rho + \partial_x(\rho v) = 0$, we use forward difference for time, i.e., $\partial_t\rho \approx (\hat{\rho}_i^{t+1} - \hat{\rho}_i^t)/\Delta t$. For the spatial flux derivative $\partial_x(\rho v)$, we propose a routing-aware graph-based finite-volume conservation residual. While the directed adjacency A_{dir} permits backward propagation of congestion shockwaves via ϵ-weighted edges for graph convolution, traffic mass routing must strictly follow the physical forward direction. We thus define a forward-flow adjacency matrix $B_{ij} = A_{ij}\mathbf{1}\{j \in \mathcal{N}_{\text{up}}(i)\}$ to isolate forward flow, and normalize the outgoing weights of each upstream node j to derive the routing ratio ω_{ji} from j to downstream node i:

$$\omega_{ji} = \frac{B_{ij}}{\sum_{k\in\mathcal{N}_{\text{down}}(j)} B_{kj}} \tag{9}$$

In practical traffic datasets, exact vehicle density ρ is not directly observed; instead, road occupancy O is collected. We introduce a scaling scalar c_i to construct a calibrated density proxy $\tilde{\rho}_{it} = c_i\hat{O}_{it}$, yielding the physical traffic flux $\tilde{F}_{it} = \tilde{\rho}_{it}\hat{v}_{it} = c_i\hat{O}_{it}\hat{v}_{it}$. We normalize the routing ratios such that $\sum_{i\in\mathcal{N}_{\text{down}}(j)} \omega_{ji} = 1$ for all internal nodes, mathematically ensuring exact conservation of upstream node j's total out-flux across its downstream neighbors and avoiding non-physical vehicle gain or loss at intersections. The generalized spatial flux derivative for node i's control volume is formulated as:

$$\partial_x(\rho v) \approx \frac{\tilde{F}_i^t - \sum_{j\in\mathcal{N}_{\text{up}}(i)} \omega_{ji}\tilde{F}_j^t}{L_i} = \frac{c_i\hat{O}_i^t\hat{v}_i^t - \sum_{j\in\mathcal{N}_{\text{up}}(i)} \omega_{ji}c_j\hat{O}_j^t\hat{v}_j^t}{L_i}. \tag{10}$$

where L_i represents the effective road segment length associated with sensor i. Our model predicts three channels $(\hat{q}, \hat{O}, \hat{v})$. To effectively link the directly predicted traffic flow $\hat{q}$ with the physical flux derived from occupancy and speed, we introduce a flux-consistency loss:

$$\mathcal{L}_{\text{cons}} = \frac{1}{NT}\sum_{i,t}\left(\hat{q}_i^t - c_i\hat{O}_i^t\hat{v}_i^t\right)^2, \tag{11}$$

where c_i is a scaling scalar that can be jointly optimized as a learnable parameter during training or pre-computed via least-squares estimation from the training set. The purely data-driven loss and the physics penalty are updated as follows:

$$\mathcal{L}_{\text{data}} = \frac{1}{NT}\sum_{i,t}\left(\|\hat{q}_i^t - q_i^t\|^2 + \gamma\left(\|\hat{v}_i^t - v_i^t\|^2 + \|\hat{O}_i^t - O_i^t\|^2\right)\right), \tag{12}$$

$$\mathcal{L}_{\text{phy}} = \frac{1}{NT} \sum_{i,t} |R_{\text{phy}}(i,t)|^2. \tag{13}$$

Here, $R^{(q)}_{\text{phy}}(i,t) = c_i \frac{\hat{O}_i^{t+1} - \hat{O}_i^t}{\Delta t} + \frac{\hat{q}_i^t - \sum_{j \in \mathcal{N}_{\text{up}}(i)} \omega_{ji} \hat{q}_j^t}{L_i}$ is the discretized LWR continuity residual rigorously scaled by c_i. The total loss is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{data}} + \lambda_{\text{phy}} \mathcal{L}_{\text{phy}} + \lambda_{\text{cons}} \mathcal{L}_{\text{cons}}. \tag{14}$$

Note that λ_{cons} is empirically fixed at 10^{-5}, as it merely serves as a soft consistency regularizer to align the physical units between the direct output and the proxy states, rather than a primary driving objective. By co-optimizing these terms, our framework encourages mass-conserving predictions.

