EVIPHYS: EVIDENCE-AWARE MULTIMODAL REPRESENTATION LEARNING WITH PHYSICS-GUIDED GRAPH MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning reliable representations from structured spatio-temporal data is crucial for traffic flow prediction. However, three major challenges remain: effectively leveraging structural and semantic information from traffic data, improving the robustness of traffic representation under noisy and uncertain observations, and incorporating physical knowledge to enhance prediction reliability. To address these challenges, this paper proposes **EviPhys**, an **Evi**dence-aware multimodal representation learning framework with **Phys**ics-guided graph modeling. First, the Multimodal Representation Fusion Module (MRFM) converts historical traffic measurements into a three-channel image that encodes traffic level, temporal variation, and sensor-axis variation, and aligns the image with a state-aware text prompt through a frozen Vision-Language Model (VLM) to obtain macroscopic traffic semantics. Second, the Dempster-Shafer Evidence Reasoning Module (DSERM) maps each sensor's historical window to a Basic Probability Assignment (BPA) and uses Deng entropy to generate reliability-aware microscopic representations. Third, the Evidence Semantic Graph Learning Module (ESGLM) fuses the macroscopic semantics and microscopic evidence to construct a directed propagation graph. Fourth, the Physics-Guided Prediction Module (PGPM) recursively predicts traffic flow increments and regularizes the predicted flow sequence with a graph-diffusion prior defined on the physical sensor graph, encouraging topology-consistent flow evolution and stable long-horizon prediction. Experiments on four PEMS benchmarks and two large-scale LargeST datasets demonstrate the effectiveness and scalability of EviPhys, achieving average reductions of 2.60%, 1.08%, and 2.17% in MAE, RMSE, and MAPE compared with the second-best results on the PEMS benchmarks. On the SD and GBA datasets, EviPhys achieves the best performance across all reported horizons and metrics, with average reductions of 9.57%, 7.73%, and 11.57% in MAE, RMSE, and MAPE on SD, and 7.80%, 4.41%, and 21.47% on GBA, respectively. Code is available at `https://anonymous.4open.science/r/EviPhys-CC32`.

1 INTRODUCTION

Learning reliable representations from structured spatio-temporal data is fundamental to accurate traffic flow prediction (Xiang et al., 2026; Wang et al., 2026a), where complex temporal dynamics and spatial dependencies must be jointly captured from evolving observations (Tian et al., 2026; Jungel et al., 2025). Accurate flow predictions not only support congestion warning and route planning but also help intelligent vehicles anticipate network-wide traffic trends (Zou et al., 2026; Philion et al., 2024). In recent years, traffic flow prediction has gradually evolved from traditional statistical models (Stock & Watson, 2001; Pan et al., 2012) to deep learning models, such as Recurrent Neural Networks (RNNs) (Yoon et al., 2025), Graph Neural Networks (GNNs) (Zhang et al., 2026a), and Transformers (Chen et al., 2026a). Related research has also begun employing multimodal attention (Chang et al., 2026), cross-modal alignment (Chen et al., 2026b), and VLMs (Kim et al., 2021) to extract complementary information from various representations of traffic sequences. However, significant modality gaps exist between numerical traffic sequences and semantic representations like images and text (Zhong et al., 2025). Existing methods mostly rely on feature concatenation or

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.

attention mechanisms, making effective connections between the spatio-temporal evolution patterns difficult to establish.

Observations in spatio-temporal data are often affected by factors such as equipment failures, communication delays, and environmental interference (Wang et al., 2026b). Such measurement errors lead to significant noise and incompleteness (Zhang et al., 2026d), which in turn degrades the accuracy of traffic flow prediction models (Zhou et al., 2026). Currently, methods such as Bayesian learning (Duan et al., 2025), fuzzy inference (Khan & Ali, 2026), and evidential fusion (Zhan et al., 2025) have been applied to traffic state estimation and risk analysis. Dempster-Shafer evidence theory (DSET) (Shafer, 1976; Dempster, 1967) can represent the degree of support for different propositions and proposition sets through BPAs (Li et al., 2025b), while preserving the ignorance resulting from a lack of information (Yi et al., 2026; Pellegrini & Brunelli, 2026). Compared with methods that rely on a single probability distribution, DSET can more flexibly characterize the randomness and epistemic uncertainty in traffic observations (Wei et al., 2026). Furthermore, Deng entropy (Deng, 2016) can measure the degree of uncertainty in the evidential distribution, providing a basis for identifying low-reliability observations (Tang & Xiao, 2025). However, most existing studies apply uncertainty handling to post-prediction confidence evaluation, and rarely embed evidential uncertainty directly into the traffic feature learning and graph topology construction processes.

Physics-informed neural networks (PINNs) (Zhang et al., 2026b) and related physics-guided learning methods (Goswami et al., 2026; Panda et al., 2026), by incorporating traffic flow conservation relations, dynamic equations, or spatio-temporal evolution laws into model training (Cong et al., 2026), can reduce implausible predictions generated by purely data-driven models, as shown in Figure 1. Physics-guided methods can also alleviate error accumulation in multi-step flow predictions (Zhang et al., 2026c). Existing studies have attempted to combine the traffic flow continuity equation (Li et al., 2025a), the kinematic wave model (Newell, 1993), and other dynamic laws (Su et al., 2026) with deep neural networks to enhance model stability in scenarios of data scarcity and long-horizon prediction (Fofanah et al., 2026). However, traffic evolution in real-world road networks exhibits strong non-linearity and non-stationarity, and there may be scale imbalances and optimization conflicts between physics-based residuals and data-fitting losses.

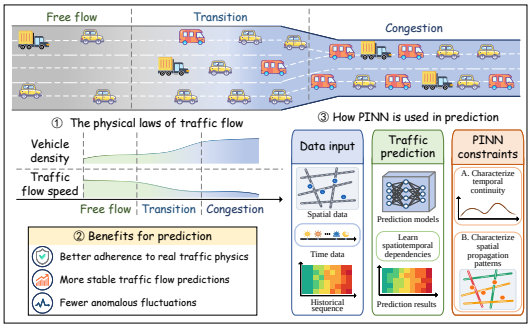


Figure 1: Illustration of physics-guided traffic flow prediction. PINN constraints characterize temporal continuity and spatial propagation over the physical road network and regularize data-driven predictions, promoting physically consistent traffic evolution.

To address the aforementioned issues, EviPhys introduces an evidence-aware multimodal framework with physics-guided learning. The framework integrates multimodal traffic representations, evidential reliability assessment, and physics-consistent graph modeling to improve the accuracy of traffic flow prediction. The main contributions are fourfold:

- **Multimodal Representation Fusion Module (MRFM).** MRFM transforms historical traffic data into three-channel images that describe traffic states, temporal changes, and cross-sensor variations. These images are aligned with text prompts through a frozen VLM to capture complementary structural and semantic traffic information.
- **Dempster-Shafer Evidence Reasoning Module (DSERM).** DSERM maps sensors' historical observations to BPAs and uses Deng entropy to construct uncertainty-sensitive local evidence representations. This mechanism reduces the influence of uncertain local information before graph construction.
- **Evidence Semantic Graph Learning Module (ESGLM).** ESGLM integrates global traffic semantics with local evidential representations to construct a directed graph. The learned graph can adapt to the current traffic state and the local evidential state.
- **Physics-Guided Prediction Module (PGPM).** PGPM combines graph recurrent forecasting with physics-guided graph-diffusion regularization, where physical sensor connectivity

is introduced during recursive decoding. A graph-diffusion prior further regularizes the predicted flow sequence, improving the stability of long-horizon prediction.

2 PRELIMINARIES

This section briefly reviews the fundamental concepts of EviPhys, including DSET for uncertainty modeling and PINNs for physics-guided learning.

2.1 DEMPSTER-SHAFER EVIDENCE THEORY (DSET)

DSET is effective for modeling epistemic uncertainty by allocating probabilities to the power set of events. Let $\Theta = \{e_1, e_2, \ldots, e_M\}$ be a finite set of mutually exclusive propositions, termed the frame of discernment, with its power set denoted as 2^Θ. A BPA represents the degree of support for a proposition, defined as a mapping $m : 2^\Theta \to [0, 1]$ satisfying $m(\emptyset) = 0$ and $\sum_{A\subseteq\Theta} m(A) = 1$ for all $A \in 2^\Theta$. The value $m(A)$ strictly quantifies the exact belief assigned to A. Subsets with $m(A) > 0$ are termed focal elements. By assigning probability masses to subsets rather than singletons, DSET flexibly represents ignorance and uncertainty.

2.2 PHYSICS-INFORMED NEURAL NETWORKS (PINNS)

PINNs achieve a joint data- and physics-driven paradigm by embedding Partial Differential Equations (PDEs) as regularization constraints into the optimization process of neural networks. Consider a general nonlinear PDE system defined on a bounded spatial domain $\mathcal{D}_x$ and time interval $[0, T]$: $\frac{\partial u}{\partial t} + \mathcal{N}_x[u] = 0$, where $u(x, t)$ denotes the latent physical field and $\mathcal{N}_x[\cdot]$ is a nonlinear spatial differential operator. This system is subject to the initial condition $u(x, 0) = h(x)$ and boundary condition $u(x, t) = g(x, t)$ for $x \in \partial\mathcal{D}_x$. A PINN approximates the latent field with a neural network $u_\theta(x, t)$ and evaluates the physical residual $r_\theta(x, t) = \partial_t u_\theta + \mathcal{N}_x[u_\theta]$ through automatic differentiation. Training combines the mean-squared residual, boundary-condition error, and initial-condition error, with nonnegative coefficients balancing the three terms. This principle is retained in EviPhys, but the continuous spatial operator is replaced by graph diffusion so that physical regularization follows the sensor connectivity induced by the road network rather than an assumed Euclidean grid.

3 METHODOLOGY

EviPhys consists of four main components: MRFM, DSERM, ESGLM, and PGPM. The overall framework is shown in Figure 2.

3.1 PROBLEM FORMULATION

Let $\mathcal{V} = \{v_1, \ldots, v_N\}$ be the set of N traffic sensors. Given a batch size of B, the historical input is $X \in \mathbb{R}^{B\times T\times N\times 3}$, where T is the number of observed steps. The last mode stores traffic values, normalized time of day, and day of week. Following one-based mathematical indexing, the traffic slice is $X_{\text{flow}} = X_{:,:,:,1} \in \mathbb{R}^{B\times T\times N}$. The target traffic tensor is $Y_{\text{flow}} \in \mathbb{R}^{B\times T'\times N\times 1}$ for T' future steps, and the known future temporal covariates are $Y_{\text{cov}} \in \mathbb{R}^{B\times T'\times N\times 2}$. Let $A_{\text{phy}} \in \mathbb{R}^{N\times N}$ be the physical sensor-adjacency matrix derived from road-network connectivity. A scalar normalization map $s(\cdot)$, fitted only to training inputs, transforms the traffic flow channel. With trainable parameter vector θ, EviPhys is written as

$$\widehat{Y}^{\,s} = f_\theta(X, Y_{\text{cov}}, A_{\text{phy}}), \quad \widehat{Y} = s^{-1}(\widehat{Y}^{\,s}), \tag{1}$$

where $\widehat{Y}^{\,s} \in \mathbb{R}^{B\times T'\times N\times 1}$ is the normalized traffic flow prediction and $\widehat{Y}$ is the prediction on the original flow scale.

3.2 MULTIMODAL REPRESENTATION FUSION MODULE (MRFM)

MRFM first transforms historical traffic sequences into structured visual representations and then integrates image and text information to extract node-aware macroscopic traffic semantics.

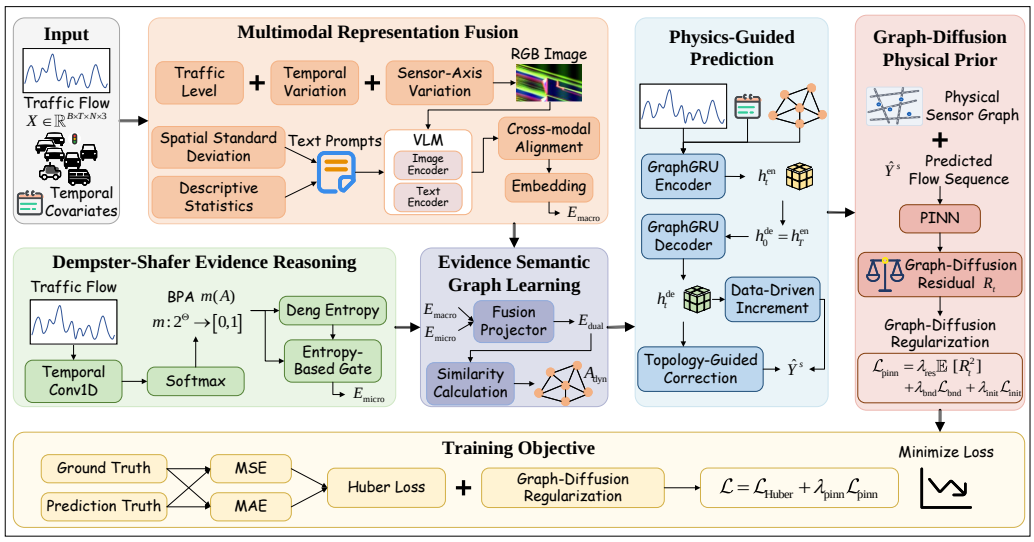


Figure 2: Overview of EviPhys. MRFM extracts macroscopic multimodal traffic semantics, DSERM constructs uncertainty-sensitive local evidence representations, ESGLM integrates both representations to construct a directed graph, and PGPM performs physics-guided graph recurrent forecasting.

