Computer Modeling in
Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.086216

Check for updates

ARTICLE

A Physics-Informed Spatial-Temporal Graph Attention Model for Traffic Forecasting and Interpretable Congestion Propagation Analysis

Yan-Wei Li and David Chunhu Li*

Information Technology and Management Program, Ming Chuan University, Taoyuan City, Taiwan

*Corresponding Author: David Chunhu Li. Email: davidli@mail.mcu.edu.tw

Received: 26 May 2026; Accepted: 22 July 2026; Published: 28 August 2026

ABSTRACT: As urban transportation systems grow increasingly complex, accurate and interpretable traffic congestion forecasting is critical. While existing deep learning models utilize graph neural networks (GNNs) and attention mechanisms, they often struggle with physical consistency under extreme scenarios. To address this, we propose the Physics-Informed Explainable Spatial-Temporal Graph Attention Network (PI-X-STGAT). Our framework models road segments as graph nodes, integrating traffic, weather, and cyclical temporal features. The architecture comprises a Context-Aware Graph Attention Network (GAT) enhanced with Node Adaptive Parameter Learning (NAPL) for capturing dynamic spatial dependencies, a Gated Recurrent Unit (GRU) layer for temporal evolution, and a physics-inspired regularization term ($\mathcal{L}_{phy}$) acting as a surrogate conservation constraint. This soft regularizer mitigates physically illogical predictions without strictly enforcing partial differential equations. Additionally, an integrated Explainable Artificial Intelligence (XAI) module extracts attention matrices to provide supportive attention-based influence analysis cues. Extensive experiments on a real-world dataset demonstrate that PI-X-STGAT exhibits competitive overall performance, achieving the lowest Jam Mean Absolute Error (MAE) among the evaluated baselines in our experimental setting while maintaining highly stable multi-step horizon forecasting. Synthetic boundary-case analyses indicate enhanced stability against extreme Out-of-Distribution (OoD) perturbations. Furthermore, the extracted attention-guided metrics offer dynamic attention-based influence analysis capabilities, providing a promising, interpretable framework for intelligent transportation decision-support.

KEYWORDS: Intelligent transportation systems; traffic forecasting; spatial-temporal graph neural network; physics-guided deep learning; Explainable AI (XAI); out-of-distribution robustness

1 Introduction

1.1 Research Background

With the rapid acceleration of urbanization, traffic congestion has emerged as a severe challenge for megacities worldwide. As the core of modern urban infrastructure, Intelligent Transportation Systems (ITS) rely heavily on accurate traffic flow forecasting for efficient road network management and sustainable urban mobility. In recent years, deep learning technologies, particularly Spatial-Temporal Graph Neural Networks (STGNNs), have achieved breakthrough progress in this domain. By abstracting urban road networks into graph structures, traditional models such as Spatio-Temporal Graph Convolutional Networks (STGCN), Diffusion Convolutional Recurrent Neural Networks (DCRNN), and attention-based spatial-temporal graph convolutional network (ASTGCN) can effectively capture complex spatio-temporal dependencies from historical traffic data, demonstrating remarkable forecasting accuracy under normal traffic conditions.

1.2 Research Motivation

Despite the academic success of existing purely data-driven deep learning models, their real-world deployment in Traffic Management Centers (TMCs) still faces three critical bottlenecks:

- **Vulnerability in Out-of-Distribution (OoD) Scenarios and Lack of Physical Consistency:** Existing models tend to overfit historical data distributions. When facing OoD events like extreme weather or sudden accidents, purely data-driven models often generate predictions that violate the macroscopic laws of traffic physics. For instance, a model might predict spontaneous downstream congestion even when an upstream accident has completely cut off the traffic flow.
- **The Opaque Nature of Deep Learning and Lack of Explainability:** The “black-box” nature of deep neural networks diminishes the transparency of predictions. When the system issues a severe congestion alert, traffic operators cannot deduce the specific correlated upstream indicators of the congestion, leading to a lack of trust in the system’s outputs.
- **The Gap Between Prediction and Traceability:** Conventional research predominantly stops at the passive prediction of future speeds or traffic states, lacking the ability to dynamically trace the origin of anomalies. A truly intelligent transportation system should not merely forecast where congestion will occur, but actively assist managers by analyzing the directional influence of kinematic waves to enable early warnings.

1.3 Research Problem & Contributions

To address the aforementioned pain points, this study proposes a novel framework: the Physics-Informed eXplainable Spatial-Temporal Graph Attention Network (PI-X-STGAT). This research innovatively introduces the macroscopic law of conservation of mass in traffic flow as a regularization constraint into the neural network architecture, while integrating eXplainable AI (XAI) techniques. This establishes an end-to-end intelligent traffic decision-making system that seamlessly bridges forecasting and influence analysis. The three core contributions of this study are summarized as follows:

1. **Physics-Guided Regularization:** We formulate a surrogate physics-inspired loss function ($\mathcal{L}_{phy}$) based on the conceptual conservation of mass. By penalizing anomalous net flow variations during training, the model’s forecasting stability is enhanced, particularly under synthetic extreme perturbations.
2. **Adaptive Context-Aware Graph Attention:** We construct a novel spatial layer that not only integrates external multi-modal features into the attention mechanism but also dynamically learns hidden spatial dependencies through Node Adaptive Parameter Learning (NAPL). This dual-graph approach generates robust, context-driven topological weights.
3. **Dynamic Attention-Based Influence Analysis via XAI:** By extracting XAI spatial attention matrices, the model highlights influential upstream bottleneck nodes in real-time. This explainable topological cue provides traffic operators with actionable early-warnings for proactive congestion mitigation.

2 Related Work

The core framework proposed in this study, PI-X-STGAT, is designed to address three major research gaps in the current fields of Intelligent Transportation Systems (ITS) and Spatial-Temporal Graph Neural Networks (STGNNs).

2.1 Spatial-Temporal Graph Neural Networks and the Lack of Physics Constraints

In recent years, traffic forecasting has transitioned from traditional statistical methods to deep learning paradigms, with Spatial-Temporal Graph Neural Networks (ST-GNNs) becoming a mainstream approach.

Notable architectures include DCRNN [1], which models spatial dependencies through bidirectional random walks on directed graphs; STGCN [2], which employs fully convolutional structures to improve training efficiency with relatively few parameters; and T-GCN [3], which integrates graph convolutional networks (GCNs) with Gated Recurrent Units (GRUs) to simultaneously capture spatial topology and dynamic temporal changes.

Furthermore, models such as ASTGCN [4] utilize spatial-temporal attention mechanisms to dynamically capture periodic features, while Graph WaveNet [5] combines dilated causal convolutions with adaptive adjacency matrices to model long-range sequences without relying exclusively on predefined graph structures.

To further reduce reliance on predefined graphs, purely data-driven models such as AGCRN [6] employ Node Adaptive Parameter Learning (NAPL) and Data Adaptive Graph Generation (DAGG) to capture hidden dependencies directly from traffic data. However, although such models can achieve high forecasting accuracy, they do not explicitly impose macroscopic physical constraints, such as traffic-flow conservation, on their predictions.

Beyond single-graph and adaptive-adjacency approaches, Promsawat et al. [7] proposed a Dynamic Multi-Graph Spatio-Temporal Graph Neural Ordinary Differential Equation Network (DMST-GNODE). Their framework constructs distance, pattern, and dynamic graphs to represent geographical proximity, similarity in historical traffic patterns, and time-varying latent relationships, respectively. These graph representations are integrated with temporal convolutional components and a graph neural ordinary differential equation module to capture complementary spatial-temporal dependencies and continuous hidden-state evolution. The model was evaluated using traffic datasets from Bangkok, PeMS08, and Los_Loop over multiple forecasting horizons. Although this multi-graph formulation improves the representation of complex and evolving spatial relationships, it remains primarily data-driven and does not explicitly introduce a traffic-flow conservation residual into the training objective or provide edge-level attention-based influence analysis.

To address the need for physical consistency, models such as CAP-STGCN [8] couple macroscopic kinematics with a congestion-aware mechanism to reduce physically implausible predictions under severe congestion and long-tail traffic conditions.

Similarly, PG-STGNN [9] incorporates traffic-engineering principles and intersection-level information, including queue formation, signal timing, travel-time functions, and the volume-to-capacity ratio, into a stepwise spatio-temporal forecasting framework. However, it does not formulate the same surrogate residual-based physics loss adopted in the present study.

Other physics-guided frameworks include STDEN [10], which represents traffic flow as a dynamic process driven by a latent potential-energy field using differential equations; PI-GRNN [11], which utilizes deep learning as a state-space model within a Kalman-filter mixture to reduce epistemic uncertainty and model drift; and GNN-based approaches that apply kinematic-wave theory to construct wave-informed anisotropic temporal graphs for directional traffic-wave modeling [12].

Additionally, Physics-Informed Neural Networks (PINNs) [13] provide a general mathematical foundation for incorporating nonlinear partial differential equations into neural-network training through automatic differentiation.

Despite these valuable efforts, many existing approaches focus primarily on static topology, adaptive graph learning, feature-level physical information, or continuous latent dynamics. Relatively few studies jointly combine adaptive spatial dependency learning with an explicit surrogate traffic-flow consistency regularizer and evaluate the resulting model under event-specific conditions such as localized traffic disruptions and holiday-related distribution shifts.

2.2 Rigidity in Context-Aware Fusion

Traffic mobility is highly situation-dependent and is significantly influenced by external and multimodal contexts, such as points of interest (POIs), land use, weather conditions, holidays, and temporal periodicity. However, conventional ST-GNNs may exhibit limited flexibility in context-aware fusion. In particular, models that rely mainly on static distance-based adjacency matrices may fail to represent hidden spatial relationships or changes associated with dynamic environmental conditions, thereby limiting their ability to model complex urban traffic patterns.

To overcome this structural rigidity, recent research has emphasized adaptive and hierarchical context fusion. The ASTAM architecture [14] addresses static graph limitations by utilizing parallel static adaptive graphs and dynamic graph attention networks, allowing the model to capture hidden and evolving spatial relationships. To integrate diverse contextual modalities, the MCGCN model [15] introduces a hierarchical spatial-embedding module that organizes POIs, road networks, and land-use information, together with an attention-based multimodal layer that fuses these representations with traffic-speed observations.

From a heterogeneous data-fusion perspective, Wang and Susanto [16] proposed an attention-enhanced Conv-BiLSTM framework that integrates historical traffic flow with weather conditions, holiday information, vehicle-type-specific traffic data, and multiple temporal periodicities. Their framework separately models current, daily, weekly, and monthly traffic patterns and uses attention mechanisms to assign different levels of importance to the learned temporal representations. Experiments conducted using traffic records from eight gantries on Taiwan National Freeway No. 3 demonstrated the value of incorporating heterogeneous contextual and periodic information into traffic forecasting. Nevertheless, the framework does not explicitly represent road-network connectivity between traffic nodes and was evaluated within a single freeway corridor, leaving graph-topology modeling, scalability, and cross-region generalization as areas for further investigation.

Furthermore, context awareness has been extended to broader mobility and traffic-management applications. For instance, the DeepSTN+ framework [17] incorporates complex region-level contextual dependencies to predict large-scale crowd flows in metropolitan areas. At a more granular level, the Fuzzy-macro long short-term memory (LSTM) model [18] uses an adaptive fuzzy-inference system to fuse macroscopic driving contexts with individual vehicle kinematics for dangerous-driving behavior assessment.