4 Experiments and Results

4.1 Experimental Setup

Data Sources and Sample Selection

We use two authoritative public benchmark datasets from the Caltrans Performance Measurement System (PeMS): PeMS04 and PeMS08. PeMS04 contains traffic data from 307 sensors in the San Francisco Bay Area over 59 days, while PeMS08 covers 170 sensors in the San Bernardino area over 62 days. Both datasets provide flow, occupancy, and speed. The raw data are collected at 30-second intervals and aggregated to 5-minute steps (288 time steps per day per sensor). We use the standard 6:2:2 chronological split for training, validation, and testing, and adopt the previous 12 time steps (1 hour) to predict the next 12 time steps (1 hour).

Baselines include HA (Williams and Hoel, 2003), ARIMA (Williams and Hoel, 2003), VAR (Zivot and Wang, 2006), LSTM (Hochreiter and Schmidhuber, 1997), GRU (Chung et al., 2014), TCN (Bai et al., 2018), STAEformer (Liu et al., 2023), Graph WaveNet (Wu et al., 2019) and standard STGCN (Yu et al., 2018). We set spatial kernel $K_s = 1$, temporal kernel $K_t = 3$, and train with RMSprop (batch size 32, 50 epochs, initial learning rate 10^{-3}, decay factor 0.7 every 1000 steps). Key hyperparameters: $\lambda_{\text{phy}} = 5 \times 10^{-4}$, $\epsilon = 10^{-3}$, $\sigma^2 = 10.0$. We adopt three standard traffic forecasting metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), with a zero-flow mask for MAPE computation.

4.2 Overall Performance Comparison

To evaluate the overall 1-hour performance, Table 1 compares the average prediction errors of our model and baselines on both datasets.

Table 1: Overall 1-hour average error comparison on PeMS04 and PeMS08.

Model	PeMS04		PeMS08	
	MAE	RMSE	MAE	RMSE
HA	36.76	54.14	29.52	44.03
ARIMA	32.11	68.13	24.04	43.30
VAR	33.76	51.73	21.41	31.21
LSTM	29.45	45.82	23.18	36.96
GRU	28.65	45.11	22.20	35.95
Graph WaveNet	25.45	39.70	19.13	31.05
STAEformer	18.23	30.30	13.58	23.43
TCN	31.11	37.25	22.69	30.60
STGCN	25.93	39.47	20.16	35.79
Ours	23.36	36.18	18.73	28.91

As shown in Table 1, Traditional time-series models (HA, ARIMA) fail to capture the nonlinear complexity of traffic flow. Deep learning methods including LSTM, GRU and TCN yield better results yet ignore spatial dependencies. Spatio-temporal models such as STGCN and Graph WaveNet further improve performance via graph convolutions. Attention-based STAEformer achieves state-of-the-art point-estimation accuracy through heavy parameterization, yet acts as a deterministic black box restricted to single-variable prediction. Because it omits critical states like occupancy and speed, it precludes the mathematical verification of macroscopic mass conservation. In contrast, our PI-Dir-STGCN explicitly predicts all three physical variables and embeds the LWR conservation law into the directed graph topology without computationally heavy attention modules. This structural coupling allows us to actively monitor and minimize physical violations, achieving a low Conservation Error (CE, $\overline{\text{CE}} = \frac{1}{NT}\sum |R_{\text{phy}}(i,t)|$) of 0.0418. The model outperforms standard STGCN, delivering competitive prediction accuracy alongside improved conservation consistency and physical reliability critical for real-world traffic management.

