HealthMamba: An Uncertainty-aware Spatiotemporal Graph State Space Model for Effective and Reliable Healthcare Facility Visit Prediction

**Dahai Yu[1], Lin Jiang[1], Rongchao Xu[1], Guang Wang[1]*

[1]Department of Computer Science, Florida State University
dahai.yu@fsu.edu, lin.jiang@fsu.edu, rxu@fsu.edu, guang@cs.fsu.edu

Abstract

Healthcare facility visit prediction is essential for optimizing healthcare resource allocation and informing public health policy. Despite advanced machine learning methods being employed for better prediction performance, existing works usually formulate this task as a time-series forecasting problem without considering the intrinsic spatial dependencies of different types of healthcare facilities, and they also fail to provide reliable predictions under abnormal situations such as public emergencies. To advance existing research, we propose HealthMamba, an uncertainty-aware spatiotemporal framework for accurate and reliable healthcare facility visit prediction. HealthMamba comprises three key components: (i) a Unified Spatiotemporal Context Encoder that fuses heterogeneous static and dynamic information, (ii) a novel Graph State Space Model called GraphMamba for hierarchical spatiotemporal modeling, and (iii) a comprehensive uncertainty quantification module integrating three uncertainty quantification mechanisms for reliable prediction. We evaluate HealthMamba on four large-scale real-world datasets from California, New York, Texas, and Florida. Results show HealthMamba achieves around 6.0% improvement in prediction accuracy and 3.5% improvement in uncertainty quantification over state-of-the-art baselines.

1 Introduction

Healthcare facilities encompass a wide range of locations providing healthcare services, from small clinics and physicians' offices to urgent care centers and large hospitals. They are indispensable to population well-being for preventing and treating disease, enabling continuous care, and forming the backbone of resilient public health systems. Consequently, predicting population visits to healthcare facilities is critical for resource allocation and policy planning [Marcusson *et al.*, 2020; Xu *et al.*, 2026a; Xu *et al.*, 2026b].

Recent studies on have made progress by using statistical methods [Pun *et al.*, 2019; Alghamdi *et al.*, 2019], deep learning methods [Piccialli *et al.*, 2020; Zhong *et al.*, 2026], and hybrid approaches [Jiang *et al.*, 2025a; Xu *et al.*, 2025; Xiao and Liu, 2025]. However, there are three key limitations of existing work. First, prior studies predominantly focus on aggregated visit prediction, overlooking fine-grained healthcare facility types such as ambulatory clinics, hospitals, nursing facilities, and social assistance providers. Second, most existing work formulates healthcare facility visit prediction as a purely time-series forecasting problem, neglecting spatial dependencies that could provide critical information for accurate prediction. Third, most state-of-the-art approaches fail to deliver reliable predictions under public health emergencies and extreme weather events, such as the COVID-19 pandemic and hurricanes.

Hence, in this paper, we aim to develop an uncertainty-aware spatiotemporal framework for type-specific healthcare facility visit prediction that explicitly models spatial dependencies while providing reliable uncertainty estimates. However, there are two key challenges to achieving this. The first challenge lies in the heterogeneity and sensitivity of health-related mobility. Our data analysis indicates that healthcare facility visits exhibit a strong correlation with population structure (e.g., age composition), spatial accessibility, and service availability. These characteristics lead to highly uneven visit patterns across regions and facility types, making it difficult for existing models to capture fine-grained spatiotemporal heterogeneity. Second, reliable uncertainty quantification under abnormal situations remains challenging. During public emergencies such as pandemics, healthcare facility visit behaviors and demand distributions shift dramatically. Without explicit uncertainty modeling, predictions become overconfident and unreliable, potentially leading to misguided decision-making.

To address the above challenges and advance existing research, we propose HealthMamba, which comprises three key components: (i) a **Unified SpatioTemporal Context Encoder (STCE)**, which fuses heterogeneous static and time-varying contextual information and historical visit data into compact node-time representations, (ii) a novel **GraphMamba Backbone (G-Mamba)**, which innovatively integrates adaptive graph learning into a UNet-style Mamba architecture for hierarchical spatiotemporal modeling, and (iii)

*Prof. Guang Wang is the corresponding author.

a **comprehensive uncertainty quantification (UQ)** module integrating node-based, distribution-based, and parameter-based UQ mechanisms, followed by a post-hoc quantile calibration to achieve higher prediction reliability.

The key contributions of this paper are as follows:

- **Conceptually**, unlike existing works that focus on aggregated healthcare facility visit prediction and treat this task as a time-series forecasting problem, we target type-specific healthcare facility visit prediction and formulate it as an uncertainty-aware spatiotemporal forecasting task with explicit spatial modeling.
- **Technically**, we propose HealthMamba, an uncertainty-aware spatiotemporal framework comprising three novel components: (i) a Unified Spatiotemporal Context Encoder for heterogeneous information fusion, (ii) a graph state space model called GraphMamba for hierarchical spatiotemporal modeling, and (iii) a comprehensive UQ module that integrates three different UQ mechanisms and a quantile calibration for reliable prediction.
- **Empirically**, we evaluate HealthMamba on four real-world datasets from California, New York, Texas, and Florida by comparing it with 13 state-of-the-art baselines across five metrics. Extensive results demonstrate that HealthMamba achieves 6.0% higher prediction accuracy and 3.5% better uncertainty quantification than the best baseline. Our implementation is available at https://github.com/UFOdestiny/HealthMamba. A longer version of this paper can be found at https://arxiv.org/pdf/2602.05286.