Collectively, these studies demonstrate that heterogeneous contextual information and adaptive spatial representations can improve the modeling of traffic dynamics. However, contextual feature fusion alone does not ensure that the resulting predictions remain consistent with traffic-flow principles. Moreover, the learned spatial relationships are not always accompanied by directly accessible edge-level indicators for examining model-internal spatial influence. These limitations motivate the integration of heterogeneous contextual features, adaptive graph attention, and surrogate physics-guided regularization within the proposed PI-X-STGAT framework.

2.3 Explainability and Dynamic Attention-Based Influence Analysis

While deep learning models achieve high predictive accuracy, their "black-box" nature poses significant challenges for deployment in critical intelligent transportation systems where decision accountability is essential. A taxonomic survey of GNN explainability by Yuan et al. [19] categorizes current methods into instance-level and model-level explanations. At the instance level, tools like GNNExplainer [20] identify a compact subgraph structure and a small subset of crucial node features that maximize mutual information with the GNN's prediction, providing human-intelligible insights.

In the specific realm of spatial-temporal modeling, frameworks like STExplainer [21] leverage the Graph Information Bottleneck (GIB) principle to evaluate fidelity and sparsity, identifying influential subgraphs for traffic predictions. However, these post-hoc explainers remain entirely data-driven and lack physical traffic flow constraints to theoretically ground their attention mechanisms.

In the specific context of mobile traffic forecasting, Explainable AI (XAI) has been uniquely leveraged for dynamic attention-based influence analysis and vulnerability analysis. The DeExp framework [22] utilizes XAI techniques, such as Layer-wise Relevance Propagation (LRP) and Gradient-weighted Class Activation Mapping (Grad-CAM), to dynamically analyze the influence of predictions across the spatio-temporal domain. This allows operators to spot influential base stations and evaluate how targeted adversarial traffic perturbations at these identified sources can severely disrupt overall forecasting accuracy.

Recently, Large Language Models (LLMs) have been introduced to further enhance explainability and dynamic attention-based influence analysis. Frameworks like xTP-LLM [23] convert multi-modal spatial-temporal data and external factors into structured natural language prompts. By employing chain-of-thought (CoT) reasoning, xTP-LLM explicitly analyzes the impact of dynamic external events (e.g., holidays, weather) and spatial attributes step-by-step. This LLM-based paradigm not only maintains competitive predictive accuracy but also outputs intuitive, language-based explanations, effectively identifying the underlying correlated indicators of traffic fluctuations and moving traffic forecasting from a black box to a transparent, reliable system.

Differentiation of PI-X-STGAT: In contrast to the aforementioned studies—which largely diverge into isolated pursuits of adaptive graph learning, physics-informed feature embeddings, or data-driven explainability—our PI-X-STGAT proposes a unified framework. It uniquely synergizes Context-Aware GAT and NAPL-style dynamic embeddings with a soft surrogate flow conservation loss inspired by macroscopic kinematic wave theory. This integration not only achieves high predictive accuracy but also empowers the global attention matrices to conduct physically-grounded attention-based influence analysis for congestion. Consequently, unlike purely data-driven baselines, the proposed framework demonstrates rigorous robustness and generalizability under out-of-distribution (OoD) evaluations. To clearly illustrate this differentiation, Table 1 summarizes the capabilities and limitations of recent representative frameworks against our proposed model.

Table 1: Comparison of PI-X-STGAT with closest related studies.

Method	Adaptive Graph Learning	Physics Constraint	Explainability	OoD Analysis	Traffic Forecasting
CAP-STGCN [8]	ASTGCN backbone with dynamic spatial weighting.	Use fundamental diagram ($q = k \times v$) and physical consistency check	None	Discusses long-tail congestion generalization, but does not explicitly emphasize OoD analysis.	Predicts next 12 steps on PeMS04/08 with MAE/RMSE
PG-STGNN [9]	Topology-based ST-GNN for signalized intersection networks; no clear learnable adaptive adjacency mechanism	Incorporates traffic flow principles and intersection-level features such as queue formation, signal timing, travel time function, and v/c ratio, but not a normalized residual-based physics loss.	Provides domain knowledge interpretability through traffic engineering features, but does not explicitly focus on fidelity, sparsity, or attribution analysis.	Limited scenario-level evaluation; no rigorous OoD tests on incidents, weather shifts, unseen intersections, or cross city transfer.	Conducts short-term traffic flow prediction using real-world data from a signalized intersection in Beijing Yizhuang District.

(Continued)

Table 1 (continued)

Method	Adaptive Graph Learning	Physics Constraint	Explainability	OoD Analysis	Traffic Forecasting
STExplainer [21]	STG attention encoder/decoder + Graph Information Bottleneck.	None.	Identifies important subgraphs via sparsity and fidelity evaluation.	Robustness analysis primarily for missing and corrupted data.	Traffic prediction on PeMS04/07/08 focusing on combined explainability.
AGCRN [6]	NAPL + DAGG + GRU; learns hidden dynamics directly from data.	Completely data driven; no physical constraints (e.g., mass conservation or LWR).	None.	No cross-domain transfer, severe incident, or domain shift testing.	Multi-step prediction on PeMSD4/PeMSD8; compared with DCRNN, ASTGCN, etc.
PI-X-STGAT (Ours)	Context-Aware GAT + GRU with NAPL style node embeddings integrated into physical topology.	Employs the LWR mass conservation concept to design a surrogate flow conservation loss.	Preserves attention matrices for source/target influence analysis and attention-based influence analysis.	Features dedicated local shock stress testing (synthetic accidents) and a real world Labor Day holiday domain shift	Uses LA HERE Maps API data (18-dim features) to predict Jam/Speed; achieves best Jam MAE

3 System Model and Algorithms

To facilitate clear mathematical formulation throughout this section, the primary notations used in this paper are summarized in Table 2.

Table 2: Summary of key mathematical notations.

Notation	Description
$\mathcal{G}, \mathcal{V}, \mathcal{E}$	Traffic road network graph, set of nodes, and set of edges
N, F, C	Number of nodes, input feature dimension, and output channels
t, L, τ	Current time step, historical sequence length, and prediction horizon
$\mathbf{X}^{(t)}$	Input multi-modal feature matrix of all nodes at time t
$\hat{\mathbf{Y}}^{(t+\tau)}$	Predicted output matrix (Jam Factor and Speed) at time $t+\tau$
$\mathbf{A}, \mathbf{A}_{adp}$	Static physical adjacency matrix and adaptive adjacency matrix
$\mathbf{W}_{(\cdot)}, \mathbf{U}_{(\cdot)}$	Learnable weight matrices in linear projections and GRU layers
$\mathbf{E}_1, \mathbf{E}_2$	Learnable node embedding dictionaries for NAPL ($N \times d$)
$\alpha_{ij}^{(t)}, \mathbf{W}_{attn}^{(t)}$	Context-fused attention weight and dynamic influence matrix
$\mathbf{h}_i^{(t)}, \mathbf{c}_i^{(t)}$	Hidden representation and context representation of node i
$\mathbf{z}_i^{(t)}$	Aggregated spatial representation of node i
$\hat{q}_i, \mathbf{Q}_{in}, \mathbf{Q}_{out}$	Surrogate flow, network inflow vector, and outflow vector
J_{max}	Maximum threshold of the Jam Factor (e.g., 10.0)
$\sigma(\cdot), \odot$	Sigmoid activation function and Hadamard (element-wise) product

3.1 System Architecture

This study proposes a congestion forecasting framework that integrates spatial correlations, temporal dynamics, traffic physics, and interpretability analysis, named PI-X-STGAT (Physics-Informed Explainable Spatial-Temporal Graph Attention Network). The proposed model primarily consists of four modules: a

Context-Aware Graph Attention Layer, a Temporal GRU Layer, a Physics Loss Check module, and an XAI Interpretability Module. These components are designed to simultaneously enhance the accuracy, physical rationality, and transparency of traffic congestion forecasting.

First, road segments are conceptualized as nodes within a graph structure, and a traffic topology graph is constructed based on spatial proximity. Subsequently, traffic states, road attributes, temporal cyclical features, and meteorological conditions at each timestamp are integrated into node input features. These features are fed into the Context-Aware GAT to extract the dynamic spatial dependencies among nodes. Next, the spatial representations of each node are fed into the GRU along the temporal dimension to learn the evolutionary characteristics of traffic congestion over time. The model outputs the predicted Jam Factor and Speed for future timestamps. During the training phase, traffic physics constraints are incorporated to encourage the data-driven model to produce predictions that are more consistent with macroscopic traffic-flow conservation tendencies. Finally, the model retains the global attention matrix and feature sensitivity information to support subsequent decision-making analysis.

Rather than constructing an overly complex graph architecture, PI-X-STGAT is deliberately designed with a lightweight and highly efficient backbone (leveraging established GAT and GRU operations). This streamlined design minimizes the risk of structural overfitting in OoD scenarios and ensures that the parameter updates remain highly sensitive to the guidance of the physics-informed regularizer, thereby striking an optimal balance between predictive accuracy and real-time computational deployment feasibility.

The overall schematic of the proposed framework is illustrated in Fig. 1. The system integrates multi-modal historical features into a hierarchical architecture designed to balance predictive accuracy with physical consistency. As shown in the diagram, the input sequence is processed through sequential context-aware spatial aggregation and temporal recurrent updates, followed by a physics-informed feedback loop during training to ensure the predictions adhere to macroscopic conservation laws.

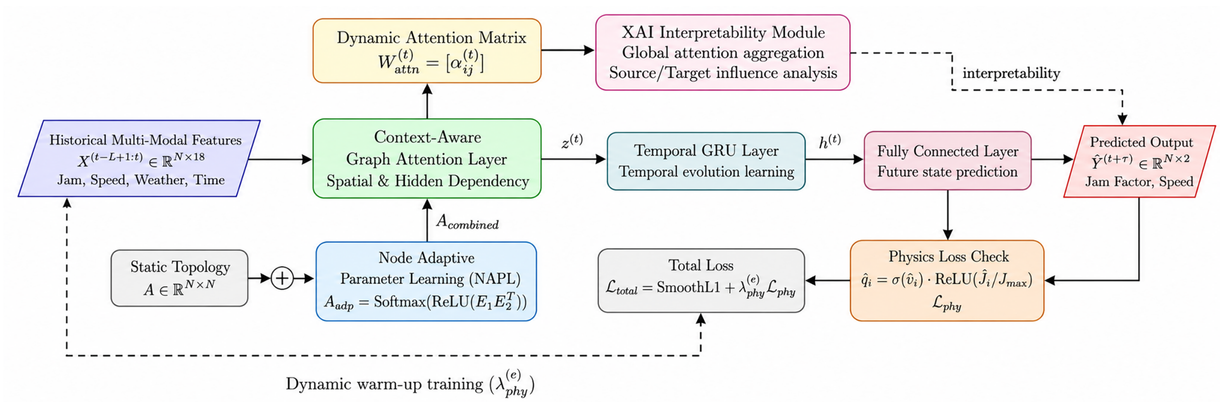

Figure 1: Detailed architectural layout of the PI-X-STGAT model, showcasing the information flow from multi-modal input processing to context-aware spatial-temporal encoding and the integrated physics-informed loss regularization.