4.3 Multi-step Long-term Prediction Evolution

Most existing studies only report overall 1-hour average errors. However, the challenge of traffic forecasting is error accumulation in long-term rolling prediction, where models recursively feed on their own outputs. To verify the stability benefit of the continuity equation, we decompose predictions of the baseline and our model into four horizons (15, 30, 45, 60 min) in Table 2. Figure 3 plots MAE and RMSE trends over prediction steps on PeMS04, with corresponding PeMS08 results given in Appendix A.3.

Table 2: Step-wise prediction errors (15/30/45/60 min).

Dataset	Horizon	Model	MAE	RMSE	MAPE
PeMS04	15 min	STGCN (Baseline)	20.82	32.39	12.76%
		PI-Dir-STGCN (Ours)	20.23	31.72	12.73%
	30 min	STGCN (Baseline)	24.82	37.55	14.72%
		PI-Dir-STGCN (Ours)	23.00	35.45	14.18%
	45 min	STGCN (Baseline)	29.75	43.82	17.14%
		PI-Dir-STGCN (Ours)	25.75	39.17	15.68%
	60 min	STGCN (Baseline)	34.34	49.67	19.39%
		PI-Dir-STGCN (Ours)	28.35	42.62	17.17%
PeMS08	15 min	STGCN (Baseline)	16.09	24.76	10.03%
		PI-Dir-STGCN (Ours)	15.89	24.53	9.90%
	30 min	STGCN (Baseline)	19.29	29.11	11.65%
		PI-Dir-STGCN (Ours)	18.44	28.23	11.39%
	45 min	STGCN (Baseline)	23.12	34.08	13.44%
		PI-Dir-STGCN (Ours)	20.90	31.74	12.74%
	60 min	STGCN (Baseline)	27.15	39.08	15.12%
		PI-Dir-STGCN (Ours)	23.24	35.01	14.04%

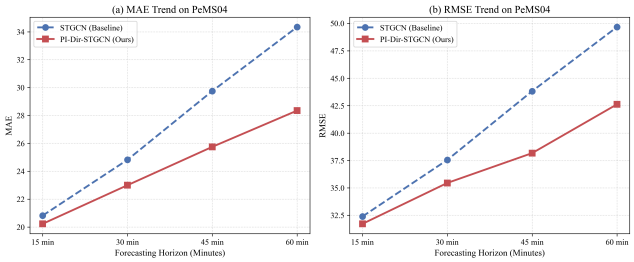

Figure 3: MAE and RMSE trends on PeMS04 over prediction horizons.

These present the step-wise results of a representative run (with a fixed random seed). As shown in Table 2 and Figure 3, at the 15-minute horizon, our model slightly improves MAE,

demonstrating effective multi-variate feature fusion. At 45-60 minutes, STGCN suffers from rapid error accumulation, while our model, constrained by LWR conservation, stays within reasonable physical bounds, reducing MAE by 5.99 on PeMS04 and 3.91 on PeMS08. This confirms that incorporating mass conservation effectively suppresses long-term error accumulation in nonlinear traffic systems. To isolate the effect of the physics regularizer from multivariate input benefits, Appendix 4 presents step-wise comparisons between STGCN-Multi and ours. Our model maintains consistently lower errors across all horizons.

4.4 Ablation Study

To isolate the contributions of our proposed modules from the benefits of richer input features, we first include the standard STGCN-Baseline in the comparison, on this basis, we establish STGCN-Multi as a controlled baseline. It utilizes the same multi-variate inputs (flow, occupancy, speed) and output channels as our model, but retains the standard undirected graph without physics regularization. Against this baseline, we evaluate three variants: Dir-only (directed graph without physics), PI-only (undirected graph with physics), and the Full model. The overall 1-hour prediction performances are summarized in Table 3. As shown in Table 3, all ablated variants outperform the original single-variable

Table 3: Ablation study results (1-hour overall MAE/RMSE).