3.2.1 RGB Traffic Image Construction

Raw traffic sequences can precisely capture numerical variations, but one-dimensional representations of traffic sequences hinder the direct identification of congestion boundaries and abrupt variations across sensors. For each sample, we therefore transpose the traffic channel into a Hovmoller field $U^{(b)} \in \mathbb{R}^{N\times T}$ and apply min-max normalization to obtain $\widetilde{U}^{(b)} \in [0,1]^{N\times T}$. First-order differences along the temporal and sensor axes are then used to describe temporal changes and cross-sensor variations, respectively. The resulting three-channel representation is

$$I^{(b)} = \mathrm{Resize}_{224\times 224}([\widetilde{U}^{(b)}, \mathrm{clip}(2|\Delta_t \widetilde{U}^{(b)}|, 0, 1), \mathrm{clip}(2|\Delta_s \widetilde{U}^{(b)}|, 0, 1)]). \tag{2}$$

The three channels encode the overall traffic level, temporal variations, and cross-sensor boundaries, providing complementary views of macroscopic traffic patterns.

3.2.2 Joint Image-Text Semantic Extraction

The three-channel traffic image captures macroscopic evolution patterns but lacks explicit descriptions of the current traffic state. In this section, two statistics are extracted: global variance characterizes temporal fluctuations, while the standard deviation of node-wise temporal means measures spatial imbalance. These statistics are then converted into text prompts describing traffic stability.

The image $I^{(b)}$ and prompt $p^{(b)}$ are jointly encoded by a locally deployed ViLT model. ViLT is frozen, and only the projection layer is optimized to obtain stable representations while reducing computational cost. The `[CLS]` representation $h_{\mathrm{mm}}^{(b)} \in \mathbb{R}^{768}$ is used as the sample semantic feature. Since $h_{\mathrm{mm}}^{(b)}$ captures global traffic semantics rather than node-specific information, this semantic feature is projected into the node-embedding space, broadcast to all nodes, and fused with learnable node embeddings:

$$E_{\mathrm{macro}}^{(b,i)} = E_{\mathrm{node}}^{(i)} + \mathrm{LayerNorm}(\mathrm{GELU}(W_m h_{\mathrm{mm}}^{(b)} + b_m)), \tag{3}$$

where $E_{\mathrm{node}}^{(i)} \in \mathbb{R}^d$ preserves the persistent identity of sensor i, while $W_m \in \mathbb{R}^{d\times 768}$ and $b_m \in \mathbb{R}^d$ project the global semantic feature to dimension d. Stacking all samples and nodes produces $E_{\mathrm{macro}} \in \mathbb{R}^{B\times N\times d}$. The broadcast semantic term provides every node with the current network-wide traffic context, whereas the base embedding preserves long-term differences among sensors.

3.3 DEMPSTER-SHAFER EVIDENCE REASONING MODULE (DSERM)

The macroscopic representation captures the overall traffic regime but cannot characterize the uncertainty of individual sensor observations. DSERM models each sensor's historical window as an evidence distribution and derives a local representation before graph construction.

Let $\Theta = \{e_1, e_2, \ldots, e_M\}$ denote a frame of discernment. Its power set contains $K = |2^\Theta| = 2^M$ propositions. For sample b and sensor i, two 1×1 convolutions with an intermediate GELU activation map the complete historical window to logits $\ell^{(b,i)} \in \mathbb{R}^{K-1}$ over all nonempty subsets. A Softmax operation then produces the BPA $m_{b,i}(A)$ for each $A \in 2^\Theta \setminus \{\emptyset\}$, with $m_{b,i}(\emptyset) = 0$. The resulting mass vector is $\mu^{(b,i)} = [m_{b,i}(A)]_{A \in 2^\Theta} \in \mathbb{R}^K$.

To quantify the uncertainty of the evidence distribution, Deng entropy is computed as $H_D^{(b,i)} = -\sum_{A \in 2^\Theta, A \neq \emptyset} m_{b,i}(A) \log \frac{m_{b,i}(A)}{2^{|A|}-1}$, where $|A|$ denotes the cardinality of subset A. The entropy value is used to control the contribution of local evidence through a learnable gate:

$$E_{\text{micro}}^{(b,i)} = g(H_D^{(b,i)})(W_e \mu^{(b,i)} + b_e), \tag{4}$$

where $g(\cdot)$ is a perceptron with sigmoid activation, and W_e and b_e are learnable parameters. Stacking all sensors yields $E_{\text{micro}} \in \mathbb{R}^{B \times N \times d}$, which provides ESGLM with local traffic representations.

3.4 EVIDENCE SEMANTIC GRAPH LEARNING MODULE (ESGLM)

ESGLM combines the network-wide traffic context in E_{macro} with the local evidential state in E_{micro} to construct a dynamic graph. The two representations are first fused as $E_{\text{dual}} = \phi_d([E_{\text{macro}} \| E_{\text{micro}}]) \in \mathbb{R}^{B \times N \times d}$, where $\|$ denotes concatenation and ϕ_d maps the resulting $2d$-dimensional features back to dimension d using a linear layer, GELU, and layer normalization. For sample b, the dynamic graph is then computed by

$$S^{(b)} = \text{ReLU}(E_{\text{dual}}^{(b)} E_{\text{dual}}^{(b)\top}) + I_N + \epsilon, \quad A_{\text{dyn}}^{(b)} = \text{RowNorm}(S^{(b)}), \tag{5}$$

where $S^{(b)} \in \mathbb{R}^{N \times N}$ is the nonnegative node-similarity matrix, I_N adds self-connections, $\epsilon > 0$ ensures numerical stability, and RowNorm normalizes each row. Stacking all samples gives $A_{\text{dyn}} \in \mathbb{R}^{B \times N \times N}$. Thus, the learned propagation weights adapt to global traffic conditions and local evidence, while A_{phy} serves as an independent prior preserving physical sensor connectivity.

3.5 PHYSICS-GUIDED PREDICTION MODULE (PGPM)

PGPM encodes historical observations with the learned graph, predicts future flow increments, and regularizes predictions via a graph-diffusion physical prior on the sensor graph.

3.5.1 GRAPH RECURRENT SPATIO-TEMPORAL ENCODER

At historical step t, the concatenated time-of-day and day-of-week embedding $q_t \in \mathbb{R}^{B \times N \times d_q}$ and the reference-time embedding $q_{\text{ref}} \in \mathbb{R}^{B \times N \times d_q}$ adapt the node embedding used to generate recurrent parameters:

$$\widetilde{E}_t = \gamma_t \odot E_{\text{dual}} + (1 - \gamma_t) \odot W_s h_{t-1}^{\text{en}}, \quad \gamma_t = \sigma([q_{\text{ref}} \| q_t] W_\gamma), \tag{6}$$

where $E_{\text{dual}} \in \mathbb{R}^{B \times N \times d}$ is the node representation from ESGLM, and $h_{t-1}^{\text{en}} \in \mathbb{R}^{B \times N \times d_h}$ is the previous encoder state. The learnable matrices $W_\gamma \in \mathbb{R}^{2d_q \times d}$ and $W_s \in \mathbb{R}^{d \times d_h}$ produce the interpolation gate $\gamma_t \in (0,1)^{B \times N \times d}$ and project the hidden state to dimension d, respectively. The resulting $\widetilde{E}_t \in \mathbb{R}^{B \times N \times d}$ combines spatial information with the current recurrent state. The operators σ, $\|$, and $\odot$ denote sigmoid activation, concatenation, and element-wise multiplication.

The adaptive embedding $\widetilde{E}_t$ generates node-specific graph-convolution parameters from shared parameter pools. Within each recurrent update, the graph convolution combines each node's own features with one-hop information propagated through A_{dyn}. Replacing the affine transformations of a standard GRU with this graph operation gives $h_t^{\text{en}} = \text{GraphGRU}_{\text{enc}}(X_t, h_{t-1}^{\text{en}}; A_{\text{dyn}}, \widetilde{E}_t)$, where $X_t \in \mathbb{R}^{B \times N \times 3}$ is the historical input, $A_{\text{dyn}} \in \mathbb{R}^{B \times N \times N}$ stacks the graphs, and $h_t^{\text{en}} \in \mathbb{R}^{B \times N \times d_h}$ is the updated hidden state. Starting from zero, the encoder processes all T historical steps, and the final state h_T^{en} summarizes the spatio-temporal history and initializes the decoder.

3.5.2 Flow Increment Spatio-Temporal Decoder

Let $g_0 = X_{:,T,:,1} \in \mathbb{R}^{B\times N\times 1}$ denote the last observed normalized traffic flow. At future step t, the decoder input $d_t \in \mathbb{R}^{B\times N\times 4}$ concatenates the previous flow estimate $g_{t-1} \in \mathbb{R}^{B\times N\times 1}$, the known time-of-day and day-of-week covariates $Y_{\mathrm{cov},t} \in \mathbb{R}^{B\times N\times 2}$, and the fixed anchor g_0. These components support autoregressive prediction, identify the future temporal context, and reduce drift from the latest observation, respectively. The decoder state is updated by $h_t^{\mathrm{de}} = \mathrm{GraphGRU}_{\mathrm{dec}}\left(d_t, h_{t-1}^{\mathrm{de}}; A_{\mathrm{dyn}}, E_{\mathrm{dual}}\right)$, where $h_{t-1}^{\mathrm{de}}, h_t^{\mathrm{de}} \in \mathbb{R}^{B\times N\times d_h}$ are the previous and updated decoder states. The decoder uses A_{dyn} to propagate information across dynamically related sensors and E_{dual} to incorporate the fused global and local evidential representations into the recurrent update. The decoder is initialized by the final encoder state, i.e., $h_0^{\mathrm{de}} = h_T^{\mathrm{en}}$, thereby transferring the encoded historical spatio-temporal context to the future prediction process.

The updated state is passed to two complementary branches. The data-driven branch combines the updated state with a learned forecast-horizon embedding to estimate a nonlinear flow change, denoted by $\Delta g_{\mathrm{data}}^{(t)} \in \mathbb{R}^{B\times N\times 1}$. The topology-guided branch compares each sensor's previous estimate with the weighted average of its physically connected neighbors under A_{phy}. A gated correction network converts this discrepancy, together with the decoder state and horizon information, into a bounded correction $\Delta g_{\mathrm{phy}}^{(t)} \in \mathbb{R}^{B\times N\times 1}$. The two terms capture data-dependent variation and topology-consistent adjustment, respectively. The prediction is updated by

$$\widehat{Y}_t^{\,s} = g_t = g_{t-1} + \Delta g_{\mathrm{data}}^{(t)} + \Delta g_{\mathrm{phy}}^{(t)}, \tag{7}$$

where $g_t \in \mathbb{R}^{B\times N\times 1}$ is the decoder state in the flow space, and $\widehat{Y}_t^{\,s} \in \mathbb{R}^{B\times N\times 1}$ is the normalized prediction at step t; the identity output mapping gives $\widehat{Y}_t^{\,s} = g_t$. Stacking all stepwise outputs gives $\widehat{Y}^{\,s} \in \mathbb{R}^{B\times T'\times N\times 1}$, which is transformed by $s^{-1}(\cdot)$ to obtain $\widehat{Y}$ on the original traffic flow scale.