3.2 PI-X-STGAT Model Construction

Let the traffic road network be represented as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$ is the set of road segment nodes, and $\mathcal{E}$ represents the connectivity between nodes. At time t, the input feature vector of node v_i is denoted as $\mathbf{x}_i^{(t)} \in \mathbb{R}^F$, which encompasses features such as Jam Factor, speed, geographical coordinates, meteorological variables, and temporal cyclical encodings. By stacking the features of all nodes at time t, we obtain:

$$\mathbf{X}^{(t)} = \left[\mathbf{x}_1^{(t)}, \mathbf{x}_2^{(t)}, \ldots, \mathbf{x}_N^{(t)}\right]^\top \in \mathbb{R}^{N\times F} \tag{1}$$

Given a historical observation sequence of length L, denoted as $\left\{\mathbf{X}^{(t-L+1)}, \ldots, \mathbf{X}^{(t)}\right\}$, the objective of the model is to predict the traffic states at the future time step $t+\tau$:

$$\hat{\mathbf{Y}}^{(t+\tau)} \in \mathbb{R}^{N\times C} \tag{2}$$

In this study, we set $C = 2$, meaning the model predicts both the Jam Factor and the speed simultaneously. In the spatial modeling module, node features are first projected into a hidden space and a context space, respectively:

$$\mathbf{h}_i^{(t)} = \mathbf{W}_h \mathbf{x}_i^{(t)}, \quad \mathbf{c}_i^{(t)} = \mathbf{W}_c \mathbf{x}_i^{(t)} \tag{3}$$

To capture hidden spatial correlations that are not reflected in the predefined physical road network, we introduce Node Adaptive Parameter Learning (NAPL). We define two learnable node embedding dictionaries, $\mathbf{E}_1, \mathbf{E}_2 \in \mathbb{R}^{N\times d}$, where d is the embedding dimension. The adaptive adjacency matrix is generated via:

$$\mathbf{A}_{adp} = \mathrm{Softmax}(\mathrm{ReLU}(\mathbf{E}_1\mathbf{E}_2^\top)) \tag{4}$$

This learned hidden topology is then combined with the static physical adjacency matrix $\mathbf{A}$ to form the combined topological weight $\mathbf{A}_{combined} = \mathbf{A} + \mathbf{A}_{adp}$. Consequently, the context-fused attention score for any given nodes i and j is calculated as follows:

$$e_{ij}^{(t)} = \mathrm{LeakyReLU}\left(\mathbf{a}_s^\top \mathbf{h}_i^{(t)} + \mathbf{a}_d^\top \mathbf{h}_j^{(t)} + \beta \mathbf{c}_i^{(t)\top} \mathbf{c}_j^{(t)} + \log(1 + A_{combined,ij})\right) \tag{5}$$

where $\mathbf{a}_s$ and $\mathbf{a}_d$ are the attention parameter vectors, and β is the weight coefficient for the context term. A softmax function is then applied to normalize these scores into attention weights:

$$\alpha_{ij}^{(t)} = \frac{\exp\left(e_{ij}^{(t)}\right)}{\sum_{k\in\mathcal{N}(i)} \exp\left(e_{ik}^{(t)}\right)} \tag{6}$$

These normalized scores $\alpha_{ij}^{(t)}$ form a dynamic node influence matrix, $\mathbf{W}_{attn}^{(t)} = [\alpha_{ij}^{(t)}] \in \mathbb{R}^{N\times N}$. This attention-based expression is carefully formulated to dynamically trace and weight the influence of neighboring nodes based on both real-time traffic states and context representations. By explicitly calculating the directional interaction between outbound node j and inbound node i, this matrix serves as the rigorous mathematical foundation for our subsequent dynamic influence analysis and XAI analysis. Weighted aggregation is then performed to obtain the spatial representation of the node:

$$\mathbf{z}_i^{(t)} = \sum_{j\in\mathcal{N}(i)} \alpha_{ij}^{(t)} \mathbf{h}_j^{(t)} \tag{7}$$

For temporal modeling, the sequence of spatial representations is fed into a GRU to learn the temporal dependencies of congestion evolution. The update process is formulated as:

$$\mathbf{r}_i^{(t)} = \sigma\left(\mathbf{W}_r \mathbf{z}_i^{(t)} + \mathbf{U}_r \mathbf{h}_i^{(t-1)}\right) \tag{8}$$

$$\mathbf{u}_i^{(t)} = \sigma\left(\mathbf{W}_u \mathbf{z}_i^{(t)} + \mathbf{U}_u \mathbf{h}_i^{(t-1)}\right) \tag{9}$$

$$\tilde{\mathbf{h}}_i^{(t)} = \tanh\left(\mathbf{W}_h\mathbf{z}_i^{(t)} + \mathbf{U}_h\left(\mathbf{r}_i^{(t)} \odot \mathbf{h}_i^{(t-1)}\right)\right) \tag{10}$$

$$\mathbf{h}_i^{(t)} = \left(1 - \mathbf{u}_i^{(t)}\right) \odot \mathbf{h}_i^{(t-1)} + \mathbf{u}_i^{(t)} \odot \tilde{\mathbf{h}}_i^{(t)} \tag{11}$$

where $\mathbf{r}_i^{(t)}$ and $\mathbf{u}_i^{(t)}$ denote the reset and update gates, respectively. $\sigma(\cdot)$ represents the sigmoid activation function, and $\odot$ denotes the Hadamard (element-wise) product. Finally, the future predictions are generated through a fully connected layer: $\hat{\mathbf{y}}_i^{(t+\tau)} = \text{FC}(\mathbf{h}_i^{(t)})$.

3.3 Physics-Guided Loss Function Design

Theoretical Justification from Macroscopic Traffic Flow: In classical macroscopic traffic flow theory, particularly the Lighthill-Whitham-Richards (LWR) kinematic wave model, traffic dynamics are governed by the conservation of mass. The one-dimensional continuity equation is defined as $\frac{\partial\rho}{\partial t} + \frac{\partial q}{\partial x} = 0$, where ρ is the traffic density and q is the traffic flow. When discretizing this continuous partial differential equation (PDE) into a graph-based network, the change in density at a specific node over time ($\Delta\rho_i$) should be proportionally balanced by the net spatial flow transitioning through that node ($Q_{in} - Q_{out}$).

However, directly applying the strict LWR PDE to deep learning models for complex urban networks is mathematically intractable due to the presence of unobservable sinks/sources (e.g., parking lots, small alleys) and dynamic capacity limits. Therefore, instead of enforcing a rigid PDE, we formulate a *surrogate conservation constraint* acting as a soft physics-guided regularizer.

It is crucial to clarify that $\hat{q}_i$ in our model is utilized as a conservation-aware surrogate representation for spatial regularization, rather than a calibrated true physical flow estimator. Due to the absence of absolute density in API-sourced data, we approximate the vehicular density using the normalized Jam Factor as a practical proxy. This surrogate formulation reasonably reflects the macroscopic conservation principle under free-flow to moderate congestion regimes, where the relationship between the Jam Factor and true density remains monotonically consistent. However, we explicitly acknowledge that this surrogate relationship may deviate during extreme gridlock or transient shockwaves. Under such conditions, phenomena like capacity drop occur, and the linear proxy assumption breaks down as speed approaches zero while the Jam Factor saturates. Recognizing these physical limitations, we do not enforce this surrogate mass-balance as a hard physical equation. Instead, it is deployed strictly as a soft regularization term $\mathcal{L}_{phy}$ acting as a spatial inductive bias. By penalizing physically implausible divergences between adjacent nodes, it guides the model toward spatially coordinated predictions without falsely enforcing strict vehicular conservation. The surrogate flow is thus calculated as:

$$\hat{q}_i = \sigma(\hat{v}_i) \cdot \text{ReLU}\left(\frac{\hat{J}_i}{J_{max}}\right) \tag{12}$$

where $\hat{v}_i$ is the predicted speed and $\sigma(\cdot)$ bounds the speed representation. Furthermore, the graph adjacency matrix $\mathbf{A}$ is leveraged to calculate the surrogate outflow and inflow for each node ($\mathbf{Q}_{out} = \mathbf{A}\hat{\mathbf{q}}$, $\mathbf{Q}_{in} = \mathbf{A}^\top\hat{\mathbf{q}}$). We define the temporal change in congestion as $\Delta\hat{\mathbf{J}} = \hat{\mathbf{J}}^{(t+\tau)} - \mathbf{J}^{(t)}$. Consequently, the physics-guided regularization term is defined as:

$$\mathcal{L}_{phy} = \text{Mean}\left(\left|(\mathbf{Q}_{in} - \mathbf{Q}_{out}) - \Delta\hat{\mathbf{J}}\right|\right) \tag{13}$$

This equation conceptually reflects the macroscopic mass conservation, indicating that the temporal evolution of congestion should correspond consistently with the spatial net flow. To balance data-fitting capability and physical rationality while maintaining robustness against extreme outliers, we replace the standard mean square error (MSE) with the Smooth L1 Loss (Huber Loss). The final total loss function is integrated as:

$$\mathcal{L}_{total} = \text{SmoothL1}(\hat{\mathbf{Y}}, \mathbf{Y}_{true}) + \lambda_{phy}^{(e)}\mathcal{L}_{phy} \tag{14}$$

where $\lambda_{phy}^{(e)}$ is a dynamic physics loss weight at training epoch e, linearly warmed up from 0 to its optimal value (e.g., 0.015) to prevent premature regularization before the network captures the base data distribution.

3.4 XAI Interpretability Module

To obtain stable interpretation results, the attention matrices generated across all time steps during the testing phase are averaged:

$$\bar{\mathbf{W}}_{attn} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{W}_{attn}^{(t)} \tag{15}$$

This averaged attention matrix serves as the core basis for identifying traffic congestion propagation relationships. To quantify the influential role of each road segment within the overall network, the outbound influence score and inbound influence score of a node are defined as follows:

$$S_j^{src} = \sum_{i=1}^{N}\bar{\alpha}_{ij} \tag{16}$$

$$S_i^{tgt} = \sum_{j=1}^{N}\bar{\alpha}_{ij} \tag{17}$$

Limitations of Attention-based Explanation: It is imperative to establish the theoretical boundaries of the proposed XAI module. While the attention mechanism provides intuitive propagation cues, the extracted weights ($\alpha_{ij}^{(t)}$) inherently represent learned statistical correlations within the latent space, not a definitive Directed Acyclic Graph (DAG) of causality. Attention does not equal causality. Therefore, the terminology "attention-guided candidate source identification" utilized in this study should be strictly interpreted as "identifying high-correlation candidate nodes for heuristic evaluation," rather than proving strict physical causality (e.g., Granger causality). The provided topological metrics act as supplementary decision-support references rather than deterministic causal derivations.

3.5 Algorithm and Implementation Details

The end-to-end training logic and the subsequent explainable feature extraction process are formalized in Algorithm 1. The procedure ensures that the model not only optimizes for data-driven accuracy via Smooth L1 loss but also converges toward physical rationality by minimizing the net flow residual. Upon completion of training, the global attention matrix is calculated to provide the foundation for subsequent influence analysis.