Model	PeMS04		PeMS08	
	MAE	RMSE	MAE	RMSE
STGCN-Baseline	25.93	39.47	20.16	35.79
STGCN-Multi	23.67 ± 0.33	36.60 ± 0.45	19.05 ± 0.08	29.41 ± 0.12
Dir-only	23.50 ± 0.28	36.24 ± 0.36	18.83 ± 0.25	29.12 ± 0.30
PI-only	23.51 ± 0.08	36.45 ± 0.15	18.86 ± 0.23	29.06 ± 0.42
Full	23.36 ± 0.13	36.18 ± 0.21	18.73 ± 0.22	28.91 ± 0.32

Note: The standard baseline is a fixed-seed single run, while others report mean ± std over 3 seeds for statistical significance.

STGCN baseline, and simply adding occupancy and speed as input features already yields noticeable improvements. Crucially, the full model still consistently outperforms STGCN-Multi across all metrics, which confirms that the performance gains do not merely come from richer input information, but more importantly from the effective synergy between the directed graph structure and physics regularization. Each module alone already improves over STGCN-Multi: the directed graph reduces errors by capturing asymmetric spatial dependencies, and applying the LWR physics constraint to the standard undirected graph also improves performance by encouraging physical consistency. When the two are combined in the full model, the directed graph provides message passing that aligns with the upwind discretization scheme, allowing the physics prior to take full effect. This structural integration further suppresses error accumulation over long prediction horizons. Together, the directed graph captures asymmetric spatial dependencies, while physics regularization improves global physical consistency and long-term stability, and their complementary benefits lead to the best overall performance.

To quantitatively verify the effectiveness of the LWR mass conservation constraint, we compare the full model against the directed-graph variant without physics loss (Dir-only). On PeMS08, the full model reduces Conservation Error from 0.0519 to 0.0418 (a 19.5% relative decrease), confirming that physics regularization effectively suppresses non-physical mass conservation violations.

4.5 Qualitative Evaluation and Error Evolution Analysis

In this section, we visualize the node-level prediction results on PeMS04, as shown in Figure 4. The network-level error distribution is deferred to Appendix A.4.

Figure 4 shows daily traffic flow curves of two arterial nodes. The baseline STGCN (blue dashed) captures overall periodicity but exhibits 5–10 min phase lag and peak underestimation during abrupt traffic changes, due to blurred upstream-downstream causality and

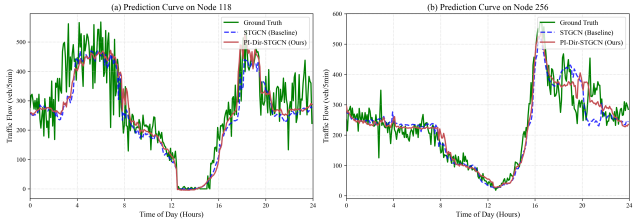

Figure 4: Prediction curves of 24-hour traffic flow on typical nodes. (a) Node 118; (b) Node 256.

feature over-smoothing from undirected graph convolution. The proposed PI-Dir-STGCN (red solid) aligns nearly perfectly with ground truth (green solid) in both phase and amplitude, even at the sharp evening peak of Node 256. Additional error distribution analysis is provided in Appendix A.4. In summary, directed graph convolution improves local dynamic response while physics regularization constrains the evolution boundary, and their synergy achieves both high prediction accuracy and strong physical interpretability.

4.6 Hyperparameter Sensitivity Analysis

The physical regularization weightλ_{phy} balances the data-driven loss and the physics constraint.We test $\lambda_{\text{phy}} = \{0, 5 \times 10^{-6}, 10^{-5}, 10^{-4}, 5 \times 10^{-4}\}$on both datasets with a fixed seed (2026) to trace the exact evolution trend. As shown in Appendix A.5, When $\lambda_{\text{phy}} = 0$, the physics constraint is disabled and the model degenerates to the directed-graph variant Dir-only, yielding the largest errors (MAE of 23.48 on PeMS04 and 18.96 on PeMS08). The MAE reaches its minimum at $\lambda_{\text{phy}} = 5 \times 10^{-4}$ (23.36 on PeMS04 and 18.73 on PeMS08), while both smaller and larger values lead to higher errors. These results validate that a moderate physical constraint best improves prediction accuracy, confirming the robustness of our chosen $\lambda_{\text{phy}} = 5 \times 10^{-4}$.