3.5.3 Training Objective

A graph-diffusion physical prior acts as a training-time regularizer for the predicted sequence $\{g_t\}_{t=1}^{T'}$, encouraging temporal changes between adjacent predictions to follow flow propagation over the physical sensor graph. Let $A_{\mathrm{rw}} \in \mathbb{R}^{N\times N}$ denote the nonnegative, row-normalized physical graph without self-loops. Given the last observed normalized flow g_0, the graph-diffusion residual at step t is $R_t = \frac{g_t - g_{t-1}}{\Delta t} + \kappa(I_N - A_{\mathrm{rw}})g_{t-1}$, where I_N is the identity matrix, $\Delta t > 0$ is the prediction interval, and $\kappa > 0$ is a learned diffusion coefficient. The first term captures temporal flow variation, while the second term measures deviations from physically connected sensors. Therefore, $R_t \in \mathbb{R}^{B\times N\times 1}$ quantifies the deviation of the predicted evolution from the graph-diffusion prior. The graph-diffusion regularization loss is $\mathcal{L}_{\mathrm{pinn}} = \lambda_{\mathrm{res}}\mathbb{E}_{b,t,i}[(R_t^{(b,i)})^2] + \lambda_{\mathrm{bnd}}\mathcal{L}_{\mathrm{bnd}}$, where $\mathbb{E}_{b,t,i}$ averages over samples, prediction steps, and sensors. The residual term penalizes graph-inconsistent flow evolution, and $\mathcal{L}_{\mathrm{bnd}}$ is the mean-squared prediction error at topological boundary sensors. Since $g_0 = X_{:,T,:,1}$, the initial condition is satisfied by construction without introducing an additional loss term. The nonnegative coefficients λ_{res} and λ_{bnd} balance the two terms.

The Huber loss $\mathcal{L}_{\mathrm{Huber}}$ measures prediction errors on the original traffic scale. The final objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{Huber}} + \lambda_{\mathrm{pinn}}\mathcal{L}_{\mathrm{pinn}}, \tag{8}$$

where λ_{pinn} controls the contribution of the graph-diffusion regularizer. Teacher forcing optimizes prediction loss, while free-running trajectories regularize graph diffusion. See Appendix A for theory and Appendix B for training and complexity.

4 Experiments

This section evaluates EviPhys under various traffic prediction scenarios, covering performance, sensitivity, ablation, physical prior selection, and efficiency.

4.1 Experimental Settings

This section describes the experimental setup for evaluating EviPhys. The datasets, baseline methods, and evaluation metrics are introduced to ensure a fair and reproducible comparison.

Table 1: Comparison of traffic flow prediction performance between EviPhys and baselines on four PEMS benchmark datasets. Bold and underline denote the best and second-best results, respectively.

Methods	PEMS03			PEMS04			PEMS07			PEMS08		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
FC-LSTM (2014)	21.33	35.11	23.33%	26.77	40.65	18.23%	29.98	45.94	13.20%	23.09	35.17	14.99%
TCN (2018)	19.32	33.55	23.33%	23.22	37.26	15.59%	32.72	42.23	14.26%	22.72	35.79	14.03%
STGCN (2018)	17.49	30.12	17.15%	22.70	35.55	14.59%	25.38	38.78	11.08%	18.02	27.83	11.40%
DCRNN (2018)	15.63	27.26	15.89%	19.63	31.28	13.57%	21.31	34.28	9.15%	15.23	24.25	10.28%
GWNet (2019)	14.60	**25.37**	14.59%	19.12	30.65	12.88%	20.63	33.74	8.67%	14.61	23.48	9.38%
AGCRN (2020)	16.01	27.93	15.37%	19.46	31.21	12.95%	21.48	34.95	9.20%	15.69	24.69	10.35%
STSGCN (2020)	17.48	29.21	16.78%	21.19	33.65	13.90%	24.26	39.03	10.21%	17.13	26.80	10.96%
STFGNN (2021)	16.77	28.34	16.30%	19.83	31.88	13.02%	22.07	35.80	9.21%	16.64	26.22	10.60%
STID (2022)	15.23	27.05	16.27%	18.37	**29.87**	12.93%	19.60	32.73	8.30%	14.18	23.35	9.23%
MegaCRN (2023)	14.57	25.87	14.75%	18.69	30.54	12.71%	19.70	32.72	8.29%	14.83	23.92	9.64%
HimNet (2024)	15.34	27.50	15.82%	18.17	29.92	12.03%	19.30	32.77	8.05%	13.56	23.36	9.02%
PDFormer (2023)	14.94	25.39	15.82%	18.36	30.03	12.00%	19.97	32.95	8.55%	13.58	23.41	9.05%
STDN (2025)	15.54	27.28	16.86%	18.40	30.22	12.50%	20.65	34.77	10.98%	14.25	24.39	10.67%
PDG2Seq (2025)	14.62	25.47	14.88%	18.24	30.08	12.09%	19.28	33.04	8.07%	13.60	23.37	8.99%
RAST (2026)	15.20	25.81	16.12%	18.39	29.93	12.43%	19.52	32.73	8.23%	14.16	23.33	9.27%
EviPhys	**14.31**	25.43	**14.33%**	**17.95**	**29.87**	**11.91%**	**18.66**	**32.46**	**7.70%**	**12.99**	**22.45**	**8.79%**

4.1.1 DATASETS

Six traffic prediction datasets are adopted for evaluation, including four PEMS benchmarks and two large-scale datasets from LargeST (Liu et al., 2023). The PEMS datasets are collected from the California Performance Measurement System with a 5-minute sampling interval. The LargeST datasets, including San Diego (SD) and Greater Bay Area (GBA), contain 716 and 2,352 sensors, respectively, providing larger traffic networks than the PEMS benchmarks for assessing model scalability. Detailed descriptions are provided in Appendix C.

4.1.2 BASELINES

To comprehensively evaluate EviPhys, we compare EviPhys with representative traffic prediction baselines covering temporal, graph-based, adaptive graph, and recent spatio-temporal approaches. FC-LSTM (Sutskever et al., 2014) and TCN (Bai et al., 2018) model temporal dependencies, while STGCN (Yu et al., 2018), ASTGCN (Guo et al., 2019), DCRNN (Li et al., 2018), GWNet (Wu et al., 2019), AGCRN (Bai et al., 2020), STSGCN (Song et al., 2020), and STFGNN (Li & Zhu, 2021) employ graph-based spatio-temporal modeling. STID (Shao et al., 2022), MegaCRN (Jiang et al., 2023b), and HimNet (Dong et al., 2024) capture identity information, adaptive structures, and heterogeneity. We further include RPMixer (Yeh et al., 2024), PDFormer (Jiang et al., 2023a), STDN (Cao et al., 2025), PDG2Seq (Fan et al., 2025), and RAST (Ruan et al., 2026) as recent advanced spatio-temporal baselines. Detailed descriptions are provided in Appendix D.

4.1.3 METRICS AND PARAMETER SETTINGS

Prediction performance is evaluated using mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). All metrics are computed after inverse normalization on the original traffic-flow scale, with zero-valued targets excluded following the evaluation implementation. Lower values indicate better performance. EviPhys is implemented in PyTorch and trained on an NVIDIA H100 GPU using Adam and the Huber loss. The best validation checkpoint is retained for test evaluation. The definitions of the evaluation metrics and additional implementation details are provided in Appendix E.

4.2 OVERALL PERFORMANCE

The performance of EviPhys is evaluated against representative traffic flow prediction methods on four PEMS benchmarks and two large-scale LargeST datasets. As shown in Tables 1 and 2, EviPhys achieves competitive or superior performance across different traffic networks.

On the four PEMS datasets, EviPhys achieves the best results on most evaluation metrics. Compared with the second-best results, EviPhys reduces the average MAE, RMSE, and MAPE by 2.60%, 1.08%, and 2.17%, respectively. These results demonstrate the effectiveness of EviPhys in model-

Table 2: Performance comparison on the LargeST datasets under different prediction horizons. Bold and underline denote the best and second-best results, respectively.

Dataset	Methods	Horizon 3			Horizon 6			Horizon 12			Average		
		MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
SD	STGCN (2018)	17.45	29.99	12.42%	19.55	33.69	13.68%	23.21	41.23	16.32%	19.67	34.14	13.86%
	DCRNN (2018)	17.14	27.47	11.12%	20.99	33.29	13.95%	26.99	42.86	18.67%	21.03	33.37	14.13%
	ASTGCN (2019)	19.56	31.33	12.18%	24.13	37.95	15.38%	30.96	49.17	21.98%	23.70	37.63	15.65%
	RPMixer (2024)	18.54	30.33	11.81%	24.55	40.04	16.51%	35.90	58.31	27.67%	25.25	42.56	17.64%
	RAST (2026)	15.84	26.41	10.15%	18.55	31.56	12.13%	22.18	39.43	15.38%	18.39	31.96	12.19%
	EviPhys	**14.21**	**24.07**	**8.78%**	**16.56**	**28.74**	**10.62%**	**20.55**	**36.90**	**13.86%**	**16.63**	**29.49**	**10.78%**
GBA	STGCN (2018)	21.05	34.51	16.42%	23.63	38.92	18.35%	26.87	44.45	21.92%	23.42	38.57	18.46%
	DCRNN (2018)	18.71	30.36	14.72%	23.06	36.16	20.45%	29.85	46.06	29.93%	23.13	36.35	20.84%
	ASTGCN (2019)	21.46	33.86	17.24%	26.96	41.38	24.22%	34.29	52.44	32.53%	26.47	40.99	23.65%
	RPMixer (2024)	20.31	33.34	15.64%	26.95	44.02	22.75%	39.66	66.44	37.35%	27.77	47.72	23.87%
	RAST (2026)	17.71	29.29	13.72%	20.86	34.26	16.63%	24.97	41.33	21.25%	20.64	34.47	16.63%
	EviPhys	**16.36**	**28.40**	**10.71%**	**19.16**	**32.88**	**12.98%**	**22.92**	**38.93**	**16.60%**	**19.03**	**32.95**	**13.06%**

Table 3: Sensitivity analysis of batch size and hidden dimension on the prediction performance of EviPhys. "–" indicates that the corresponding configuration exceeded the available GPU memory due to memory-intensive dynamic graph construction and node-level intermediate representations.

Hyperparameters	Values	PEMS03			PEMS04			PEMS07			PEMS08		
		MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
Batch Size	4	**14.31**	25.85	14.80%	18.17	30.27	12.56%	18.68	32.63	7.70%	13.13	22.66	8.97%
	8	**14.31**	**25.43**	**14.33%**	18.13	30.23	12.09%	**18.66**	**32.46**	7.70%	13.05	22.54	8.76%
	16	14.54	25.74	14.86%	18.03	**29.79**	12.19%	18.67	32.53	**7.68%**	13.02	22.47	**8.61%**
	32	14.44	25.57	14.62%	**17.95**	29.87	**11.91%**	–	–	–	**12.99**	**22.45**	8.79%
Hidden Dim	32	14.39	**24.97**	14.72%	18.18	29.93	12.20%	18.80	32.71	7.80%	13.03	22.49	**8.60%**
	64	**14.31**	25.43	**14.33%**	**17.95**	**29.87**	**11.91%**	**18.66**	**32.46**	**7.70%**	**12.99**	**22.45**	8.79%
	96	14.41	25.68	14.62%	–	–	–	18.80	32.66	7.73%	13.13	22.60	8.74%

ing spatio-temporal dependencies under different traffic conditions. On the large-scale SD and GBA datasets, EviPhys achieves the best performance across all reported prediction horizons and evaluation metrics. On SD, EviPhys reduces the average MAE, RMSE, and MAPE by 9.57%, 7.73%, and 11.57%, respectively, compared with RAST. On GBA, EviPhys achieves corresponding reductions of 7.80%, 4.41%, and 21.47%. The consistent improvements across both datasets demonstrate the scalability and effectiveness of EviPhys on large-scale traffic networks. Moreover, Appendix F provides a case study on PEMS04 traffic prediction, while Appendix G visualizes the learned contextual representations.

4.3 Parameter Sensitivity Analysis

The sensitivity of EviPhys to batch size and hidden dimension is evaluated on four PEMS datasets, as shown in Table 3. EviPhys maintains stable performance across different batch sizes, while the optimal setting varies with dataset characteristics. Batch size 8 is selected for PEMS03 and PEMS07, whereas batch size 32 is adopted for PEMS04 and PEMS08. For the hidden dimension, a value of 64 achieves the best overall performance across most datasets. Increasing the dimension to 96 provides no consistent improvement, while a dimension of 32 yields better results only on a few individual metrics. Therefore, 64 is selected as the default hidden dimension.

4.4 Ablation Study

The contributions of the VLM semantic branch, evidence modeling branch, and dynamic graph learning are evaluated on four PEMS datasets. The w/o VLM, w/o Evi, and w/o Dyn variants remove the multimodal semantic representation, evidence modeling components, and dynamic graph learning, respectively. As shown in Figure 3, EviPhys achieves the best overall performance across all datasets. Removing either the VLM or evidence branch degrades performance, demonstrating their complementary roles, while the w/o Dyn variant causes the largest degradation, highlighting the importance of dynamic spatial dependency learning.