Algorithm 1: Operational workflow for the joint training and XAI attention extraction of the PI-X-STGAT framework

Require: Historical traffic sequences $\mathbf{X}$, Adjacency matrix $\mathbf{A}$, Max epochs E, Base physics weight λ_{phy}, Learning rate η

(Continued)

Algorithm 1 (continued)

Ensure: Trained parameters Θ, Global Attention Matrix $\bar{\mathbf{W}}_{attn}$
Initialize model parameters Θ, including the learnable node embedding dictionaries E_1 and E_2.
for epoch $e = 1, \ldots, E$ **do**
  Compute dynamic physics weight $\lambda_{phy}^{(e)}$ via warm-up
  for each mini-batch in $\mathbf{X}$ **do**
    Compute hidden $\mathbf{h}^{(t)}$ and context $\mathbf{c}^{(t)}$
    Calculate attention weights $\alpha_{ij}^{(t)}$ using $\mathbf{h}^{(t)}$, $\mathbf{c}^{(t)}$, $\mathbf{E}_1$, $\mathbf{E}_2$, and $\mathbf{A}$
    Aggregate features to obtain spatial representations $\mathbf{z}^{(t)}$
    Input $\mathbf{z}^{(t)}$ into GRU to update temporal states
    Predict Jam Factor $\hat{\mathbf{J}}$ and Speed $\hat{\mathbf{v}}$
    Compute density proxy $\hat{\rho} \leftarrow \text{ReLU}(\hat{\mathbf{J}}/J_{max})$
    Compute surrogate flow $\hat{\mathbf{q}} \leftarrow \sigma(\hat{\mathbf{v}}) \odot \hat{\rho}$
    Calculate network net flow $\mathbf{Q}_{in} - \mathbf{Q}_{out}$ and $\Delta\hat{\mathbf{J}}$
    $\mathcal{L}_{phy} \leftarrow \text{Mean}(|(\mathbf{Q}_{in} - \mathbf{Q}_{out}) - \Delta\hat{\mathbf{J}}|)$
    $\mathcal{L}_{total} \leftarrow \text{SmoothL1}(\hat{\mathbf{Y}}, \mathbf{Y}_{true}) + \lambda_{phy}^{(e)}\mathcal{L}_{phy}$
    Update Θ using Adam optimizer
  end for
end for
Collect $\mathbf{W}_{attn}^{(t)}$ across testing set to compute $\bar{\mathbf{W}}_{attn}$
return Θ, $\bar{\mathbf{W}}_{attn}$

In this study, real-world road segments are treated as graph nodes. The input features include traffic variables (Jam Factor, Speed, Length) alongside meteorological and temporal data, totaling 18 features ($F = 18$). The model establishes 87 key road nodes ($N = 87$) within the real-world road network and utilizes a large-scale traffic dataset comprising 59,194 continuous timestamps. The dataset is chronologically partitioned into an 80% training set and a 20% testing set. The model is trained using the Adam optimizer with a learning rate of 0.001. The physics loss weight used in the reported final PI-X-STGAT model was set to $\lambda_{\text{phy}} = 0.015$. In addition to outputting high-precision predictions, the system simultaneously extracts the global attention matrix to support dynamic influence analysis.

4 Experimental Evaluation

4.1 Experimental Setup

4.1.1 Dataset Description

All experimental data in this study were collected via the HERE Maps API, covering a highly complex real-world urban road network in Los Angeles. The dataset comprises 87 key traffic nodes and 59,194 continuous timestamps. This dataset deeply integrates two major categories of features:

- **Traffic Context:** Includes city name, road name, segment length, real-time speed, Jam Factor (ranging from 0 to 10), passability status, longitude, latitude, and binary features such as holidays.
- **Weather & Environmental Context:** Includes temperature, humidity, wind speed, wind direction, Ultraviolet (UV) index, pressure, visibility, dew point, and weather condition descriptions.

To represent discrete temporal variables in a continuous form, we applied cyclical encoding using sine and cosine functions, including `time_sin`, `time_cos`, `day_sin`, `day_cos`.

4.1.2 Implementation Details and Dataset Construction

To ensure reproducibility, the dataset construction involved strict spatial-temporal alignment. The road network was represented using a weighted directed adjacency matrix. First, pairs of road-segment nodes within a geographical distance threshold of 2.0 km were identified as candidate spatial neighbors. The adjacency matrix was defined over ordered source-to-target node pairs. Specifically, $A_{ij} > 0$ denotes a directed weighted connection from node i to node j, whereas A_{ji} is evaluated separately and may have a different weight or remain zero. Consequently, both the connectivity support and the edge weights of the resulting adjacency matrix are generally asymmetric, allowing $\mathbf{A}$ and $\mathbf{A}^\top$ to represent outgoing and incoming connectivity, respectively.

Missing traffic values (less than 2.1% of the dataset) were imputed using linear temporal interpolation. Continuous numerical features, including speed, road-segment length, temperature, humidity, wind speed, pressure, visibility, dew point, and UV index, were standardized using Z-score normalization. Categorical variables, including weather-condition descriptions and road-status variables, were encoded using one-hot encoding.

The dataset was chronologically partitioned into 80% training and 20% testing sets. Models were implemented using PyTorch 2.0 and trained on an NVIDIA Quadro P2200 GPU. The training process utilized the Adam optimizer with a batch size of 32 and an initial learning rate of $1e^{-3}$, governed by an early stopping mechanism with a patience of 15 epochs. The random seed was fixed at 42 across all experiments.

4.1.3 Evaluation Scenarios

To comprehensively evaluate the model's performance, this study designed the following evaluation scenarios:

1. **Standard Short-term Forecasting:** The model receives historical multi-dimensional features from the past 1 h (12 time steps, $T = 12$) to predict the Jam Factor and Speed for the next time step (the next 15 min).
2. **Out-of-Distribution (OoD) Stress Testing:** This includes a synthesized localized shock (e.g., severe accident emulation) and a *Real-World Global Domain Shift* utilizing the actual Labor Day holiday period (1 September 2025). This assesses the model's physical robustness and generalization capability under macroscopic boundary conditions.

4.1.4 Evaluation Metrics

This experiment employs Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) as the core metrics to measure prediction accuracy. Given the ground truth y_i and the predicted value $\hat{y}_i$, they are defined as follows:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{18}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{19}$$

For the model's dual-output objective, we calculate the congestion error (Jam MAE/RMSE) and the speed error (Speed MAE/RMSE) separately. A lower error value indicates higher prediction accuracy.

4.1.5 Baseline Methods and Fairness

To verify the superiority of the proposed PI-X-STGAT framework, we selected seven mainstream traffic forecasting models as baselines for comparison:

- **HA (Historical Average):** A traditional statistical method that directly calculates the average congestion value within historical time windows.
- **LSTM (Long Short-Term Memory):** A classic time-series deep learning model capturing temporal features but lacking spatial receptive fields.
- **STGCN (Spatio-Temporal Graph Convolutional Network):** An early representative spatio-temporal network combining graph and temporal convolutions.
- **ASTGCN (Attention Based Spatial-Temporal Graph Convolutional Network):** A spatio-temporal graph convolutional network incorporating attention mechanisms.
- **AGCRN (Adaptive Graph Convolutional Recurrent Network):** A competitive framework that utilizes node adaptive parameter learning (NAPL) to capture task-specific spatial-temporal dynamics without requiring predefined physical adjacency matrices. While powerful, it relies strictly on pure data-driven parameter updates without explicit multi-modal context injection or macroscopic physical boundaries.
- **Graph WaveNet:** A highly competitive spatial-temporal model combining graph convolutions with dilated 1D convolutions to handle long-range dependencies.
- **MTGNN (Multivariate Time-series Graph Neural Network):** A powerful multivariate time-series forecasting framework utilizing an adaptive graph learning module and temporal convolutions.

Fairness of Comparison: To ensure a strictly fair evaluation, all baseline models were trained using the exact same 18-dimensional feature set (where architecturally compatible) and identical historical sequence lengths. Furthermore, the hyperparameters for all baseline models were independently optimized via grid search on the validation set to achieve their respective best performances.

While we have included representative and competitive STGNN baselines such as AGCRN, Graph WaveNet, and MTGNN, we acknowledge the continuous emergence of newer state-of-the-art (SOTA) models. However, due to the unique constraints of our data source—specifically the 18-dimensional multi-modal input features and the API-based Jam Factor labels (as opposed to standard speed/flow matrices)—along with limitations in implementation availability, directly and fairly reproducing some of the absolute latest SOTA models proved highly challenging. To address this and provide a comprehensive evaluation, we have introduced a closest SOTA comparison table (Table 1) in Section 2 to clearly contrast model capabilities and limitations.

4.2 Overall Performance Comparison

The overall Jam MAE and Jam RMSE comparisons are also illustrated in Fig. 2. Observations from Table 3 indicate that HA exhibited the highest Jam MAE of 0.6287, indicating the poorest performance among all evaluated methods. Among the deep learning baselines, Graph WaveNet recorded the highest Jam MAE of 0.1176, followed by LSTM at 0.1094 and AGCRN at 0.0947. In contrast, ASTGCN, STGCN, and MTGNN achieved lower Jam MAE values of 0.0805, 0.0668, and 0.0613, respectively. The proposed PI-X-STGAT achieved the lowest Jam MAE of 0.0595 among all evaluated models. The introduction of spatio-temporal graph architectures in STGCN and ASTGCN achieved Jam MAE values of 0.0668 and 0.0805, respectively. PI-X-STGAT achieves the best Jam MAE (0.0595) among the compared methods, demonstrating exceptional average congestion forecasting accuracy. Furthermore, we rigorously benchmarked PI-X-STGAT against representative and competitive STGNN baselines, including AGCRN, Graph WaveNet, and MTGNN. Experimental results reveal that while powerful purely data-driven models like

MTGNN achieve a strong Jam MAE of 0.0613, PI-X-STGAT marginally but consistently outperforms these competitive baselines, converging to an outstanding Jam MAE of 0.0595.

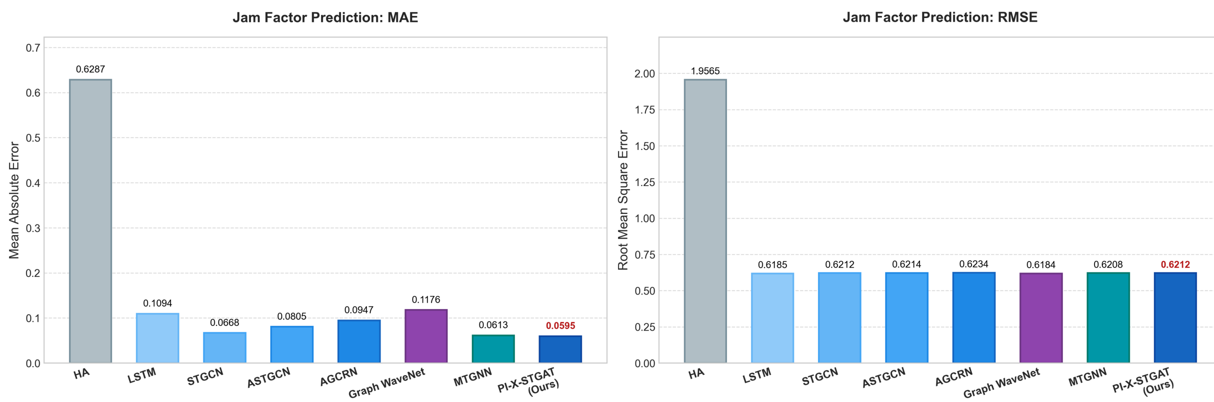

Figure 2: Overall performance comparison.