4.7 Model Complexity and Runtime Efficiency Analysis

To evaluate the computational cost of our model, we compare the parameter count, training time, and single-step inference latency of standard STGCN and PI-Dir-STGCN on PeMS04 and PeMS08. Our model requires only a marginal 6.2% increase in parameter count. Training time rises by 42.2% (PeMS04) and 53.9% (PeMS08) due to multivariate input and per-step physics residual calculation. However, inference latency remains nearly unchanged ($\approx$ 1.2 ms/sample) since physics constraints are only applied during training. Detailed quantitative comparisons of parameter counts and computational costs are provided in Appendix A.6. This confirms that the added complexity is limited to offline training and does not affect real-time online deployment.

5 Conclusion and Future Work

This work proposes PI-Dir-STGCN for traffic flow forecasting. It adopts asymmetric directed graph convolution to match traffic directionality and integrates the LWR mass conservation law into the spatio-temporal graph convolution framework via a graph-topology generalized upwind difference scheme. Physical residuals serve as regularization to mitigate long-horizon error accumulation, and a flux-consistency constraint ensures internal physical consistency across the three prediction channels. Evaluations on real-world datasets deliver a favorable accuracy-consistency trade-off: while attention-based models like STAEformer achieve lower point-estimation errors, our model effectively resolves the non-conservation issue and suppresses long-horizon error accumulation, with inference efficiency comparable to STGCN. Future work will incorporate external factors (weather, accidents, holidays) and explore reinforcement learning integration for traffic signal control.

AI Use Statement

Generative AI tools were used for language refinement, code debugging, and critical discussion of methodological presentation. All model formulations, implementation decisions, experiments, and reported results were independently verified and approved by the authors.

Ethics Statement

This research uses publicly available traffic datasets (PeMS04 and PeMS08) and does not involve human subjects. The proposed model aims to improve traffic flow prediction and management, which contributes to societal welfare. No potential conflicts of interest, discrimination/bias/fairness concerns, privacy/security issues, or legal compliance problems are identified.

Reproducibility Statement

To ensure reproducibility, we have provided detailed descriptions of the model architecture, hyperparameters, and experimental settings in Sections 3 and 4. The source code will be made publicly available upon acceptance. All datasets used are publicly accessible. Data processing steps and evaluation protocols are described in Section 4.1.

References

B. Yu, H. Yin, and Z. Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI), pages 3634–3640, 2018.

S. Guo, Y. Lin, N. Feng, C. Song, and H. Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence, volume 33, pages 922–929, 2019.

M. J. Lighthill and G. B. Whitham. On kinematic waves II. A theory of traffic flow on long crowded roads. Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences, 229(1178):317–345, 1955.

P. I. Richards. Shock waves on the highway. Operations Research, 4(1):42–51, 1956.

A. Aw and M. Rascle. Resurrection of "second order" models of traffic flow. SIAM Journal on Applied Mathematics, 60(3):916–938, 2000.

H. M. Zhang. A non-equilibrium traffic model devoid of gas-like behavior. Transportation Research Part B: Methodological, 36(3):275–290, 2002.

M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics, 378:686–707, 2019.

Y. Li, R. Yu, C. Shahabi, and Y. Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In International Conference on Learning Representations (ICLR), 2018.

Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang. Graph wavenet for deep spatial-temporal graph modeling. In Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI), pages 1907–1913, 2019.

Z. Wu, S. Pan, G. Long, J. Jiang, X. Chang, and C. Zhang. Connecting the dots: Multivariate time series forecasting with graph neural networks. In Advances in Neural Information Processing Systems (NeurIPS), volume 33, pages 7588–7598, 2020.