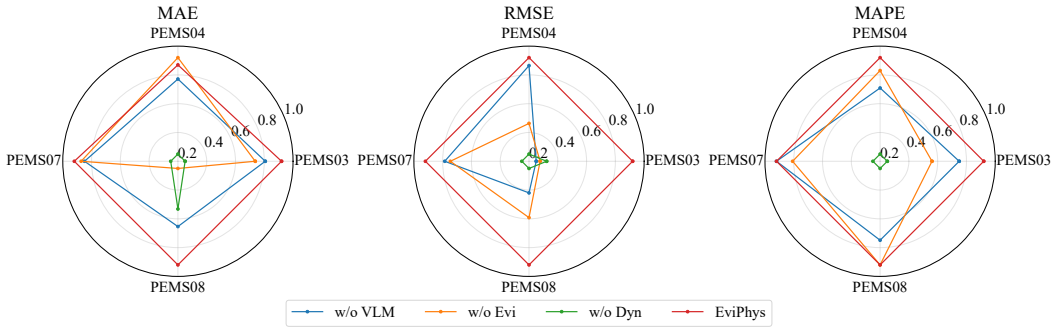


Figure 3: Radar visualization of the ablation results on PEMS datasets. The normalized MAE, RMSE, and MAPE values demonstrate the contributions of different components.

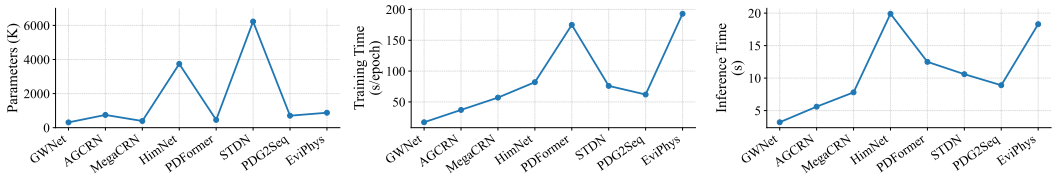


Figure 4: Computational efficiency comparison among different methods on the PEMS04 dataset.

4.5 PHYSICAL PRIOR COMPARISON

To evaluate the influence of different physical priors, the graph-diffusion physical prior in EviPhys is compared with the Lorenz system, wave equation, and Burgers' equation on four PEMS datasets, while all other model settings remain unchanged. All physics-informed variants achieve competitive prediction performance, whereas the graph-diffusion formulation obtains the best overall results across most metrics. The results indicate that directly coupling temporal flow evolution with propagation over the physical road graph provides a more suitable physical prior for traffic flow prediction. Detailed formulations and experimental results are provided in Appendix H.

5 EFFICIENCY COMPARISON

The computational efficiency of EviPhys is evaluated on the PEMS04 dataset using a single NVIDIA H100 GPU. As shown in Figure 4, EviPhys contains 877K parameters, achieving a moderate model size among the compared methods. Although multimodal fusion, evidence reasoning, dynamic graph construction, and physics-guided learning introduce additional training overhead, the inference time remains comparable to existing spatio-temporal prediction models since the physics-guided graph-diffusion regularization is only applied during training. These results demonstrate that EviPhys achieves a balance between representation capability and computational efficiency. Additional robustness evaluation and statistical analysis are provided in Appendices I and J.

6 CONCLUSION

This paper presents EviPhys, an evidence-aware multimodal framework with physics-guided learning for traffic flow prediction. The framework integrates multimodal traffic semantics, evidential uncertainty modeling, dynamic graph learning, and a graph-diffusion physical prior to improve traffic representation and long-horizon prediction. Despite these advantages, several limitations remain. First, the multimodal semantic branch relies on a pretrained vision-language model, which introduces additional computational overhead. Second, the current graph-diffusion physical prior is simplified and cannot fully characterize the complex nonlinear traffic dynamics. Future work will explore lightweight multimodal semantic modeling and more realistic graph-diffusion physical priors, while extending evaluation to more diverse real-world traffic scenarios.

AI USE STATEMENT

In this work, we used generative AI tools to improve the readability and language quality of the manuscript. We did not use generative AI tools for other tasks requiring disclosure, and the remaining required-disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. AI-assisted suggestions, including grammar corrections and writing improvements, were manually reviewed and validated by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study does not involve human subjects, personally identifiable information, or sensitive data. All experiments are conducted on publicly available datasets, which are widely used in transportation research and do not contain any personally identifiable information.

REPRODUCIBILITY STATEMENT

The reproducibility of EviPhys is supported by detailed descriptions of the model architecture, training procedure, datasets, evaluation metrics, and experimental settings in the main paper and supplementary materials. The experiments use publicly available PEMS and LargeST datasets, including PEMS03, PEMS04, PEMS07, PEMS08, SD, and GBA. Detailed hyperparameter settings and implementation configurations are provided to facilitate replication of the reported results. The source code, including data preprocessing, model implementation, training, and evaluation scripts, is maintained in an anonymous repository, and the complete implementation has been organized to support reproduction of the reported experiments.

REFERENCES

Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. Adaptive graph convolutional recurrent network for traffic forecasting. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17804–17815, 2020.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Lingxiao Cao, Bin Wang, Guiyuan Jiang, Yanwei Yu, and Junyu Dong. Spatiotemporal-aware trend-seasonality decomposition network for traffic flow forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11463–11471, 2025.

Wenxia Chang, Chao Zhang, Wentao Li, and Deyu Li. MIGC-CMamba: Cross-domain Mamba with multi-scale imaging and granular-ball computing for traffic flow prediction. In *Proceedings of the ACM Web Conference 2026*, pp. 7167–7177, 2026.

Geng Chen, Xiantao Du, Fei Shen, Qingtian Zeng, and Yu-Dong Zhang. Phased spatial-temporal targeted networks based on transformer and data augmentation for cellular traffic prediction. *IEEE Internet of Things Journal*, 13(7):13573–13589, 2026a.

Jinwen Chen, Hao Miao, Chenxi Liu, Yan Zhao, and Kai Zheng. VisionST: Coordinating cross-modal traffic prediction with interactive geo-image encoding. In *Proceedings of the ACM Web Conference 2026*, pp. 7307–7317, 2026b.

Andong Cong, Shijun Wang, Yuhong Jin, Haiming Yi, Zeyuan Chang, Yifan Jiang, Lei Hou, and Nasser A Saeed. Respecting causality segmental time-domain training accelerates physics-informed DeepONet on time-dependent PDEs. *Neurocomputing*, 694:133915, 2026.

A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 38(2):325–339, 1967.

Yong Deng. Deng entropy. *Chaos, Solitons & Fractals*, 91:549–553, 2016.

Zheng Dong, Renhe Jiang, Haotian Gao, Hangchen Liu, Jinliang Deng, Qingsong Wen, and Xuan Song. Heterogeneity-informed meta-parameter learning for spatiotemporal time series forecasting. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 631–641, 2024.

Wenchang Duan, Zhenguo Gao, Jiwan He, and Jinguo Xian. Bayesian critique-tune-based reinforcement learning with adaptive pressure for multi-intersection traffic signal control. *IEEE Transactions on Intelligent Transportation Systems*, 26(10):14968–14983, 2025.

Jin Fan, Wenchao Weng, Qikai Chen, Huifeng Wu, and Jia Wu. PDG2Seq: Periodic dynamic graph to sequence model for traffic flow prediction. *Neural Networks*, 183:106941, 2025.

Abdul Joseph Fofanah, Lian Wen, and David Chen. PIMCST: Physics-informed multi-phase consensus and spatio-temporal few-shot learning for traffic flow forecasting. *arXiv preprint arXiv:2602.01936*, 2026.

Atmadip Goswami, Chayan Banerjee, Dhivya Bharathi, and Lelitha Vanajakshi. Adaptive physics-informed machine learning for bus travel time prediction: Solving the peak prediction problem. *Applied Soft Computing*, 200:115435, 2026.

Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 922–929, 2019.

Jiawei Jiang, Chengkai Han, Wayne Xin Zhao, and Jingyuan Wang. PDFormer: Propagation delay-aware dynamic long-range transformer for traffic flow prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 4365–4373, 2023a.

Renhe Jiang, Zhaonan Wang, Jiawei Yong, Puneet Jeph, Quanjun Chen, Yasumasa Kobayashi, Xuan Song, Shintaro Fukushima, and Toyotaro Suzumura. Spatio-temporal meta-graph learning for traffic forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8078–8086, 2023b.

Kai Jungel, Dario Paccagnan, Axel Parmentier, and Maximilian Schiffer. WardropNet: Traffic flow predictions via equilibrium-augmented learning. In *Proceedings of the International Conference on Learning Representations*, 2025.

Amjid Khan and Jawad Ali. Artificial intelligence-based smart traffic management system selection in urban cities using T-spherical fuzzy Z-numbers and neutrality aggregation operators. *Engineering Applications of Artificial Intelligence*, 182:115992, 2026.

Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 5583–5594, 2021.

Donghe Li, Jiacheng Zhang, Yixian Liu, Jiaming Zhang, Huan Xi, and Qingyu Yang. Embedding fluid dynamics into neural networks: Toward interpretable traffic flow prediction via physics-informed learning. *IEEE Internet of Things Journal*, 12(16):34528–34547, 2025a.

Mengzhang Li and Zhanxing Zhu. Spatial-temporal fusion graph neural networks for traffic flow forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 4189–4196, 2021.

Rui Li, Chao Zhang, Deyu Li, Wentao Li, and Jianming Zhan. Improved evidential three-way decisions in incomplete multi-scale information systems. *International Journal of Approximate Reasoning*, 181:109417, 2025b.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *Proceedings of the International Conference on Learning Representations*, 2018.

Xu Liu, Yutong Xia, Yuxuan Liang, Junfeng Hu, Yiwei Wang, Lei Bai, Chao Huang, Zhenguang Liu, Bryan Hooi, and Roger Zimmermann. LargeST: A benchmark dataset for large-scale traffic forecasting. In *Advances in Neural Information Processing Systems*, volume 36, pp. 75354–75371, 2023.

Gordon F Newell. A simplified theory of kinematic waves in highway traffic, part I: General theory. *Transportation Research Part B: Methodological*, 27(4):281–287, 1993.

Bei Pan, Ugur Demiryurek, and Cyrus Shahabi. Utilizing real-world transportation data for accurate traffic prediction. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*, pp. 595–604, 2012.

Arabinda Panda, Debi Prosad Dogra, and Partha Pratim Dey. Physics-constrained dynamic graph and multi-resolution temporal modeling for traffic forecasting. *Information Sciences*, 757:123911, 2026.

Tommaso Pellegrini and Matteo Brunelli. Uncertainty measures in evidence theory: A genealogy. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 56(3):1973–1987, 2026.

Jonah Philion, Xue Bin Peng, and Sanja Fidler. Trajeglish: Traffic modeling as next-token prediction. In *Proceedings of the International Conference on Learning Representations*, 2024.

Weilin Ruan, Xilin Dang, Ziyu Zhou, Sisuo Lyu, and Yuxuan Liang. A retrieval augmented spatio-temporal framework for traffic prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 39163–39172, 2026.

Glenn Shafer. *A mathematical theory of evidence*. Princeton University Press, Princeton, NJ, 1976.

Zezhi Shao, Zhao Zhang, Fei Wang, Wei Wei, and Yongjun Xu. Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 4454–4458, 2022.

Chao Song, Youfang Lin, Shengnan Guo, and Huaiyu Wan. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 914–921, 2020.

James H Stock and Mark W Watson. Vector autoregressions. *Journal of Economic Perspectives*, 15 (4):101–115, 2001.

Xinyu Su, Majid Sarvi, Feng Liu, Egemen Tanin, and Jianzhong Qi. Generalising traffic forecasting to regions without traffic observations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 15761–15769, 2026.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27, pp. 3104–3112, 2014.

Chen Tang and Fuyuan Xiao. Complex Deng entropy for uncertainty measure in complex evidence theory. *Engineering Applications of Artificial Intelligence*, 141:109696, 2025.

Zhen Tian, Yaqiong Zhang, Zhihao Lin, Fujiang Yuan, Yijun Lu, Wangjie Lang, Xinyu Wang, Ning Lyu, Zhiguo Tao, Kaijie Chen, and Aaron Wang. Physics-guided multimodal multi-agent learning for intelligent transportation systems. In *Proceedings of the International Conference on Learning Representations*, 2026.

Anna Wang, Chao Zhang, Mingwei Lin, Junbo Zhang, Zeshui Xu, Wentao Li, Pengfei Zhang, and Oscar Castillo. Towards vision-spatiotemporal fusion in traffic forecasting: A survey on cross-modal alignment. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 8032–8040, 2026a.