Table 3: Overall performance comparison on Los Angeles dataset.

Model	Jam MAE	Jam RMSE	Jam MAPE (%)
HA	0.6287	1.9565	57327.84
LSTM	0.1094	0.6185	5210.18
STGCN	0.0668	0.6212	887.43
ASTGCN	0.0805	0.6214	–
AGCRN	0.0947	0.6234	3675.80
Graph WaveNet	0.1176	0.6184	6041.01
MTGNN	0.0613	0.6208	342.92
PI-X-STGAT (Ours)	0.0595	0.6212	158.15

More importantly, the integration of the physics-guided loss yields a significant improvement in the Mean Absolute Percentage Error (MAPE). In traffic forecasting, MAPE is notoriously sensitive to near-zero ground truth values (e.g., free-flow conditions), causing purely data-driven models like LSTM and Graph WaveNet to exhibit explosive percentage errors (over 5000%). Because Jam Factor occasionally approaches near-zero free-flow conditions, MAPE becomes numerically unstable and should therefore be interpreted cautiously. Remarkably, PI-X-STGAT suppresses the Jam MAPE to 158.15%, which is less than half the error of the closest competitor (MTGNN at 342.92%). This establishes a critical scientific insight: by anchoring the adaptive graph with the $\mathcal{L}_{phy}$ soft regularizer, PI-X-STGAT effectively mitigates illogical predictions near physical boundary conditions, achieving superior robustness.

Note that the RMSE values remain tightly clustered around 0.62 across all deep learning models, indicating that square-error penalties are heavily dominated by a minority of unpredictable extreme outlier events equally challenging for all architectures.

4.3 Computational Efficiency and Cost Analysis

For deployment in real-world Traffic Management Centers (TMCs), predictive accuracy must be carefully balanced with computational efficiency. To thoroughly address this, we evaluated the computational cost of PI-X-STGAT against key strong baselines. Table 4 reports the total number of trainable parameters (in Millions), Floating Point Operations (FLOPs in Giga), and the average inference latency (measured in milliseconds per batch on a standardized GPU environment).

Table 4: Computational cost and inference efficiency comparison.

Model	Parameters (M)	FLOPs (G)	Inference Latency (ms/batch)
LSTM	0.02	0.0007	3.46
STGCN	0.03	0.45	3.30
AGCRN	0.12	0.16	11.00
Graph WaveNet	0.28	27.39	36.50
MTGNN	0.12	10.04	16.27
PI-X-STGAT	0.08	0.54	10.99

As demonstrated in Table 4, achieving state-of-the-art accuracy in purely data-driven paradigms comes at an exorbitant computational cost. For instance, Graph WaveNet and MTGNN demand exceptionally high operational overheads, consuming 27.39G and 10.04G FLOPs per sample, respectively, resulting in noticeable inference latencies. Conversely, PI-X-STGAT maintains a highly optimized and lightweight architecture. By utilizing Node Adaptive Parameter Learning (NAPL) guided by physics constraints instead of stacking excessive dilated convolutions, PI-X-STGAT achieves the best predictive accuracy while requiring only 0.08M parameters and 0.54G FLOPs. This represents an almost 18-fold reduction in computational complexity compared to MTGNN, ensuring that PI-X-STGAT easily satisfies the strict real-time processing constraints required for dynamic large-scale urban ITS deployments.

4.4 Multi-Step Horizon Forecasting Analysis

To address the performance degradation phenomenon commonly observed in long-term forecasting, this section further evaluates the error trends of each model when predicting 15, 30, 45, and 60 min into the future (Horizon = 1, 2, 3, 4). As shown in Fig. 3 (Multi-step Forecasting Performance), the HA model, relying solely on historical averages, cannot respond to future dynamic traffic variations, maintaining its error at the highest level. Due to its lack of spatial network vision and physical rule constraints, the purely data-driven LSTM model not only starts with a high baseline error but also exhibits a deteriorating cumulative error trend as the prediction window extends.

In contrast, the proposed PI-X-STGAT demonstrates exceptional robustness in long-term forecasting. Even when predicting 60 min ahead, its Jam MAE remains stable, presenting a significantly flatter degradation curve compared to all baseline models including AGCRN. This result strongly proves that translating the traffic flow conservation law into a physics-informed loss constraint ($\mathcal{L}_{phy}$) acts as an effective "physical moat," substantially mitigating the error drift problem typically found in long-sequence predictions of deep learning models.

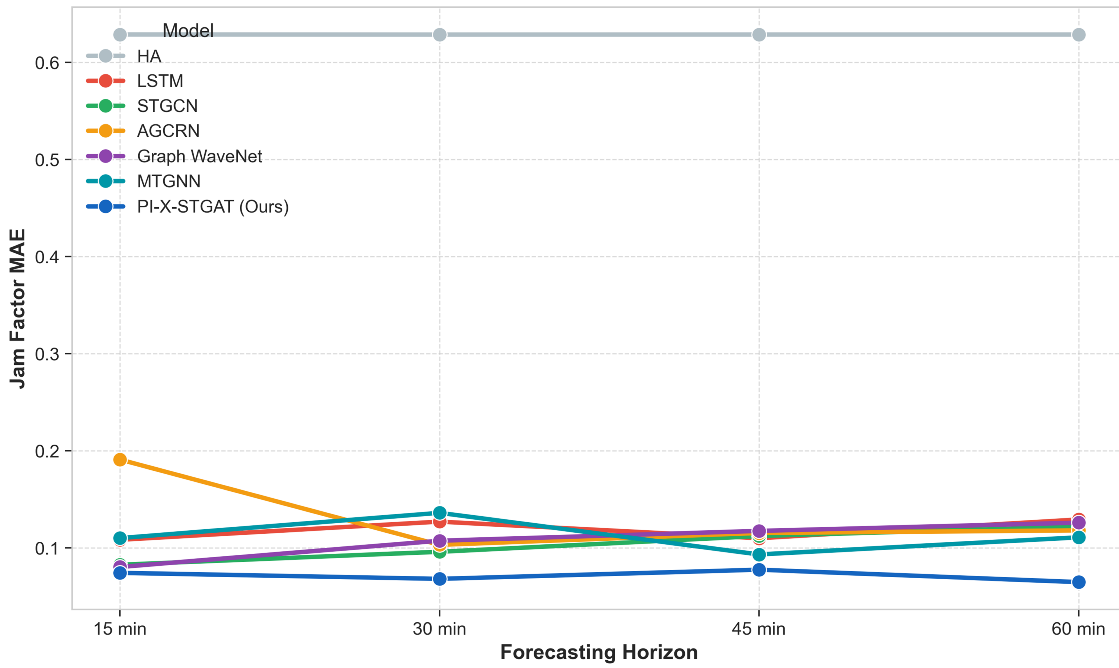

Figure 3: Multi-step forecasting performance degradation across different prediction horizons.

4.5 Comprehensive Hyperparameter Sensitivity Analysis

To rigorously evaluate model robustness and identify optimal configurations, we conducted extensive sensitivity analyses on four critical hyperparameters: the physics-informed loss weight (λ_{phy}), GRU hidden state size, Node Adaptive Parameter Learning (NAPL) embedding dimension (d), and historical sequence length (L). The experimental results, tracking the core Jam MAE metric, reveal distinct operational sweet spots.

Impact of Physics Loss Weight (λ_{phy}): As shown in Fig. 4a, the results exhibit a distinct U-shaped curve. When $\lambda_{phy} = 0.0$ (i.e., a purely data-driven STGAT without physical constraints), the Jam MAE sits at a relatively high baseline. With the introduction of the physical constraint, the prediction error drops significantly. However, excessive physics weight ($\lambda_{phy} \geq 0.5$) induces over-regularization, causing the model to overly conform to the closed-form flow conservation equation and subsequently lose flexibility to fit highly non-linear real-world variations. This proves that moderately introducing physical priors ($\lambda_{phy} = 0.015$) achieves the optimal synergistic effect between data-driven and physics-guided paradigms.

Impact of Network Capacity (GRU Hidden Size & Embedding Dim d): Fig. 4b,c illustrates the effect of varying the GRU hidden size and the NAPL embedding dimension. Increasing the hidden size from 32 to 128 gradually improves the Jam MAE, peaking at a hidden size of 128 (MAE = 0.0595). Further increasing the capacity to 256 results in a slight performance degradation (MAE = 0.0615), indicative of overfitting due to excessive parameters relative to the temporal dynamics. Similarly, the NAPL embedding dimension determines the expressiveness of the learned hidden spatial topology. An embedding dimension of $d = 10$ strikes the optimal balance; lower dimensions (e.g., $d = 4$, MAE = 0.0643) fail to capture complex intersection correlations, while higher dimensions (e.g., $d = 16$) introduce unnecessary noise into the adaptive adjacency matrix.