L. Bai, L. Yao, C. Li, X. Wang, and C. Wang. Adaptive graph convolutional recurrent network for traffic forecasting. In Advances in Neural Information Processing Systems (NeurIPS), volume 33, pages 17804–17815, 2020.

Y. A. Pan, F. Li, A. Li, Z. Niu, and Z. Liu. Urban intersection traffic flow prediction: A physics-guided stepwise framework utilizing spatio-temporal graph neural network algorithms. Multimodal Transportation, 4(2):100207, 2025.

T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In International Conference on Learning Representations (ICLR), 2017.

P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph attention networks. In International Conference on Learning Representations (ICLR), 2018.

S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural Computation, 9(8):1735–1780, 1997.

J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In NIPS 2014 Workshop on Deep Learning, 2014.

B. M. Williams and L. A. Hoel. Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results. Journal of Transportation Engineering, 129(6):664–672, 2003.

E. Zivot and J. Wang. Vector autoregressive models for multivariate time series. In Modeling Financial Time Series with S-PLUS, pages 385–429. Springer, 2006.

C. Chen, K. Petty, A. Skabardonis, P. Varaiya, and Z. Jia. Freeway performance measurement system: Mining loop detector data. Transportation Research Record, 1748(1):96–102, 2001.

X. Yin, G. Wu, J. Wei, Y. Shen, H. Qi, and Y. Yin. Deep learning on traffic prediction: A survey. IEEE Transactions on Intelligent Transportation Systems, 22(8):4927–4943, 2021.

S. Bai, J. Z. Kolter, and V. Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271, 2018.

H. Shao, J. Yao, D. Liu, Y. Wang, and Y. Liu. Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting. In Proceedings of the 36th AAAI Conference on Artificial Intelligence, volume 36, pages 8590–8598, 2022.

Z. Fang, Q. Long, G. Song, and K. Xie. Spatial-temporal graph ODE networks for traffic flow forecasting. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), pages 1907–1913, 2021.

H. Liu, Z. Dong, R. Jiang, J. Deng, J. Deng, Q. Chen, and X. Song. Spatio-temporal adaptive embedding makes vanilla transformer SOTA for traffic forecasting. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM), pages 4125–4129, 2023.

A Appendix

A.1 Derivation of the Routing-Aware Conservation Residual

In macroscopic traffic flow theory, the LWR model is defined by the 1D continuity equation:

$$\frac{\partial \rho}{\partial t} + \frac{\partial q}{\partial x} = 0, \tag{1}$$

where q is the traffic flow. In practical traffic datasets, vehicle density ρ is not directly observed; instead, road occupancy O is collected. We therefore introduce a node-specific scaling scalar c_i to construct a calibrated density proxy:

$$\tilde{\rho}_i^t = c_i \hat{O}_i^t. \tag{2}$$

To generalize the continuous PDE to a non-Euclidean directed graph, we treat each sensor node i as a discrete control volume with an effective road segment length L_i. Let m_i^t denote the total number of vehicles within the segment at time t, approximated by:

$$m_i^t \approx \tilde{\rho}_i^t L_i = c_i \hat{O}_i^t L_i. \tag{3}$$

According to the principle of mass conservation, the change in the number of vehicles is given by the net influx minus the net outflux. For node i, the influx comes from its upstream neighbors $j \in \mathcal{N}_{\text{up}}(i)$. Let ω_{ji} be the routing ratio of flow from node j that splits into node i. Since our model directly predicts the traffic flow $\hat{q}$, we can explicitly formulate the discrete mass balance equation as:

$$\frac{m_i^{t+1} - m_i^t}{\Delta t} = \sum_{j \in \mathcal{N}_{\text{up}}(i)} \omega_{ji} \hat{q}_j^t - \hat{q}_i^t + s_i^t, \tag{4}$$

where s_i^t represents unobserved sources/sinks (e.g., on/off-ramps). Under the closed-mainline assumption, we take $s_i^t \approx 0$. Substituting $m_i^t = c_i \hat{O}_i^t L_i$ gives:

$$\frac{c_i \hat{O}_i^{t+1} L_i - c_i \hat{O}_i^t L_i}{\Delta t} = \sum_{j \in \mathcal{N}_{\text{up}}(i)} \omega_{ji} \hat{q}_j^t - \hat{q}_i^t. \tag{5}$$