Bingjie Wang, Chao Zhang, Wentao Li, and Deyu Li. MF3: Multimodal federated learning with dual-path Mamba-Transformer for metro flow prediction. In *Proceedings of the ACM Web Conference 2026*, pp. 5451–5462, 2026b.

Zhaoyang Wei, Guoliang Wang, Guohua Gao, Yanchao Hao, Mingda Li, Wenchao Ding, Xi Chen, Shizhu He, and Xuehui Yu. Seeing is believing: Grounding long-video understanding in spatio-temporal visual evidence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10584–10592, 2026.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph WaveNet for deep spatial-temporal graph modeling. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 1907–1913, 2019.

Bairan Xiang, Chenguang Zhao, and Huan Yu. Micro-macro coupled Koopman modeling on graph for traffic flow prediction. In *Proceedings of the International Conference on Learning Representations*, 2026.

Chin-Chia Michael Yeh, Yujie Fan, Xin Dai, Uday Singh Saini, Vivian Lai, Prince Osei Aboagye, Junpeng Wang, Huiyuan Chen, Yan Zheng, Zhongfang Zhuang, Liang Wang, and Wei Zhang. RPMixer: Shaking up time series forecasting with random projections for large spatial-temporal data. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3919–3930, 2024.

Xiaoyang Yi, Li Peng, Yuru Bao, and Jian Zhang. Bayesian evidence-driven prototype evolution for federated domain adaptation. In *Proceedings of the International Conference on Learning Representations*, 2026.

Chanyoung Yoon, Soobin Yim, Sangbong Yoo, Chanyoung Jung, Hanbyul Yeon, and Yun Jang. V-DCRNN: Virtual network-based diffusion convolutional recurrent neural network for estimating unobserved traffic data. *IEEE Transactions on Intelligent Transportation Systems*, 26(7):10336–10352, 2025.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 3634–3640, 2018.

Tianxiang Zhan, Yuanpeng He, Yong Deng, Zhen Li, Wenjie Du, and Qingsong Wen. Time evidence fusion network: Multi-source view in long-term time series forecasting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(12):11220–11233, 2025.

Shiqi Zhang, Zhen Liu, Chenliang Wo, Jialong Qian, Yonghong Liu, and H Oliver Gao. Spatiotemporal graph neural network for traffic flow prediction based on a two-stage pre-training and fine-tuning framework. *Advanced Engineering Informatics*, 73:104572, 2026a.

Weiwei Zhang, Wei Suo, Jiahao Song, and Wenbo Cao. Physics-informed neural networks (PINNs) as intelligent computing technique for solving partial differential equations: Limitation and future prospects. *Science China Physics, Mechanics & Astronomy*, 69:214602, 2026b.

Yuhang Zhang, Meng Ma, and Ping Wang. TRACER: Physics-guided causal evidence construction for zero-shot traffic anomaly diagnosis. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 6653–6664, 2026c.

Yunhao Zhang, Wenyao Hu, Jiale Zheng, Lujia Pan, and Junchi Yan. MMPD: Diverse time series forecasting via multi-mode patch diffusion loss. In *Proceedings of the International Conference on Learning Representations*, 2026d.

Siru Zhong, Weilin Ruan, Ming Jin, Huan Li, Qingsong Wen, and Yuxuan Liang. Time-VLM: Exploring multimodal vision-language models for augmented time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 78478–78497, 2025.

Chengyang Zhou, Zijian Zhang, Chunxu Zhang, Hao Miao, Yulin Zhang, Kedi Lyu, and Juncheng Hu. FedDis: A causal disentanglement framework for federated traffic prediction. In *Proceedings of the ACM Web Conference 2026*, pp. 7541–7551, 2026.

Yiqing Zou, Hanning Yuan, Qianyu Yang, Ziqiang Yuan, Shuliang Wang, and Sijie Ruan. Meta dynamic graph for traffic flow prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 16584–16592, 2026.

A THEORETICAL ANALYSIS OF GRAPH-DIFFUSION REGULARIZATION

This appendix explains why the graph-diffusion regularization term in EviPhys is useful for multi-step traffic prediction. In long-horizon prediction, the model must generate future states recursively. Therefore, an unreasonable update at an early step may be propagated and amplified at later steps. The proposed physics-guided graph-diffusion regularization constrains this recursive prediction process according to the physical road graph.

For simplicity, we omit the batch dimension and consider one traffic flow channel. Let $g_t \in \mathbb{R}^N$ denote the free-running prediction at prediction step t, and let $A_{\mathrm{rw}} \in \mathbb{R}^{N\times N}$ denote the nonnegative row-normalized physical road graph. The graph-diffusion residual is

$$R_t = \frac{g_t - g_{t-1}}{\Delta t} + \kappa(I_N - A_{\mathrm{rw}})g_{t-1}, \tag{9}$$

where $\Delta t > 0$ is the temporal interval and $\kappa > 0$ is the diffusion coefficient. We define $\eta = \kappa\Delta t$.

If the predicted trajectory exactly follows the graph-diffusion equation, each future node value is formed from a weighted combination of its previous value and its physical neighbors. Thus, the physical update itself does not amplify the largest traffic flow magnitude across prediction steps.

Proposition A.1 (Conditional stability of graph-diffusion prediction)**.** *Assume that* $0 \le \eta \le 1$ *and* $R_t = 0$*. Then the predicted traffic state satisfies*

$$g_t = P_\eta g_{t-1}, \qquad P_\eta = (1-\eta)I_N + \eta A_{\mathrm{rw}}. \tag{10}$$

Moreover, P_η *is row-stochastic, and the prediction at node* i *is*

$$g_{t,i} = (1-\eta)g_{t-1,i} + \eta\sum_{j=1}^{N} A_{\mathrm{rw},ij}g_{t-1,j}. \tag{11}$$

Consequently, the graph-diffusion update is non-expansive:

$$\|g_t\|_\infty \le \|g_{t-1}\|_\infty. \tag{12}$$

Proof. Setting $R_t = 0$ in Equation (9) gives

$$g_t = g_{t-1} - \kappa\Delta t(I_N - A_{\mathrm{rw}})g_{t-1} = \left[(1-\eta)I_N + \eta A_{\mathrm{rw}}\right] g_{t-1}. \tag{13}$$

Since $0 \le \eta \le 1$, $A_{\mathrm{rw},ij} \ge 0$, and A_{rw} is row-normalized, every entry of P_η is nonnegative and

$$\sum_{j=1}^{N} P_{\eta,ij} = (1-\eta) + \eta\sum_{j=1}^{N} A_{\mathrm{rw},ij} = 1. \tag{14}$$

Hence, P_η is row-stochastic, and Equation (11) is a convex combination of the previous node states. Therefore,

$$|g_{t,i}| \le \sum_{j=1}^{N} P_{\eta,ij}|g_{t-1,j}| \le \|g_{t-1}\|_\infty \sum_{j=1}^{N} P_{\eta,ij} = \|g_{t-1}\|_\infty. \tag{15}$$

Taking the maximum over all nodes proves Equation (12). □

Proposition A.1 explains the role of the graph-diffusion physical prior: a prediction that follows the graph-diffusion dynamics cannot create an unreasonably large traffic flow value solely through recursive propagation. This is particularly relevant for long-horizon prediction, where errors may otherwise accumulate step by step.

In practice, the graph-diffusion regularization loss minimizes the residual rather than requiring $R_t = 0$ exactly. The following proposition shows that the residual is the direct source of deviation from the ideal graph-diffusion transition. Therefore, reducing the graph-diffusion residual limits the cumulative deviation from the graph-diffusion prior of the free-running prediction.

Proposition A.2 (Residual-controlled deviation over the prediction horizon)**.** *Under* $0 \leq \eta \leq 1$*, the free-running prediction can be written as*

$$g_t = P_\eta g_{t-1} + \Delta t R_t. \tag{16}$$

Its one-step deviation from the ideal graph-diffusion update is

$$\|g_t - P_\eta g_{t-1}\|_\infty = \Delta t \|R_t\|_\infty. \tag{17}$$

Furthermore, its magnitude over the prediction horizon satisfies

$$\|g_t\|_\infty \leq \|g_0\|_\infty + \Delta t \sum_{\tau=1}^{t} \|R_\tau\|_\infty. \tag{18}$$

Proof. Rearranging Equation (9) yields Equation (16). Subtracting $P_\eta g_{t-1}$ from both sides gives

$$g_t - P_\eta g_{t-1} = \Delta t R_t, \tag{19}$$

which proves Equation (17). By the triangle inequality and Proposition A.1,

$$\|g_t\|_\infty \leq \|P_\eta g_{t-1}\|_\infty + \Delta t \|R_t\|_\infty \leq \|g_{t-1}\|_\infty + \Delta t \|R_t\|_\infty. \tag{20}$$

Applying this inequality recursively from step 1 to step t proves Equation (18). □

Proposition A.2 clarifies the role of the graph-diffusion regularization loss in the training objective. The residual R_t can be viewed as an additive perturbation to the ideal graph-diffusion prediction. A smaller residual means that the free-running trajectory remains closer to the graph-diffusion evolution over multiple prediction steps. The supervised Huber loss fits the observed traffic values, while the graph-diffusion regularization loss discourages temporal transitions inconsistent with the graph-diffusion prior. The combination of evidence-aware learning and physics-guided modeling improves the reliability of recursive prediction.

B Algorithm Framework and Complexity Analysis

This appendix provides the algorithm framework and computational complexity analysis of EviPhys. The algorithm summarizes the complete training procedure, including multimodal semantic extraction, evidence-aware representation learning, dynamic graph construction, physics-guided prediction, and joint optimization. The complexity analysis describes the computational cost of the main components and the overall training and inference procedures.

B.1 Training Procedure

Algorithm 1 presents the training procedure of EviPhys. The framework first extracts multimodal traffic semantics through the frozen ViLT backbone and constructs evidence-aware representations from historical traffic observations. The learned dual representations are then used to build a dynamic graph for traffic propagation. During prediction, two decoding branches are adopted: the supervised branch optimizes prediction accuracy with teacher forcing, while the free-running branch generates recursive trajectories for physics-guided graph-diffusion regularization. Finally, the prediction loss and graph-diffusion regularization loss are jointly optimized to update the model parameters.

B.2 Complexity Analysis

The computational complexity of EviPhys is analyzed as follows. Let B, N, T, and T' denote the batch size, number of sensors, historical length, and prediction horizon, respectively. Let d and d_h denote the embedding dimension and GraphGRU hidden dimension. Let K denote the number of focal elements in DSERM, k denote the graph propagation order, and L denote the number of GraphGRU layers. C_{ViLT} denotes the computational cost of the frozen ViLT encoder for one sample.

The computational cost of MRFM mainly comes from multimodal feature extraction and semantic projection, requiring $\mathcal{O}(BC_{\mathrm{ViLT}} + BNT + BNd)$ operations. For DSERM, evidence extraction,

Algorithm 1 Training procedure of EviPhys.

Require: Historical traffic observations X, future covariates Y_{cov}, traffic labels Y_{s}, physical graph A_{phy}, and model parameters θ.
Ensure: Updated model parameters θ.
Construct traffic images and state-aware prompts from historical traffic observations.
Extract global traffic semantic representations using the frozen ViLT backbone.
Generate macro-level traffic representations from multimodal features.
Extract local traffic evidence representations by modeling uncertainty of historical traffic observations.
Fuse macro semantic representations and micro evidence representations.
Construct the evidence-driven dynamic traffic graph from fused representations.
Encode historical traffic states with the graph-based encoder.
Initialize supervised and free-running decoding branches.
for $t = 1, \ldots, T'$ **do**
    Generate supervised traffic prediction with teacher forcing.
    Generate free-running prediction recursively without future labels.
end for
Compute prediction loss from supervised predictions.
Compute graph-diffusion residual from free-running trajectories.
Calculate physics-guided graph-diffusion regularization loss.
Combine prediction loss and graph-diffusion regularization loss.
Update model parameters θ using back-propagation.

uncertainty estimation, and evidence projection introduce $\mathcal{O}(BN(Td_h + d_hK + Kd))$ complexity. The dual-evidence fusion operation requires $\mathcal{O}(BNd^2)$ operations, while the dynamic graph construction in ESGLM introduces $\mathcal{O}(BN^2d)$ complexity. For the GraphGRU-based prediction module, a single recurrent update requires $\mathcal{O}(B(kN^2d_h + Ndd_h^2))$ operations. Since the training procedure contains one encoder, one supervised decoder, and one free-running decoder, the overall recurrent computation requires $\mathcal{O}(BL(T+2T')(kN^2d_h+Ndd_h^2))$ operations. The computation of graph-diffusion residuals introduces an additional $\mathcal{O}(BT'N^2)$ cost.