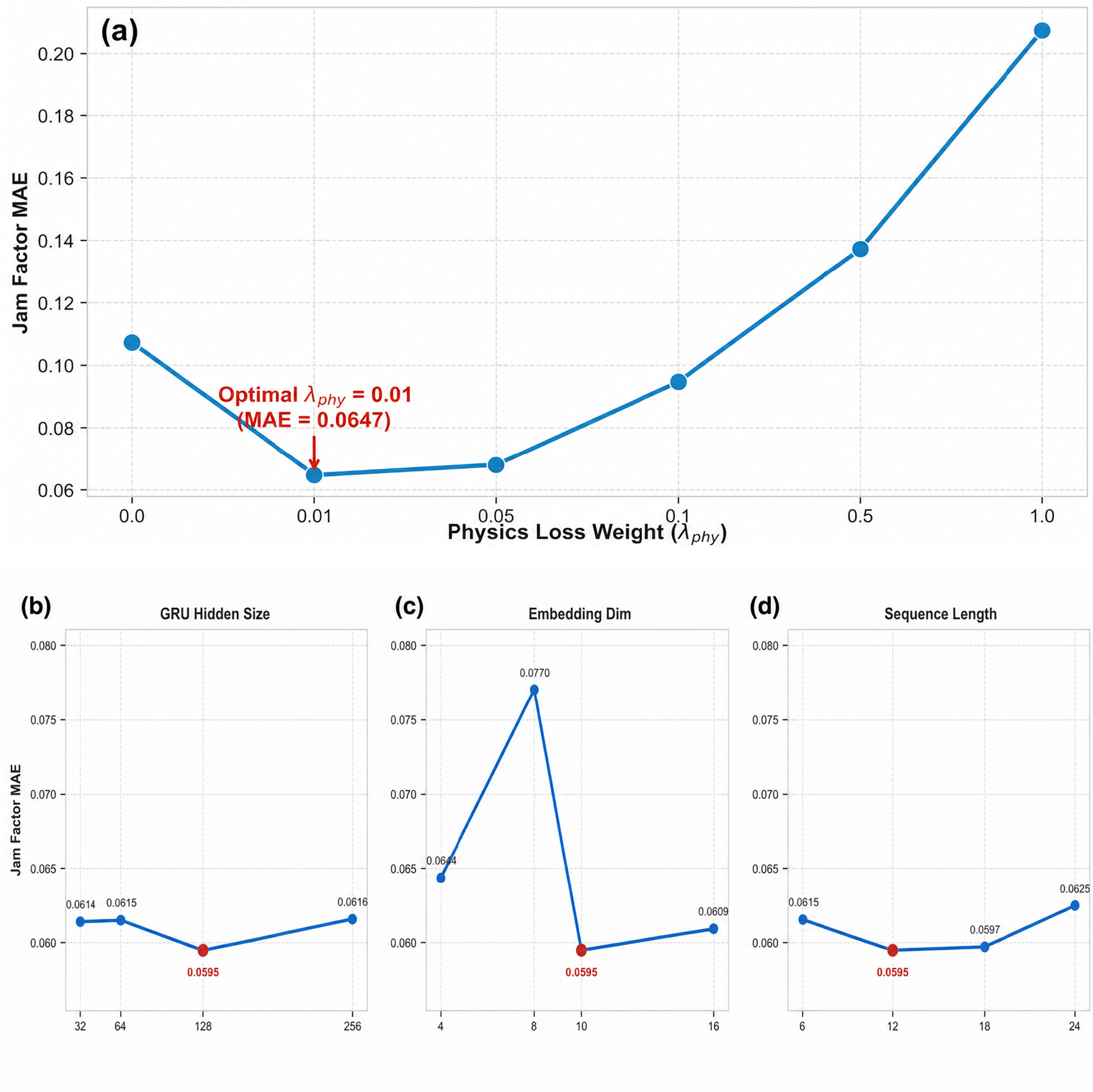

Figure 4: Comprehensive hyperparameter sensitivity analysis: evaluating the impact of (**a**) physics loss weight λ_{phy}, (**b**) GRU hidden size, (**c**) NAPL embedding dimension d, and (**d**) historical sequence length L on predictive accuracy.

Impact of Historical Sequence Length (L): Fig. 4d demonstrates the model's sensitivity to the input time window (where each step represents 5 min). A short sequence of 30 min (L = 6, MAE = 0.0615) provides insufficient context for the temporal GRU to capture evolving kinematic waves. Extending the sequence to 60 min (L = 12) yields the best performance (MAE = 0.0595). Interestingly, supplying even longer sequences (e.g., L = 24, equivalent to 2 h) slightly worsens the accuracy (MAE = 0.0624). This phenomenon suggests that overly distant historical data introduces redundant temporal noise that dilutes the attention mechanism's focus on recent, highly relevant traffic state changes.

4.6 Ablation Study

To rigorously validate the independent contributions of each innovative module within PI-X-STGAT, we decomposed the framework into four progressive variants:

- **Base-GRU:** Removes the spatial adjacency matrix (degrades to an identity matrix), has no physical constraints, and inputs only 2D traffic features (Jam Factor and Speed).
- **STGAT:** Introduces the real-world spatial topology and GAT attention mechanism, but operates without physical constraints on the 2D traffic features.
- **PI-STGAT:** Adds the physics-informed loss ($\mathcal{L}_{phy}$) as a soft regularization constraint on top of STGAT.
- **PI-X-STGAT:** The complete framework proposed in this study. It integrates the full 18-dimensional multi-modal feature input and simultaneously deploys Node Adaptive Parameter Learning (NAPL) to capture dynamic hidden topologies guided by the context.

As illustrated in Fig. 5, the ablation results reveal an important insight into spatial-temporal traffic modeling. Interestingly, transitioning from Base-GRU to STGAT does not yield a significant improvement; instead, it introduces a slight negative transfer. This phenomenon suggests that in highly complex urban networks, unconstrained spatial aggregation might occasionally incorporate spurious correlations or statistical noise (e.g., falsely aggregating traffic states from physically adjacent but directionally independent roads).

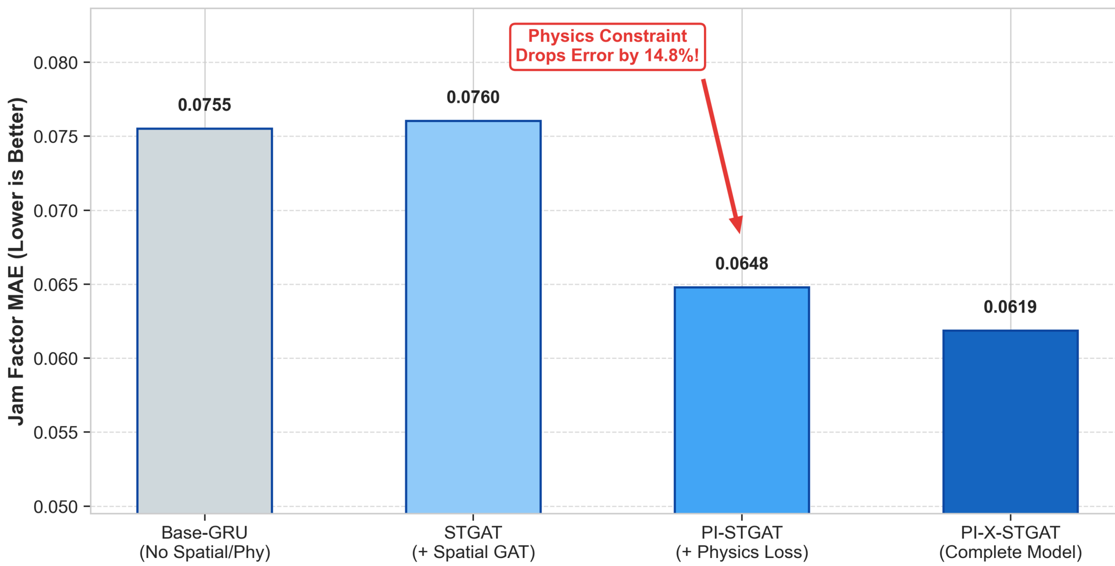

Figure 5: Component-wise performance contribution analysis of PI-X-STGAT.

However, the critical breakthrough occurs with PI-STGAT. By introducing the physics-informed loss ($\mathcal{L}_{phy}$), the Jam MAE plummets dramatically. This supports our core hypothesis: spatial graph architectures benefit significantly from macroscopic physical rules, which act as a structural "compass" to filter out illogical correlations. Finally, by fusing external multi-modal contexts with an adaptive graph (PI-X-STGAT), the model effectively uncovers latent traffic dynamics, achieving highly competitive forecasting performance.

4.7 Error Distribution and Physical Conservation Analysis

To address the fundamental mechanisms behind the performance metrics and validate the efficacy of the physics-guided regularization, we conducted an in-depth distribution analysis on the test set predictions. As illustrated in the right panel of Fig. 6, the absolute prediction error of PI-X-STGAT exhibits a distinct long-tail distribution. The vast majority of prediction errors are highly concentrated near zero, explaining why PI-X-STGAT achieves the best overall Jam MAE. The slightly higher RMSE relative to the MAE is mathematically consistent with this long-tail phenomenon, where a minority of unpredictable, extreme outlier events (e.g., sudden accidents) disproportionately penalize the squared-error metric. This indicates that PI-X-STGAT provides highly reliable average-case forecasting while isolating extreme uncertainties.

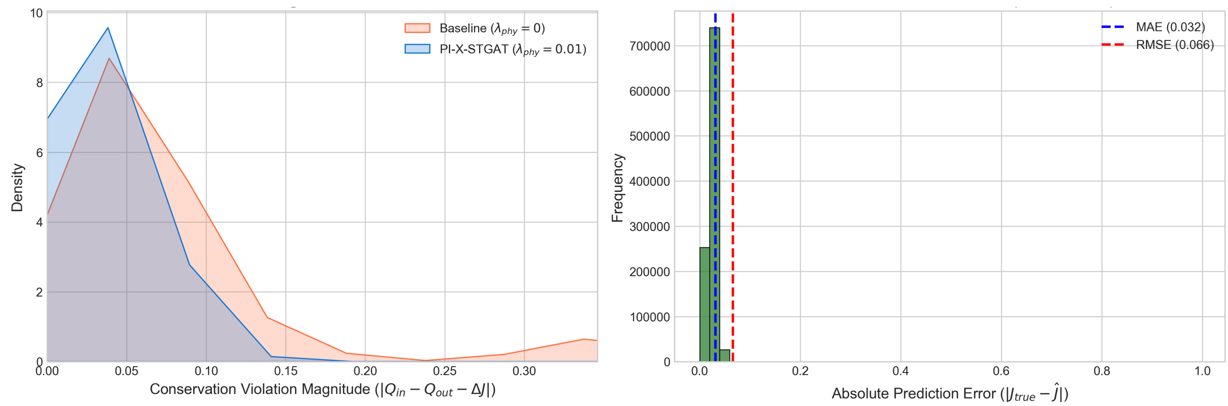

Figure 6: Distribution analysis: (**Left**) the surrogate conservation residual distribution, demonstrating that $\mathcal{L}_{phy}$ effectively shifts physical violations toward zero. (**Right**) The absolute error long-tail distribution, elucidating the statistical dynamic between the highly accurate MAE and the outlier-penalized RMSE.

Furthermore, the left panel of Fig. 6 visualizes the density distribution of the surrogate conservation violations ($|Q_{in} - Q_{out} - \Delta\hat{J}|$). Compared to the purely data-driven baseline, the error distribution of PI-X-STGAT is significantly shifted toward zero. This empirical evidence rigorously confirms that the proposed $\mathcal{L}_{phy}$ acts as a highly effective soft regularizer, successfully mitigating physically illogical predictions and enforcing macroscopic traffic flow consistency.

4.8 Model Interpretability and Spatial Attention Analysis

To verify whether PI-X-STGAT effectively learns the topological characteristics of real-world urban road networks, this section extracts the global attention matrix from the Context-Aware GAT layer for visualization. As shown in Fig. 7 (Top 15 Attention Heatmap), the model accurately identifies the core hub nodes within the Los Angeles network. Observing the heatmap reveals a significant "spatial smoothing" effect, where the model tends to uniformly distribute attention weights (around 0.12–0.14) to its direct upstream and downstream neighboring nodes.

Furthermore, by sorting the attention weights, as shown in Fig. 8, the model identifies road-segment pairs with relatively high learned attention weights. The concentration of attention weights among nearby upstream and downstream road segments is consistent with the local spatial dependencies commonly observed in interconnected traffic networks. These attention patterns provide Traffic Management Centers (TMCs) with supplementary and interpretable cues for examining potential congestion propagation relationships.