Dividing both sides by L_i and rearranging terms yields the physical residual used in our loss function:

$$R_{\text{phy}}^{(q)}(i,t) = c_i \frac{\hat{O}_i^{t+1} - \hat{O}_i^t}{\Delta t} + \frac{\hat{q}_i^t - \sum_{j \in \mathcal{N}_{\text{up}}(i)} \omega_{ji} \hat{q}_j^t}{L_i}. \tag{6}$$

This formulation can be viewed as a graph-topology generalized first-order upwind difference scheme, bridging one-dimensional continuous traffic physics and arbitrary directed network topologies.

A.2 Step-wise Comparison with STGCN-Multi

Table 4 reports the step-wise prediction errors of STGCN-Multi and PI-Dir-STGCN. The results confirm that the physics regularizer provides additional long-horizon stability beyond the multi-variate inputs.

A.3 Additional Prediction Trend Results on PeMS08

Figure 5 presents the MAE and RMSE trends on PeMS08 over prediction horizons, complementing the PeMS04 results in the main text. The same pattern is observed: STGCN exhibits rapidly increasing errors beyond 30 minutes, while our model maintains significantly lower and more stable errors throughout the entire 60-minute horizon.

A.4 Additional Qualitative Results

Figure 6 presents the 60-minute absolute error distribution across all 307 nodes. STGCN shows strong spatio-temporal error clustering in peak hours and congested segments, reflecting unconstrained error propagation in pure data-driven models. In contrast, the proposed model achieves uniformly low errors with only isolated local outliers, demonstrating that LWR mass conservation regularization serves as a strong global inductive bias to block error propagation and suppress long-term accumulation.

Table 4: Step-wise prediction errors (15/30/45/60 min) of STGCN-Multi and PI-Dir-STGCN on PeMS04 and PeMS08.

Dataset	Horizon	Model	MAE	RMSE	MAPE
PeMS04	15 min	STGCN-Multi	20.36	31.87	12.73%
		PI-Dir-STGCN	20.23	31.72	12.73%
	30 min	STGCN-Multi	23.39	35.93	14.21%
		PI-Dir-STGCN	23.00	35.45	14.18%
	45 min	STGCN-Multi	26.45	40.07	15.84%
		PI-Dir-STGCN	25.75	39.17	15.68%
	60 min	STGCN-Multi	29.54	44.05	17.59%
		PI-Dir-STGCN	28.35	42.62	17.17%
PeMS08	15 min	STGCN-Multi	15.96	24.68	9.93%
		PI-Dir-STGCN	15.89	24.53	9.90%
	30 min	STGCN-Multi	18.61	28.59	11.45%
		PI-Dir-STGCN	18.44	28.23	11.39%
	45 min	STGCN-Multi	21.30	32.64	12.96%
		PI-Dir-STGCN	20.90	31.74	12.74%
	60 min	STGCN-Multi	23.82	36.31	14.31%
		PI-Dir-STGCN	23.24	35.01	14.04%

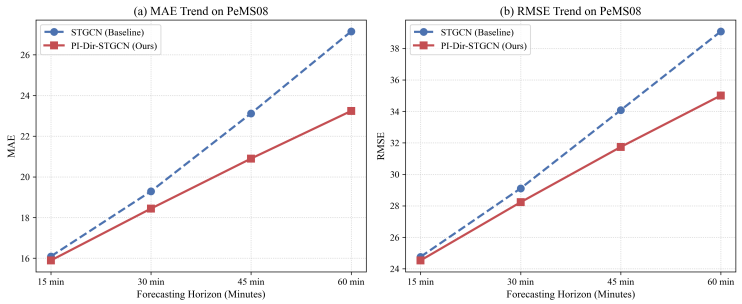

Figure 5: MAE and RMSE trends on PeMS08 over prediction horizons.