Therefore, the overall training complexity of EviPhys is $\mathcal{O}(BC_{\text{ViLT}} + BNT + BN(d_hK + Kd + d^2) + BN^2d + BL(T + 2T')(kN^2d_h + Ndd_h^2) + BT'N^2)$.

During inference, the free-running decoding branch and graph-diffusion residual computation are removed. The recurrent factor is reduced from $T + 2T'$ to $T + T'$. With fixed model dimensions and prediction settings, the dominant complexity is $\mathcal{O}(BN^2)$, mainly resulting from dynamic graph construction and graph propagation.

C Detailed Description of Datasets

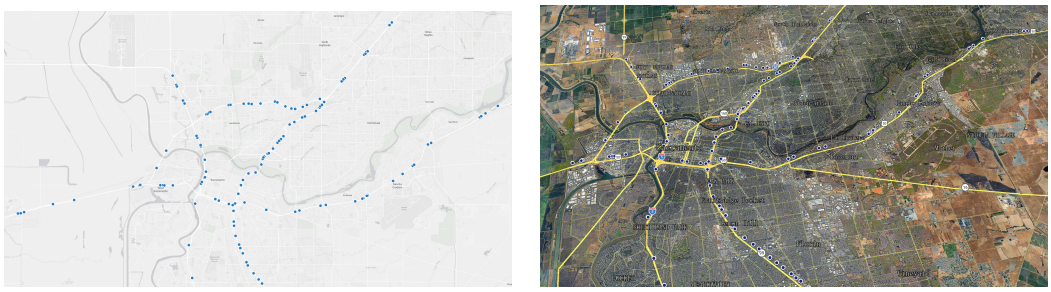
This section provides detailed descriptions of the six traffic flow prediction datasets used in this study. The evaluation datasets include four widely used PEMS benchmark datasets and two large-scale datasets from the LargeST benchmark. All datasets originally provide traffic observations collected from real-world sensor networks with a sampling interval of 5 minutes. The PEMS datasets are mainly used to evaluate the prediction capability under different traffic network scales, while the LargeST datasets are adopted to further examine the scalability of EviPhys on large-scale traffic systems. The statistics of all datasets are summarized in Table 4.

Table 4: Statistics of traffic flow prediction datasets.

	PEMS03	PEMS04	PEMS07	PEMS08	SD	GBA
Nodes	358	307	883	170	716	2,352
Interval	5 min	5 min	5 min	5 min	5 min	5 min

C.1 PEMS Benchmark Datasets

The PEMS datasets are collected from the California Performance Measurement System, which provides real-world traffic flow observations collected by roadside sensors deployed across Califor-



(a) Two-dimensional geographical distribution of traffic sensors in the PEMS dataset, illustrating the spatial deployment of roadside sensors across the urban road network.

(b) Three-dimensional geographical visualization of traffic sensors in the PEMS dataset, providing an enhanced perspective of sensor deployment over the urban road network.

Figure 5: Spatial visualization of traffic sensor deployment in the PEMS dataset.

nia. An example of sensor deployment in the PEMS datasets is illustrated in Figure 5. The two-dimensional visualization presents the geographical distribution of traffic sensors across the urban road network, while the three-dimensional visualization provides an enhanced perspective of sensor deployment by representing the same spatial distribution from a three-dimensional viewpoint. Each sensor records traffic measurements at a 5-minute interval, forming fine-grained spatio-temporal traffic sequences. Due to the diversity of network sizes and traffic patterns, PEMS datasets have been widely adopted as standard benchmarks in the traffic prediction literature, providing representative evaluation scenarios for comparing spatio-temporal prediction models.

In this study, four representative PEMS datasets are employed, including PEMS03, PEMS04, PEMS07, and PEMS08. These datasets contain different numbers of traffic sensors and represent traffic networks with different spatial scales, providing comprehensive evaluation scenarios for traffic prediction models.

- **PEMS03:** PEMS03 is collected from the California Performance Measurement System and contains traffic observations from 358 sensors. The dataset represents a medium-scale traffic network and provides a benchmark for evaluating whether prediction models can effectively capture temporal variations and spatial dependencies among traffic sensors.
- **PEMS04:** PEMS04 contains traffic observations from 307 sensors. Compared with larger traffic networks, PEMS04 provides a relatively compact sensor distribution while maintaining complex temporal traffic variations. PEMS04 is commonly used to evaluate the effectiveness of spatio-temporal prediction methods under regional traffic scenarios.
- **PEMS07:** PEMS07 contains traffic observations from 883 sensors and represents one of the larger-scale datasets among the PEMS benchmarks. The increased number of sensors introduces more complex spatial relationships and heterogeneous traffic patterns, making PEMS07 suitable for evaluating models with dynamic graph learning and large-scale spatial dependency modeling capabilities.
- **PEMS08:** PEMS08 contains traffic observations from 170 sensors. Although PEMS08 has a relatively smaller spatial scale compared with other PEMS datasets, PEMS08 still contains diverse traffic evolution patterns and provides an additional scenario for evaluating prediction performance under compact sensor networks.

C.2 LargeST Benchmark Datasets

To further evaluate the scalability of EviPhys, two large-scale datasets from the LargeST benchmark are adopted, including San Diego (SD) and Greater Bay Area (GBA). LargeST provides representative large-scale traffic prediction benchmarks with substantially larger sensor networks than conventional PEMS datasets, enabling the evaluation of model scalability under more complex spatial dependencies and heterogeneous traffic patterns.

- **San Diego (SD):** The San Diego dataset contains traffic observations collected from 716 sensors. Compared with conventional PEMS benchmarks, SD provides a larger traffic network scale and more complicated spatial interactions among sensors. Therefore, the SD dataset is used to evaluate whether the proposed framework can maintain effective representation learning and prediction performance on larger traffic systems.
- **Greater Bay Area (GBA):** The Greater Bay Area dataset contains traffic observations collected from 2,352 sensors, making GBA the largest dataset used in this study. The large number of sensors introduces stronger spatial complexity and traffic heterogeneity, providing a challenging evaluation scenario for scalable traffic prediction models. The GBA dataset is particularly suitable for testing the ability of EviPhys to construct effective traffic representations and learn large-scale spatial dependencies.

D Detailed Description of Baselines

This section provides detailed descriptions of the baseline methods used for comparison. The selected baselines cover a wide range of traffic flow prediction approaches, including recurrent neural network-based methods, temporal convolution-based methods, graph neural network-based methods, dynamic graph learning approaches, and recent advanced spatio-temporal prediction models.

These baseline methods represent different stages of development in traffic prediction research. Early approaches mainly focus on modeling temporal dependencies through sequence learning mechanisms. Later methods introduce graph-based representations to capture spatial correlations among traffic sensors. Recent studies further explore dynamic spatial relationships, adaptive representations, and advanced spatio-temporal modeling strategies. The diversity of these baselines provides a comprehensive evaluation of the effectiveness of EviPhys under different modeling paradigms.

Recurrent Neural Network-based Methods

- **FC-LSTM** (Sutskever et al., 2014) is a sequence-to-sequence learning framework based on Long Short-Term Memory (LSTM) networks. The method uses a multilayered LSTM encoder to map the input sequence into a fixed-dimensional representation and employs another deep LSTM decoder to generate the target sequence. The encoder-decoder architecture enables FC-LSTM to model temporal dependencies and perform end-to-end sequence prediction.
- **TCN** (Bai et al., 2018) is a sequence modeling framework based on convolutional architectures. The method uses causal convolutions to preserve the temporal order of sequential data and adopts dilated convolutions to enlarge the receptive field for capturing long-range temporal dependencies. TCN constructs hierarchical temporal representations through stacked convolutional layers and performs sequence prediction with parallel convolution operations.

Graph Neural Network-based Methods

- **STGCN** (Yu et al., 2018) is a spatio-temporal graph convolutional framework designed for traffic flow prediction. The method formulates traffic data as graph-structured sequences and replaces traditional convolutional and recurrent operations with graph convolution and temporal convolution components. STGCN captures spatial dependencies among traffic sensors through graph convolutions and models temporal dynamics through convolutional structures. The framework constructs multi-scale spatio-temporal representations for traffic network prediction.
- **ASTGCN** (Guo et al., 2019) is an attention-based spatial-temporal graph convolutional network framework for traffic flow prediction. The method consists of multiple components to model different temporal dependencies, including recent, daily periodic, and weekly periodic patterns. Each component incorporates a spatial-temporal attention mechanism to capture dynamic correlations among traffic nodes and employs spatial-temporal convolution operations to jointly model spatial and temporal features. The outputs from different temporal components are fused to generate the final predictions.

- **DCRNN** (Li et al., 2018) is a spatio-temporal prediction framework that models traffic flow as a diffusion process on directed graphs. The method captures spatial dependencies through bidirectional random walks on the graph and models temporal dependencies through an encoder-decoder recurrent architecture. DCRNN integrates diffusion convolution with recurrent neural networks to learn complex spatio-temporal representations from traffic networks and uses scheduled sampling during sequence generation to improve autoregressive prediction.
- **GWNet** (Wu et al., 2019) is a spatial-temporal graph modeling framework that captures hidden spatial dependencies and long-range temporal patterns in graph-structured data. The method introduces an adaptive dependency matrix learned through node embeddings to discover latent relationships between entities without relying on predefined graph structures. GWNet employs stacked dilated one-dimensional convolution layers to expand the receptive field and model long-range temporal dependencies. The adaptive graph learning component and temporal convolution component are integrated into an end-to-end framework for spatio-temporal prediction.
- **AGCRN** (Bai et al., 2020) is an adaptive graph convolutional recurrent framework for traffic prediction. The method introduces a node adaptive parameter learning module to capture node-specific patterns and a data adaptive graph generation module to automatically infer inter-dependencies among traffic series. AGCRN integrates adaptive graph convolution with recurrent networks to learn fine-grained spatial and temporal correlations without relying on predefined graph structures.
- **STSGCN** (Song et al., 2020) is a spatial-temporal graph convolutional framework for spatial-temporal network data prediction. The method introduces a spatial-temporal synchronous modeling mechanism to jointly capture localized spatial and temporal correlations. STSGCN constructs localized spatial-temporal graphs and designs multiple modules for different time periods to model heterogeneous patterns in spatial-temporal network data. The framework integrates graph convolution operations with synchronous spatial-temporal representations for prediction.
- **STFGNN** (Li & Zhu, 2021) is a spatial-temporal fusion graph neural network framework for traffic flow prediction. The method introduces a data-driven temporal graph generation strategy to supplement missing spatial dependencies in predefined graphs. STFGNN learns hidden spatial-temporal correlations through a fusion operation that integrates multiple spatial and temporal graphs corresponding to different time periods. The framework further combines the fusion graph module with a gated convolution module to construct a unified spatial-temporal learning architecture for modeling complex traffic patterns.

Adaptive and Dynamic Graph-based Methods

- **STID** (Shao et al., 2022) is a spatial-temporal identity-based framework for multivariate time series prediction. The method introduces spatial and temporal identity information to distinguish samples along spatial and temporal dimensions. STID incorporates spatial identity embeddings and temporal identity embeddings into a Multi-Layer Perceptron (MLP)-based architecture to learn spatial-temporal patterns from multivariate time series data. The framework provides a simple prediction paradigm without relying on complex graph structures.
- **MegaCRN** (Jiang et al., 2023b) is a meta-graph convolutional recurrent network framework for traffic prediction. The method introduces spatio-temporal meta-graph learning as a graph structure learning mechanism to capture spatio-temporal heterogeneity and non-stationarity in traffic data. MegaCRN incorporates a meta-graph learner with a meta-node bank into the Graph Convolutional Recurrent Network (GCRN) encoder-decoder architecture. The framework learns adaptive graph structures through meta-node representations and integrates graph convolution with recurrent networks to model dynamic spatial-temporal dependencies.
- **HimNet** (Dong et al., 2024) is a heterogeneity-informed spatiotemporal meta-network framework for spatio-temporal time series prediction. The method introduces a heterogeneity-informed meta-parameter learning scheme to capture spatial and temporal heterogeneity through learned spatial and temporal embeddings. HimNet regards the

learned embeddings as implicit representations of heterogeneous patterns and develops a spatio-temporal meta-parameter learning paradigm to generate specific parameters from meta-parameter pools. The framework incorporates heterogeneity-aware meta-parameter learning into a neural network architecture to model diverse spatio-temporal dependencies.