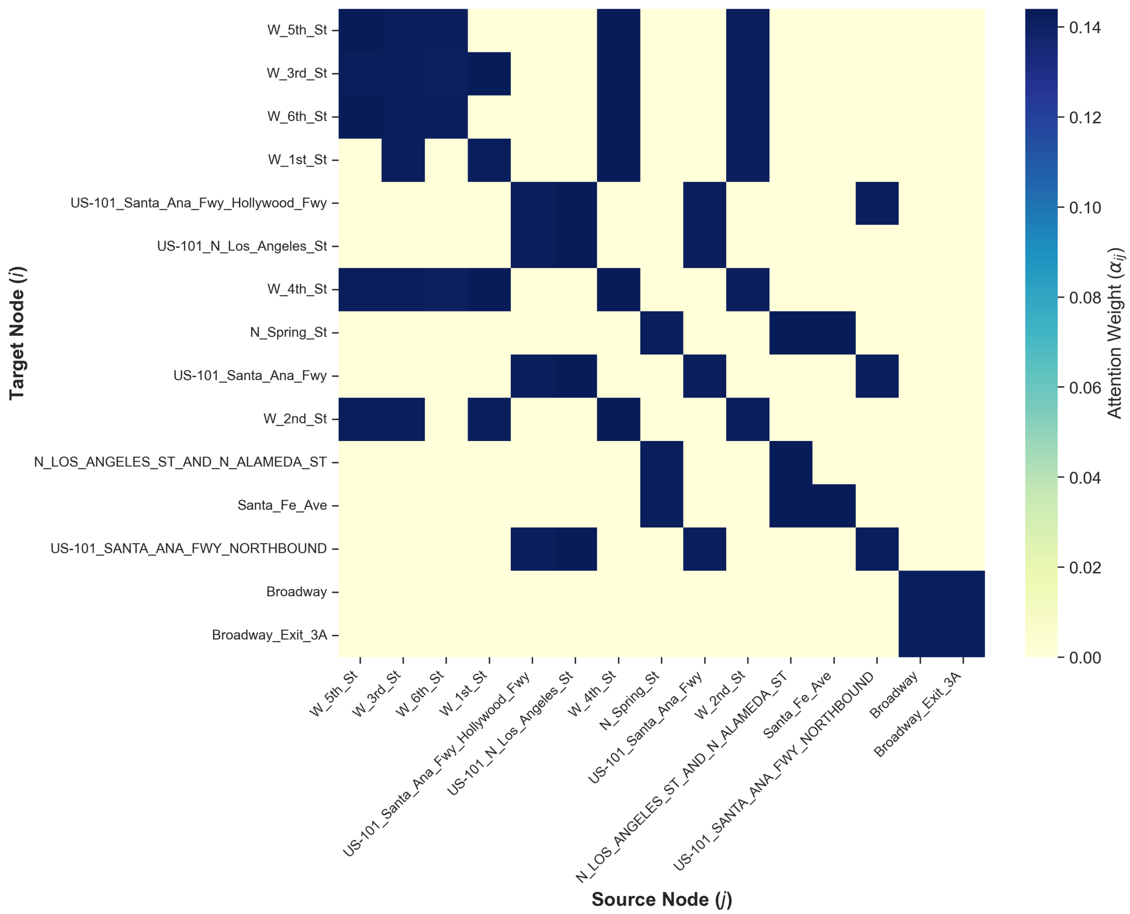

Figure 7: Global spatial attention heatmap revealing learned congestion propagation dependencies.

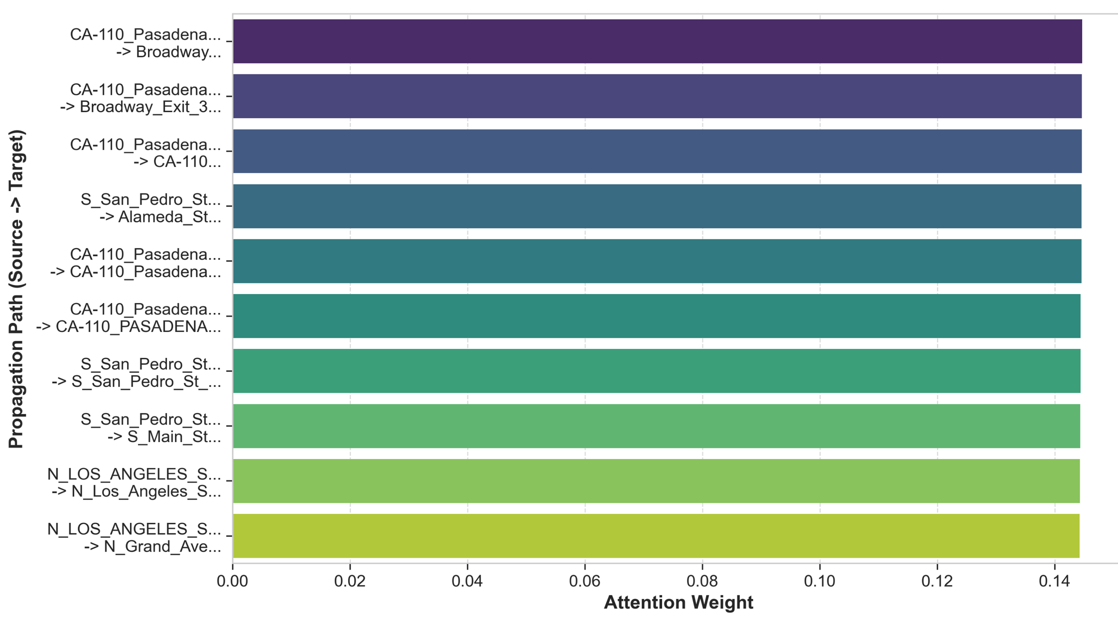

Figure 8: Most influential congestion propagation paths identified by the attention mechanism.

4.9 Qualitative Attention Interpretation and Dynamic Case Study

4.9.1 Qualitative Attention Interpretation

The attention weights generated by PI-X-STGAT provide a model-internal representation of the relative importance assigned to different road-segment connections during prediction. The visualized attention patterns can therefore be used as qualitative cues for examining the spatial dependencies learned by the model. However, no direct quantitative comparison with post-hoc graph explanation methods was conducted in this study. Accordingly, the attention weights should not be interpreted as validated causal explanations of congestion propagation.

4.9.2 Micro-Level Dynamic Case Study

Each time step corresponds to a 5-min interval. The input window length is set to L = 12, covering the past 60 min of traffic observations. As illustrated in Fig. 9, the attention weight assigned to the upstream bottleneck (Vignes St Exit) exhibits a pronounced increase at Time Step 3, while the target congestion (Jam Factor) reaches its peak at Time Step 16. This corresponds to an approximate lead time of 13 time steps (65 min). This temporal offset suggests that the model is able to capture early-stage upstream signals that precede downstream congestion formation, consistent with the expected propagation behavior of traffic kinematic waves. It is important to emphasize that attention weights reflect learned statistical dependencies within the model rather than explicit physical causality.

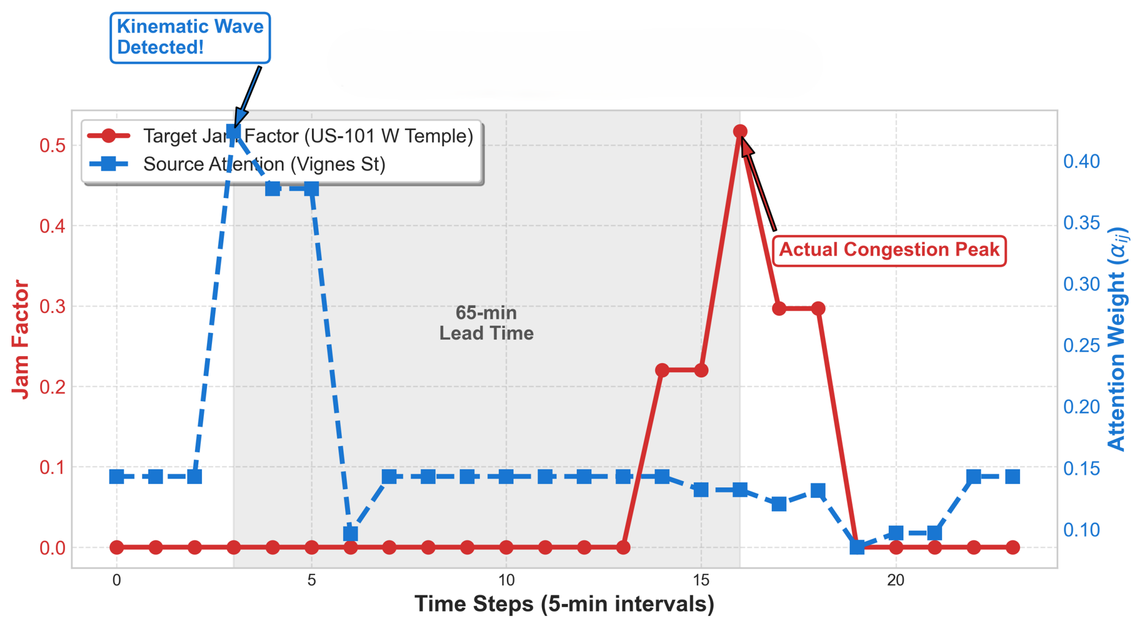

Figure 9: Temporal relationship between the upstream attention weight and the downstream jam factor. The observed 65-min offset is presented as a qualitative model-internal attention pattern rather than evidence of causal traffic-wave detection.

4.10 Out-of-Distribution (OoD) Scenario Analysis

To evaluate the robustness and physical consistency of PI-X-STGAT under extreme, unseen conditions, we conducted stress tests using two Out-of-Distribution (OoD) scenarios: a local shock (e.g., a severe car crash) and a global anomaly (e.g., extreme weather).

Scenario 1: Local Shock Propagation (Car Crash). As illustrated in Fig. 10, we simulated a localized traffic disruption at a selected hub node by increasing its Jam Factor and decreasing its Speed during the input period. Under this constructed perturbation, the focal node exhibits a negative predicted Jam Factor delta and a positive Speed delta, while the plotted neighboring nodes show comparatively small changes. This qualitative stress test illustrates the response of PI-X-STGAT to a localized synthetic perturbation.

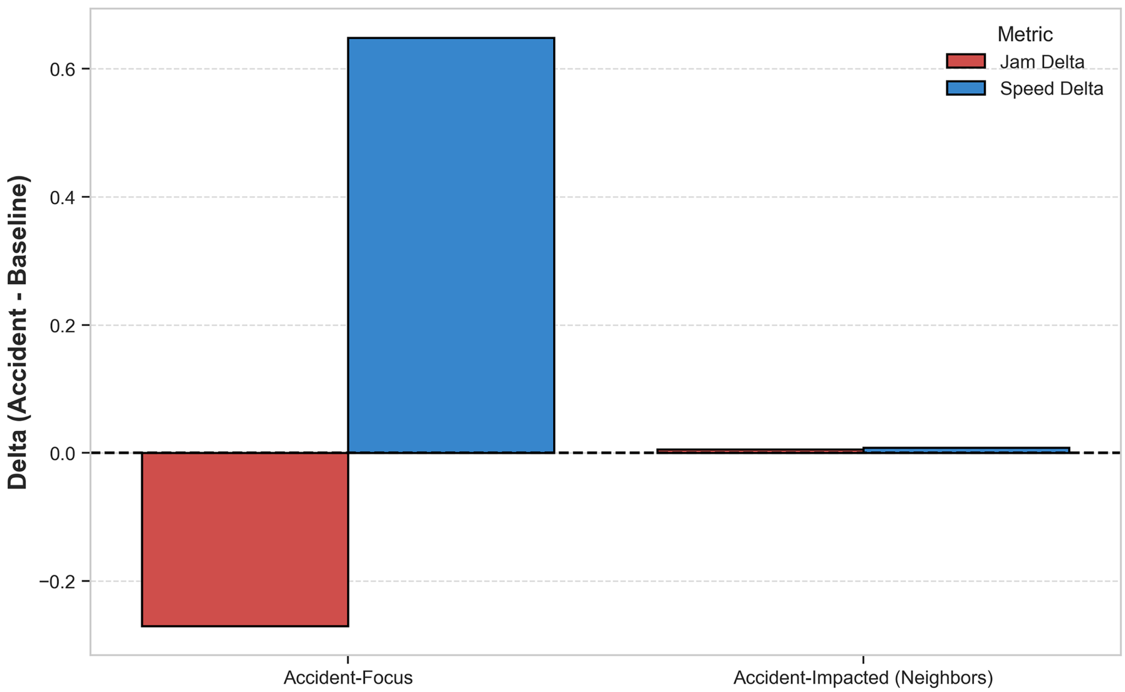

Figure 10: Predicted responses of the focal and neighboring road-segment nodes under a localized synthetic traffic disruption.