A.5 Hyperparameter Sensitivity Analysis

To verify the rationality of our chosen $\lambda_{\text{phy}} = 5 \times 10^{-4}$, we test $\lambda_{\text{phy}} = \{0, 5 \times 10^{-6}, 10^{-5}, 10^{-4}, 5 \times 10^{-4}\}$ on PeMS04 and PeMS08, using a fixed random seed to trace the exact evolution trend, avoiding the confounding effect of random initialization. Figure 7 reports the overall 1-hour MAE results.

When $\lambda_{\text{phy}} = 0$, the physics constraint is completely disabled, and the model degenerates to the directed-graph version Dir-only, yielding the largest errors. As λ_{phy} increases to 5×10^{-4}, the MAE reaches its minimum (23.36 on PeMS04 and 18.73 on PeMS08). Both increasing or decreasing λ_{phy} from this optimal point leads to higher errors. These results confirm that a moderate physical constraint best improves prediction accuracy, validating the robustness of $\lambda_{\text{phy}} = 5 \times 10^{-4}$ across both datasets.

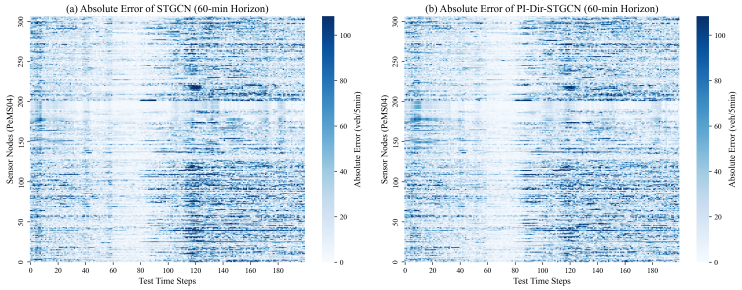

Figure 6: Absolute error distribution of 60-minute prediction over the whole network. (a) STGCN; (b) PI-Dir-STGCN.

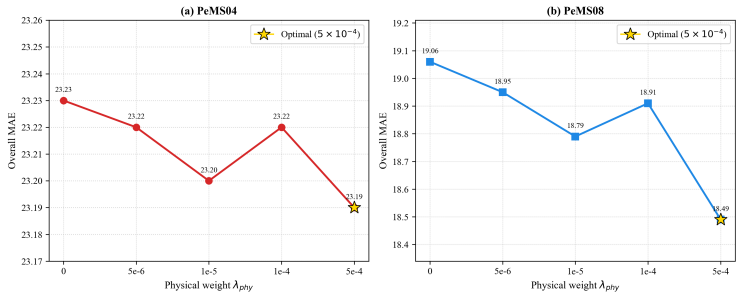

Figure 7: Overall MAE under different λ_{phy} on PeMS04 and PeMS08.

A.6 Detailed Model Complexity and Efficiency Analysis

Table 5: Model complexity comparison.

Model	Params (M)	Training time (50 epochs)		Inference (ms/sample)
		PeMS04	PeMS08	
STGCN	0.48	2023.92 s	1410.61 s	1.20
Ours	0.51	2879.15 s	2171.3 s	1.22

Note: Inference latency is the average forward time per sample per step, measured on the test set with batch prediction. Training time is the sum of per-epoch durations.

To evaluate the computational cost of our model, we compare the parameter count, training time, and single-step inference latency of standard STGCN and PI-Dir-STGCN on PeMS04 and PeMS08. All experiments are conducted in the same environment with 50 training epochs and identical optimizer, learning rate, and batch size settings. Standard STGCN relies solely on univariate flow as input, while our PI-Dir-STGCN processes three-channel multivariate input (flow, occupancy, speed) and continuously computes the physics-informed regularization during backpropagation. This expanded computational graph increases per-epoch overhead, which directly accounts for the difference in total training time, as summarized in Table 5.