Recent Spatio-Temporal Prediction Methods

- **RPMixer** (Yeh et al., 2024) is an all-MLP architecture for spatio-temporal prediction. The method applies a general time series prediction paradigm without explicitly modeling spatial relationships among nodes. RPMixer constructs the prediction architecture based on stacked MLP blocks and utilizes residual connections to enhance feature learning. The framework introduces random projection layers to increase the diversity of intermediate representations and improve the learning capability of the network.
- **PDFormer** (Jiang et al., 2023a) is a propagation delay-aware dynamic long-range transformer framework for traffic flow prediction. The method introduces a spatial self-attention mechanism to capture dynamic spatial dependencies among traffic nodes. PDFormer adopts graph masking strategies to model short-range and long-range spatial relationships from different perspectives. The framework further incorporates a traffic delay-aware feature transformation module to explicitly model the time delay of spatial information propagation and capture complex spatio-temporal dependencies in traffic systems.
- **STDN** (Cao et al., 2025) is a spatio-temporal-aware trend-seasonality decomposition network framework for traffic prediction. The method constructs a dynamic graph structure to represent traffic flow and introduces spatio-temporal embeddings to capture global traffic dynamics. STDN employs a trend-seasonality decomposition module to separate traffic patterns into trend-cyclical components and seasonal components for different traffic nodes and time periods. The decomposed components are further processed through an encoder-decoder network to generate the final predictions.
- **PDG2Seq** (Fan et al., 2025) is a periodic dynamic graph to sequence model for traffic flow prediction. The method introduces a periodic feature selection module to extract periodic traffic patterns based on temporal indices. PDG2Seq further develops a periodic dynamic graph convolutional gated recurrent unit to integrate periodic features with dynamic traffic information and construct periodic dynamic graphs for spatio-temporal feature extraction. The framework combines periodic feature modeling, dynamic graph convolution, and sequence decoding to capture temporal periodicity and dynamic spatial-temporal dependencies in traffic data.
- **RAST** (Ruan et al., 2026) is a retrieval-augmented framework for traffic prediction that integrates retrieval mechanisms with spatio-temporal modeling. The method introduces a decoupled encoder and query generator to extract spatial and temporal features and construct fusion queries through residual fusion. RAST maintains a spatio-temporal retrieval store to preserve vectorized fine-grained traffic patterns and retrieves relevant patterns during prediction. The framework further incorporates a universal backbone predictor that supports different spatio-temporal prediction architectures for modeling complex traffic dependencies.

E Detailed Evaluation Metrics and Parameter Settings

E.1 Evaluation Metrics

All predictions are transformed back to the original traffic-flow scale before evaluation. Let B_{te} denote the number of test samples, T' the prediction horizon, and N the number of traffic sensors. For the b-th test sample, $Y_{b,t,i}$ and $\widehat{Y}_{b,t,i}$ denote the ground-truth and predicted traffic flows, respectively, at prediction step t and sensor i.

MAE, RMSE, and MAPE are widely adopted evaluation metrics in traffic-flow forecasting studies and provide complementary measurements of prediction accuracy. MAE measures the average absolute deviation between predictions and observations, RMSE assigns larger penalties to substantial prediction errors, and MAPE evaluates relative prediction errors for scale-independent comparison

across datasets. MAE, RMSE, and MAPE jointly provide a comprehensive evaluation of prediction accuracy and have been extensively used in traffic-flow forecasting benchmarks.

Following the evaluation implementation, entries whose ground-truth traffic flow is zero are excluded. The set of valid evaluation entries is defined as

$$\mathcal{I} = \{(b,t,i) \,|\, Y_{b,t,i} \neq 0,\ 1 \le b \le B_{\mathrm{te}},\ 1 \le t \le T',\ 1 \le i \le N\}, \qquad M = |\mathcal{I}|, \tag{21}$$

where $\mathcal{I}$ is the valid-entry index set and M is the number of entries in this set.

The MAE is defined as

$$\mathrm{MAE} = \frac{1}{M} \sum_{(b,t,i)\in\mathcal{I}} \left| Y_{b,t,i} - \widehat{Y}_{b,t,i} \right|. \tag{22}$$

Here, $Y_{b,t,i} - \widehat{Y}_{b,t,i}$ is the prediction error, and the absolute value measures its magnitude regardless of direction.

The RMSE is defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{M} \sum_{(b,t,i)\in\mathcal{I}} \left(Y_{b,t,i} - \widehat{Y}_{b,t,i} \right)^2}. \tag{23}$$

The squared error assigns a larger penalty to predictions with large deviations, while the square root restores the original traffic-flow unit.

The MAPE is defined as

$$\mathrm{MAPE} = \frac{100}{M} \sum_{(b,t,i)\in\mathcal{I}} \left| \frac{Y_{b,t,i} - \widehat{Y}_{b,t,i}}{Y_{b,t,i}} \right|. \tag{24}$$

The denominator $Y_{b,t,i}$ converts the absolute error into a relative error, and the factor 100 expresses the result as a percentage. Lower MAE, RMSE, and MAPE values indicate better prediction performance.

E.2 Data Processing and Evaluation Protocol

Consecutive sliding windows are used to construct the prediction samples. For the PEMS datasets, 12 historical observations are used to predict the subsequent 12 observations at a 5-minute sampling interval. Therefore, both the observation window and prediction horizon cover one hour. The samples are chronologically divided into training, validation, and test sets using a ratio of 70%/10%/20%.

For the LargeST experiments, the original traffic observations are aggregated into 15-minute intervals using mean aggregation. The aggregated flow values are rounded, and missing observations are filled with zero. The same 12-step input and 12-step output setting is adopted, corresponding to three-hour observation and prediction windows. The resulting samples are chronologically divided into training, validation, and test sets using a ratio of 60%/20%/20%.

For every dataset, scalar Z-score normalization is applied to the traffic-flow channel. The normalization statistics are estimated exclusively from the training inputs to prevent information leakage from the validation or test set. Time-of-day and day-of-week information is retained as additional temporal covariates. The physical road-network adjacency matrix is row-normalized before being used by the model. The main data-processing settings are summarized in Table 5.

For the PEMS datasets, the reported MAE, RMSE, and MAPE values are averaged over all 12 prediction steps. For SD and GBA, the metrics are additionally reported at prediction horizons 3, 6, and 12, together with the average performance over all prediction steps.

E.3 Optimization Settings

EviPhys is implemented in PyTorch and trained on an NVIDIA H100 GPU. Unless otherwise stated, the main experiments use random seed 42. Adam is employed with $\epsilon = 10^{-3}$, and the learning rate is exponentially decayed after every epoch. The training objective uses the Huber loss with threshold

Table 5: Data-processing and evaluation protocols for the six traffic prediction datasets. "Input/output" denotes the numbers of historical and future steps, respectively.

Dataset	Sampling interval	Input/output	Train/val/test	Batch size
PEMS03	5 min	12/12	70%/10%/20%	8
PEMS04	5 min	12/12	70%/10%/20%	32
PEMS07	5 min	12/12	70%/10%/20%	8
PEMS08	5 min	12/12	70%/10%/20%	32
SD	15 min	12/12	60%/20%/20%	8
GBA	15 min	12/12	60%/20%/20%	8

$\delta = 1$. The maximum number of training epochs is 200, and the gradient norm is clipped to 5. An exponential moving average (EMA) of the model parameters is activated from the fifth epoch with a target decay coefficient of 0.995.

The recurrent encoder and decoder retain the same overall architecture across the evaluated datasets. Only a small number of optimization, regularization, and embedding settings are adjusted according to the scale and characteristics of each traffic network. The ViLT backbone remains frozen, while the multimodal projection, evidential representation, dynamic graph construction, recurrent prediction, and physics-guided components are optimized jointly. The SD and GBA experiments follow the same general configuration as the PEMS experiments, with adaptations to the temporal resolution, data split, graph size, and batch size of each dataset.

In addition to the all-horizon Huber loss, an auxiliary tail loss is applied to prediction steps 8-12:

$$\mathcal{L}_{\text{sup}} = \mathcal{L}_{\text{Huber}} + \omega_{\text{tail}} \mathcal{L}_{\text{Huber}}^{8:12}. \tag{25}$$

The effective PINN coefficient is adjusted according to the relative scales of the supervised and graph-diffusion regularization losses:

$$\lambda_{\text{pinn}}^{(e)} = \min\left(\lambda_{\max}, \frac{\rho \ \text{sg}\left(\mathcal{L}_{\text{Huber}}\right)}{\text{sg}\left(\mathcal{L}_{\text{pinn}}\right) + 10^{-6}}\right) \min\left(1, \frac{e}{E_{\text{w}}}\right), \tag{26}$$

where $\text{sg}(\cdot)$ denotes stop-gradient, e is the current epoch, and the PINN warm-up duration is $E_{\text{w}} = 100$ epochs. The implemented training loss is therefore

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda_{\text{pinn}}^{(e)} \mathcal{L}_{\text{pinn}}. \tag{27}$$

The auxiliary evidential loss coefficient is set to zero in the reported experiments.

F Case Study

A case study is conducted on the PEMS04 dataset to analyze the prediction behavior of EviPhys under realistic traffic conditions. Different from the averaged evaluation metrics, the case study provides an intuitive analysis of temporal traffic variations and prediction reliability.

A representative sensor from PEMS04 is selected for visualization over a continuous one-week period from January 15 to January 21, 2018. The traffic flow observations are collected with a 5-minute interval, resulting in 2,016 time steps during the case study period. The comparison between the ground truth and EviPhys predictions is illustrated in Figure 6.

The traffic flow curve exhibits clear periodic characteristics, including low traffic volumes during nighttime periods and significant increases during morning and evening peak hours. Rapid traffic transitions caused by changing traffic conditions are also observed throughout the selected period. The prediction curve generated by EviPhys closely follows the ground truth curve, indicating accurate modeling of temporal traffic dynamics.

During peak traffic periods with intensive fluctuations, EviPhys maintains consistent prediction trends and captures abrupt traffic variations effectively. The stable prediction performance demonstrates the capability of EviPhys to learn complex temporal dependencies from traffic observations.

The reliable prediction behavior benefits from the integration of multimodal semantic representation learning, evidence-aware uncertainty modeling, and physics-guided graph propagation.

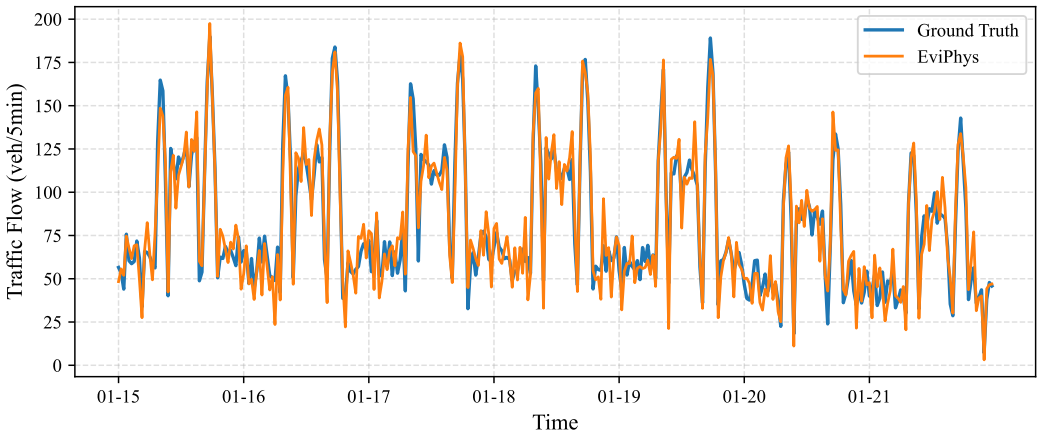


Figure 6: Case study of traffic flow prediction on PEMS04. The comparison between ground truth and EviPhys predictions demonstrates the capability of EviPhys in capturing periodic traffic patterns and dynamic variations.

G VISUALIZATION

To provide an interpretable view of EviPhys, this section visualizes the learned contextual node representations on PEMS04 for a fixed test sample. Each representation integrates global multimodal semantics, local micro-evidence, and learnable node information.

Figure 7(a) presents a t-SNE projection of the contextual node representations. The high-affinity node pair (Nodes 193 and 200) is located closely in the representation space, whereas the low-affinity pair (Nodes 44 and 297) is clearly separated. This result indicates that the learned representations distinguish nodes with different contextual traffic characteristics.

Figure 7(b) shows the cosine-similarity matrix of the contextual node representations. The matrix contains both localized high-similarity regions and low-similarity regions, demonstrating that the learned node relationships are non-uniform and sample-dependent. The selected node pairs are highlighted to connect the learned semantic affinity with observable traffic patterns.