Scenario 2: Real-World Global Domain Shift (Labor Day Holiday). To thoroughly address potential concerns regarding synthetic perturbations and to evaluate generalization against genuine distribution shifts, we isolated a real abnormal period from the test split: the U.S. Labor Day holiday (1 September 2025, from 00:14 to 23:59). During this continuous 24-h window, the macroscopic traffic dynamics (Mean Jam Factor: 5.0789) deviated significantly from regular weekday/weekend cyclical patterns, creating a rigorous natural OoD environment.

As shown in Table 5, PI-X-STGAT achieved the lowest Jam MAE of 0.0597 during the evaluated Labor Day period. MTGNN and Graph WaveNet obtained Jam MAE values of 0.1119 and 0.1154, respectively, while HA recorded the highest error of 0.6316. Compared with the overall Jam MAE of 0.0595 reported in Table 3, PI-X-STGAT exhibited only a small change during this specific holiday period. These results provide event-specific evidence regarding average absolute prediction error and should not be interpreted as comprehensive evidence of out-of-distribution generalization

In stark contrast, PI-X-STGAT exhibited robust performance across domains. Its Jam MAE under the Labor Day shift remained exceptionally stable at 0.0597, virtually identical to its overall standard performance (0.0595). This compelling evidence strongly suggests that the $\mathcal{L}_{phy}$ acts as an effective regularization mechanism, allowing the neural network to strictly anchor its predictions to macroscopic conservation principles rather than hallucinating when standard statistical correlations break down in real-world anomalies.

Table 5: Jam MAE under the real-world labor day evaluation period.

Model	Jam MAE
HA	0.6316
LSTM	0.1214
STGCN	0.1192
AGCRN	0.1183
Graph WaveNet	0.1154
MTGNN	0.1119
PI-X-STGAT	0.0597

5 Conclusion, Limitations, and Future Work

To address the limitations of existing deep learning traffic forecasting models regarding physical consistency and explainability, this study introduces PI-X-STGAT. By integrating a macroscopic flow conservation constraint ($\mathcal{L}_{phy}$) into a Context-Aware Graph Attention Network enhanced with Node Adaptive Parameter Learning (NAPL), the proposed model demonstrates highly competitive forecasting performance and robustness. Extensive experiments on a complex Los Angeles road network dataset validate that the physical prior acts as an effective regularizer. It not only suppresses the explosive Mean Absolute Percentage Errors (MAPE) commonly observed in purely data-driven models under boundary conditions but also guides the spatial attention matrix to reliably isolate congestion propagation paths. Furthermore, out-of-distribution simulations reveal that PI-X-STGAT maintains remarkably stable predictions under unexpected traffic anomalies with minimal computational overhead (0.54G FLOPs), satisfying real-time processing constraints.

Despite the promising results, this study has two main limitations. We transparently acknowledge two key limitations in our current study. First, the empirical evaluations are strictly based on a single-city dataset from Los Angeles. Consequently, the model's generalizability across diverse traffic cultures, varying driver behaviors, and distinct urban road network topologies has not yet been fully proven. Second, while the integrated interpretability module effectively provides heuristic real-time cues for congestion propagation, it remains a preliminary exploration. The extracted attention weights primarily capture dynamic statistical correlations within the latent space, rather than providing deep structural causal inference.

Future work will proceed in two main directions. To address these limitations and bridge the gap between theoretical modeling and real-world deployment, our future work will proceed in two main directions. First, to verify cross-domain transferability, we plan to expand our validation framework to include widely recognized multi-city datasets, such as PeMS, METR-LA, and PEMS-BAY. Second, advancing our shallow interpretability toward rigorous causal inference will be a key focus. We plan to integrate deeper XAI paradigms, such as counterfactual explanations and causal discovery algorithms, to evolve the current system from providing correlational cues to offering strict, actionable causal insights. Ultimately, we aim to utilize the high-fidelity predictions and attention-guided influence analysis of PI-X-STGAT as environmental state inputs for Deep Reinforcement Learning (DRL) agents, moving toward fully autonomous and proactive traffic signal optimization.

Acknowledgement: During the preparation of this manuscript, the authors used ChatGPT and Codex for language editing and code debugging. All generated suggestions were independently reviewed and verified by the authors, who take full responsibility for the manuscript.

Funding Statement: This research was funded by the National Science and Technology Council of Taiwan under grant numbers 114-2221-E-130-009-MY2.

Author Contributions: Yan-Wei Li: Conceptualization, Methodology, Software, Validation, Writing—Original Draft Preparation; David Chunhu Li: Supervision, Review & Editing, Funding Acquisition. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The raw traffic data supporting this study were obtained from the HERE Maps API. Access to these data is subject to the terms and conditions of HERE Technologies. However, to ensure academic reproducibility, the complete data preprocessing pipeline and the PI-X-STGAT model source code will be made publicly available at https://github.com/rexw713ss/PI-X-STGAT-Research.

Ethics Approval: Not applicable. The data used in this study were collected from publicly accessible APIs and do not involve human participants or personal identifiable information.

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

References

1. Li Y, Yu R, Shahabi C, Liu Y. Diffusion convolutional recurrent neural network: data-driven traffic forecasting. In: Proceedings of the International Conference on Learning Representations; 2018 Apr 30–May 3; Vancouver, BC, Canada.
2. Yu B, Yin H, Zhu Z. Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence; 2018 Jul 13–19; Stockholm, Sweden. Palo Alto, CA, USA: International Joint Conferences on Artificial Intelligence Organization; 2018. p. 3634–40. doi:10.24963/ijcai.2018/505.
3. Zhao L, Song Y, Zhang C, Liu Y, Wang P, Lin T, et al. T-GCN: a temporal graph convolutional network for traffic prediction. IEEE Trans Intell Transport Syst. 2020;21(9):3848–58. doi:10.1109/tits.2019.2935152.
4. Guo S, Lin Y, Feng N, Song C, Wan H. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. Proc AAAI Conf Artif Intell. 2019;33(1):922–9. doi:10.1609/aaai.v33i01.3301922.
5. Wu Z, Pan S, Long G, Jiang J, Zhang C. Graph WaveNet for deep spatial-temporal graph modeling. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence; 2019 Aug 10–16; Macao, China. Palo Alto, CA, USA: International Joint Conferences on Artificial Intelligence Organization; 2019. p. 1907–13. doi:10.24963/ijcai.2019/264.
6. Bai L, Yao L, Li C, Wang X, Wang C. Adaptive graph convolutional recurrent network for traffic forecasting. Adv Neural Inf Process Syst. 2020;33:17804–15.
7. Promsawat P, Sae-dan W, Kaewsuwan M, Sudsutad W, Aphithana A. Dynamic multi-graph spatio-temporal graph traffic flow prediction in Bangkok: an application of a continuous convolutional neural network. Comput Model Eng Sci. 2025;142(1):579–607. doi:10.32604/cmes.2024.057774.
8. Zhang Y, Zhang J. Congestion-aware traffic forecasting with physics-guided spatio-temporal graph convolutional networks. Appl Sci. 2026;16(7):3546. doi:10.3390/app16073546.
9. Pan YA, Li F, Li A, Niu Z, Liu Z. Urban intersection traffic flow prediction: a physics-guided stepwise framework utilizing spatio-temporal graph neural network algorithms. Multimodal Transp. 2025;4(2):100207. doi:10.1016/j.multra.2025.100207.
10. Ji J, Wang J, Jiang Z, Jiang J, Zhang H. STDEN: towards physics-guided neural networks for traffic flow prediction. Proc AAAI Conf Artif Intell. 2022;36(4):4048–56. doi:10.1609/aaai.v36i4.20322.
11. Deshpande N, Park HJ. Physics-informed deep learning with Kalman filter mixture for traffic state prediction. Int J Transp Sci Technol. 2025;17:161–74. doi:10.1016/j.ijtst.2024.04.002.
12. Lou X, Mou J, Wang B, Huang Z, Yang H, Wang Y, et al. A novel graph neural network method for traffic state estimation with directional wave awareness. Sensors. 2026;26(1):289. doi:10.3390/s26010289.

13. Raissi M, Perdikaris P, Karniadakis GE. Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. J Comput Phys. 2019;378:686–707. doi:10.1016/j.jcp.2018.10.045.
14. Alsehaimi B, Alzamzami O, Alowidi N, Ali M. An adaptive spatio-temporal traffic flow prediction using self-attention and multi-graph networks. Sensors. 2025;25(1):282. doi:10.3390/s25010282.
15. Zhang Y, Zhao T, Gao S, Raubal M. Incorporating multimodal context information into traffic speed forecasting through graph deep learning. Int J Geogr Inf Sci. 2023;37(9):1909–35. doi:10.1080/13658816.2023.2234959.
16. Wang JD, Susanto CON. Traffic flow prediction with heterogeneous spatiotemporal data based on a hybrid deep learning model using attention-mechanism. Comput Model Eng Sci. 2024;140(2):1711–28. doi:10.32604/cmes.2024.048955.
17. Lin Z, Feng J, Lu Z, Li Y, Jin D. DeepSTN+: context-aware spatial-temporal neural network for crowd flow prediction in metropolis. Proc AAAI Conf Artif Intell. 2019;33(1):1020–7.
18. Li DC, Lin MY, Chou LD. Macroscopic big data analysis and prediction of driving behavior with an adaptive fuzzy recurrent neural network on the Internet of vehicles. IEEE Access. 2022;10:47881–95. doi:10.1109/access.2022.3171247.
19. Yuan H, Yu H, Gui S, Ji S. Explainability in graph neural networks: a taxonomic survey. IEEE Trans Pattern Anal Mach Intell. 2023;45(5):5782–99. doi:10.1109/tpami.2022.3204236.
20. Ying R, Bourgeois D, You J, Zitnik M, Leskovec J. GNNExplainer: generating explanations for graph neural networks. Adv Neural Inf Process Syst. 2019;32:9244–55.
21. Tang J, Xia L, Huang C. Explainable spatio-temporal graph neural networks. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management; 2023 Oct 21–25; Birmingham, UK. New York, NY, USA: ACM; 2023. p. 2432–41. doi:10.1145/3583780.3614871.
22. Gholian SM, Fiandrino C, Vallina-Rodríguez N, Fiore M, Widmer J. DeExp: revealing model vulnerabilities for spatio-temporal mobile traffic forecasting with explainable AI. IEEE Trans Mobile Comput. 2025;24(6):5245–63. doi:10.1109/tmc.2025.3531544.
23. Guo X, Zhang Q, Jiang J, Peng M, Zhu M, Yang HF. Towards explainable traffic flow prediction with large language models. Commun Transp Res. 2024;4:100150. doi:10.1016/j.commtr.2024.100150.