Figure 7(c) compares the daily traffic flow curves of the two selected node pairs. Nodes 193 and 200 have a cosine similarity of 1.000, and the corresponding traffic curves almost completely overlap, with a Pearson correlation of 1.000. In contrast, Nodes 44 and 297 have a low cosine similarity of -0.316. Although the two nodes share a broad daily trend, the corresponding traffic magnitudes and local fluctuations differ substantially, resulting in a lower Pearson correlation of 0.887. These results show that EviPhys learns contextual node relationships that are consistent with, while more informative than, raw traffic flow similarity.

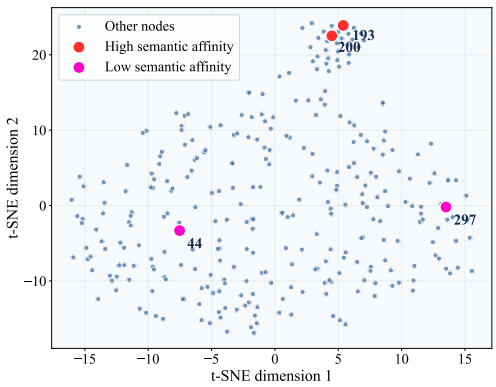
H DETAILED PHYSICAL PRIOR COMPARISON

To further examine the influence of physical priors on traffic prediction, different physical priors are compared by replacing only the graph-diffusion residual term while keeping all other model settings unchanged. Experiments are conducted on PEMS03, PEMS04, PEMS07, and PEMS08.

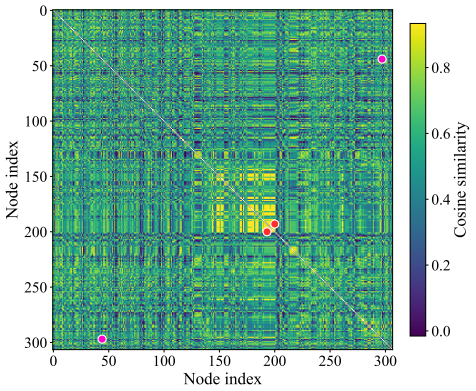
Let g_t denote the free-running traffic flow prediction at prediction step t, and let $g_0 = X_{:,T,:,1}$ represent the last observed normalized flow. A_{rw} denotes the row-normalized physical road graph, and the graph-diffusion operator is defined as

$$\mathcal{L}_{\mathrm{rw}}(v) = (I_N - A_{\mathrm{rw}})v. \tag{28}$$

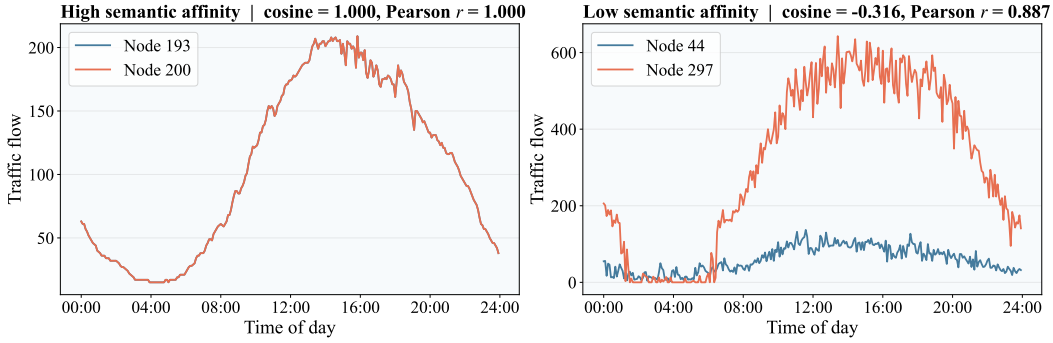
All variants adopt the same initial condition and road-endpoint boundary constraint.



(a) t-SNE projection of contextual node representations.



(b) Cosine-similarity matrix of contextual node representations.



(c) Daily traffic flow curves of the selected node pairs.

Figure 7: Visualization of contextual node representations learned by EviPhys on PEMS04. (a) t-SNE projection of contextual node representations. (b) Cosine-similarity matrix among traffic nodes. (c) Daily traffic flow patterns of the selected high- and low-affinity node pairs.

Table 6: Comparison of different physical priors in EviPhys on the PEMS03, PEMS04, PEMS07, and PEMS08 datasets.

Methods	PEMS03			PEMS04			PEMS07			PEMS08		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
Lorenz system	14.33	25.74	14.65%	18.09	29.95	12.05%	18.72	32.54	7.71%	13.02	22.51	8.80%
Wave equation	14.48	25.95	14.83%	17.99	**29.87**	11.97%	18.80	32.96	7.75%	13.03	22.53	8.84%
Burgers' equation	14.42	26.21	14.68%	17.99	30.16	12.01%	18.75	32.77	7.74%	13.04	22.54	8.92%
EviPhys	**14.31**	**25.43**	**14.33%**	**17.95**	**29.87**	**11.91%**	**18.66**	**32.46**	**7.70%**	**12.99**	**22.45**	**8.79%**

Lorenz system. The graph-based states are defined as $x_t = g_t$, $y_t = A_{\text{rw}} g_t$, and $z_t = \mathcal{L}_{\text{rw}}(g_t)$. The Lorenz system is formulated as

$$\dot{x} = \sigma(y - x), \qquad \dot{y} = x(\rho - z) - y, \qquad \dot{z} = xy - \beta z, \tag{29}$$

where $\sigma = 10$, $\rho = 28$, and $\beta = 8/3$.

Wave equation. To characterize the propagation of traffic disturbances over the physical road graph, the wave equation is defined as

$$\frac{\partial^2 g_t}{\partial t^2} + c^2 \mathcal{L}_{\text{rw}}(g_t) = 0, \tag{30}$$

where $c = 1$ controls the propagation speed. The initial flow is fixed by g_0, and the initial velocity is constrained to zero.

Table 7: Robustness evaluation under different noise levels on PEMS04.

Noise Level	MAE	RMSE	MAPE(%)
0%	17.95	29.87	11.91
5%	18.21	30.24	12.08
10%	18.56	30.71	12.31
20%	19.04	31.52	12.74

Table 8: Statistical significance analysis over 10 independent runs.

Dataset	PDFormer	EviPhys	p-value
PEMS03	14.94±0.083	14.31±0.061	0.003
PEMS04	18.36±0.091	17.95±0.072	0.008
PEMS07	19.97±0.112	18.66±0.086	0.002
PEMS08	13.58±0.074	12.99±0.059	0.004

Burgers' equation. To describe nonlinear traffic propagation with graph diffusion, the Burgers' equation is formulated as

$$\frac{\partial g_t}{\partial t} + g_t \odot \nabla_{\mathrm{rw}}(g_t) + \nu \mathcal{L}_{\mathrm{rw}}(g_t) = 0, \tag{31}$$

where ∇_{rw} denotes the directed graph difference operator, $\odot$ represents element-wise multiplication, and $\nu = 0.01/\pi$ controls the graph-diffusion strength.

During training, temporal derivatives are approximated using finite differences. All variants are evaluated under the same EviPhys prediction framework using MAE, RMSE, and MAPE, thereby isolating the effect of different physical priors.

As shown in Table 6, all physics-informed variants achieve competitive performance across the four datasets, confirming the regularization benefit of these physical priors. The graph-diffusion formulation achieves the best overall performance across most metrics. Unlike the generic nonlinear dynamics of the Lorenz system, disturbance propagation of the wave equation, and advection-diffusion dynamics of Burgers' equation, the graph-diffusion residual directly couples temporal flow changes with propagation over the physical sensor graph. The results demonstrate stronger numerical performance and better consistency with road-network topology.

I Robustness Analysis

Traffic flow measurements collected by sensors may inevitably contain noise caused by sensor failures, communication errors, or environmental disturbances. Such uncertainties introduce deviations between observed traffic states and underlying physical patterns, making stable spatio-temporal representation learning challenging for prediction models. Furthermore, noisy observations may propagate through graph-based modeling and result in unreliable dependency learning among traffic sensors. Therefore, a robust traffic prediction model should maintain reliable performance when historical observations are contaminated by uncertainties. To evaluate the robustness of EviPhys under uncertain observations, Gaussian noise with different intensities is injected into input traffic sequences, and performance variations under various noise conditions are analyzed.

Gaussian noise with different intensities is introduced into the input traffic sequences during inference to simulate uncertain observations. Specifically, the noise-corrupted traffic sequence is formulated as

$$X^{noise} = X + \epsilon, \tag{32}$$

where X denotes the original traffic sequence and ϵ represents Gaussian noise sampled from

$$\epsilon \sim \mathcal{N}(0, \sigma^2). \tag{33}$$

Four noise levels are considered, including 0%, 5%, 10%, and 20%, where the noise ratio controls the magnitude of perturbations added to the original observations. All models are evaluated under identical noisy conditions, while keeping the remaining experimental settings unchanged. The robustness results on the PEMS04 dataset are reported in Table 7.

As shown in Table 7, the prediction performance gradually decreases as the noise intensity increases, which is consistent with the increasing uncertainty in input observations. However, EviPhys still maintains stable performance under different perturbation levels. When the noise ratio increases from 0% to 20%, the MAE only increases from 17.95 to 19.04, demonstrating the strong robustness of the proposed framework.

This robustness mainly benefits from the uncertainty-aware representation and physics-guided modeling strategies. DSERM forms uncertainty-sensitive local representations through entropy-gated evidence modeling. Meanwhile, PGPM introduces traffic evolution constraints, preventing the prediction process from excessively relying on noisy historical patterns. Therefore, EviPhys can effectively maintain reliable predictions even under degraded observation conditions.

J Statistical Significance Analysis

Deep learning models usually involve randomness caused by parameter initialization, mini-batch sampling, and optimization procedures. Therefore, evaluating a model based on a single training run may introduce uncertainty into performance comparison. To verify the statistical reliability of the improvements achieved by EviPhys, this section conducts multiple independent experiments with different random seeds.

Specifically, EviPhys and the strongest baseline model PDFormer are trained independently for 10 runs. The mean value and standard deviation of MAE are reported. Furthermore, paired t-tests are performed between EviPhys and PDFormer to determine whether the performance improvements are statistically significant. The statistical results on four PEMS datasets are presented in Table 8.

The results demonstrate that EviPhys consistently achieves lower prediction errors with smaller performance variations across different datasets. Specifically, compared with PDFormer, EviPhys reduces the mean MAE by 0.63, 0.41, 1.31, and 0.59 on PEMS03, PEMS04, PEMS07, and PEMS08, corresponding to relative reductions of 4.22%, 2.23%, 6.56%, and 4.34%, respectively. The largest improvement is observed on PEMS07, while consistent gains are maintained across all four datasets. Compared with PDFormer, EviPhys also obtains reduced standard deviations, indicating better training stability and robustness against random initialization. The standard deviations decrease from 0.083 to 0.061, 0.091 to 0.072, 0.112 to 0.086, and 0.074 to 0.059 on the four datasets, respectively, corresponding to reductions of approximately 20.27%–26.51%.

Moreover, all paired t-test results achieve statistical significance with $p < 0.01$, with p-values of 0.003, 0.008, 0.002, and 0.004 on PEMS03, PEMS04, PEMS07, and PEMS08, respectively, confirming that the improvements of EviPhys are unlikely to be caused by random fluctuations. These findings further validate the effectiveness and reliability of the proposed evidence-aware multimodal representation learning and physics-guided graph modeling framework.

K Appendix Summary

In summary, this appendix provides supplementary analyses and implementation details to further support the theoretical foundation, reproducibility, and empirical evaluation of EviPhys. Appendix A presents the theoretical analysis of the graph-diffusion regularization, explaining how the graph-diffusion physical prior constrains recursive prediction over the physical road-sensor graph and improves the stability of long-horizon forecasting. Appendix B describes the complete training procedure, including the supervised and free-running decoding branches, and analyzes the computational complexity of the main components during training and inference. Appendices C–E provide detailed descriptions of the datasets, baseline methods, evaluation metrics, data-processing protocols, and optimization settings, offering additional information for reproducing the reported experiments.

Appendices F and G provide complementary interpretability analyses. The PEMS04 case study illustrates the temporal prediction behavior of EviPhys under representative traffic conditions, while the visualization of learned contextual node representations examines how multimodal semantics and local evidential information are reflected in the learned representation space. Appendix H further investigates the influence of different physical priors by comparing the proposed graph-diffusion formulation with alternative dynamics-based formulations under the same prediction framework. Appendix I evaluates prediction robustness under different levels of input noise, and Appendix J examines the statistical significance of the performance improvements through multiple independent runs. Together, these supplementary analyses provide additional evidence for the theoretical motivation, implementation transparency, large-scale applicability, robustness, and empirical reliability of EviPhys for traffic flow prediction.