2 Problem Formulation

Based on data analysis using real-world data, we found that there is a disparity in healthcare facility distribution and accessibility among different regions. For example, there are fewer healthcare facilities in rural areas, and rural residents need to travel longer distances to visit healthcare facilities. In addition, the visit patterns of different types of healthcare facilities are also different. Hence, we intuitively build graphs to model spatial dependencies to predict multi-type healthcare facility visits.

2.1 Spatiotemporal Graph Construction

We define a spatiotemporal graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$, where $\mathcal{V} = \{v_1, \dots, v_N\}$ denotes N nodes (e.g., counties), $\mathcal{E}$ is the edge set, and $\mathbf{A} \in \mathbb{R}^{N\times N}$ is the weighted adjacency matrix. At each time step t, node i has a feature vector $\mathbf{v}_{i,t} \in \mathbb{R}^C$ representing visit counts across C healthcare facility categories. We construct $\mathbf{A}$ using a Gaussian kernel with sparsity thresholding:

$$a_{ij} = \begin{cases} \exp\left(-d_{ij}^2/\sigma^2\right), & i \neq j \text{ and } \exp\left(-d_{ij}^2/\sigma^2\right) \geq \epsilon, \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where d_{ij} is the centroid distance between nodes v_i and v_j, $\sigma > 0$ controls spatial decay, and $\epsilon \in [0, 1]$ is a sparsity threshold that removes weak connections.

2.2 Healthcare Facility Visit Prediction

Let $\mathbf{V} \in \mathbb{R}^{N\times T_{\text{in}}\times C}$ denote the input visit tensor including features of all N nodes over T_{in} historical time steps across C healthcare facility types. In addition to historical visits, we also incorporate static attributes $\mathbf{D} \in \mathbb{R}^{N\times d_{\text{dem}}}$ (e.g., demographics) and time-varying external factors $\mathbf{E} \in \mathbb{R}^{N\times T_{\text{in}}\times d_{\text{ext}}}$ (e.g., weather, accessibility) as auxiliary inputs, where d_{dem} and d_{ext} denote the dimensions of demographic attributes and external factors, respectively. The goal of healthcare facility visit prediction is to predict future visits for the next T_{out} steps given historical observations and auxiliary information.

Deterministic Prediction

Conventional deterministic prediction learns a mapping f from inputs to point forecasts:

$$\{\mathbf{V}, \mathbf{D}, \mathbf{E}, \mathbf{A}\} \xrightarrow{f} \hat{\mathbf{Y}} \in \mathbb{R}^{N\times T_{\text{out}}\times C}. \tag{2}$$

Probabilistic Prediction

In this work, we focus on probabilistic prediction, which quantifies uncertainty using prediction intervals. Given a target miscoverage rate α (e.g., $\alpha = 0.1$ for 90% coverage), we predict lower, upper, and median quantiles:

$$\{\mathbf{V}, \mathbf{D}, \mathbf{E}, \mathbf{A}\} \xrightarrow{\mathcal{F}} \left[\hat{\mathbf{L}}, \hat{\mathbf{U}}, \hat{\mathbf{M}}\right], \tag{3}$$

where $\hat{\mathbf{L}} = \hat{q}_{\alpha/2}(\mathbf{V})$, $\hat{\mathbf{U}} = \hat{q}_{1-\alpha/2}(\mathbf{V})$, and $\hat{\mathbf{M}} = \hat{q}_{0.5}(\mathbf{V})$ represent the lower bound, upper bound, and median predictions, respectively, all in $\mathbb{R}^{N\times T_{\text{out}}\times C}$. The median $\hat{\mathbf{M}}$ also serves as the point prediction, analogous to $\hat{\mathbf{Y}}$ in Eq. (2).

3 Methodology

In this paper, we propose HealthMamba, a spatiotemporal framework for effective and reliable healthcare facility visit prediction. An overview of HealthMamba is shown in Figure 1, which comprises three key components: (i) a **Unified SpatioTemporal Context Encoder (STCE)**, which fuses heterogeneous contextual information and historical visit data into compact node-time representations, (ii) a novel **GraphMamba Backbone (G-Mamba)**, which innovatively integrates adaptive graph learning into a UNet-style Mamba architecture for hierarchical spatiotemporal modeling, and (iii) a **comprehensive uncertainty quantification** module integrating node-based, distribution-based, and parameter-based uncertainty quantification, followed by a post-hoc quantile calibration to achieve higher prediction reliability.

3.1 Unified Spatiotemporal Context Encoder

Given the input visit tensor $\mathbf{V}$, static attributes $\mathbf{D}$, dynamic external factors $\mathbf{E}$, and the adjacency matrix $\mathbf{A}$ as defined in Section 2, STCE produces a unified node-time representation $\mathbf{R} \in \mathbb{R}^{N\times T_{\text{in}}\times d_{\text{model}}}$ that fuses these heterogeneous inputs while considering spatial structure and temporal dynamics:

$$\mathbf{R} = \text{STCE}(\mathbf{V}, \mathbf{D}, \mathbf{E}, \mathbf{A}). \tag{4}$$

We utilize d_{hid} and d_{model} to denote the hidden-layer dimension and model output embedding dimension, respectively.

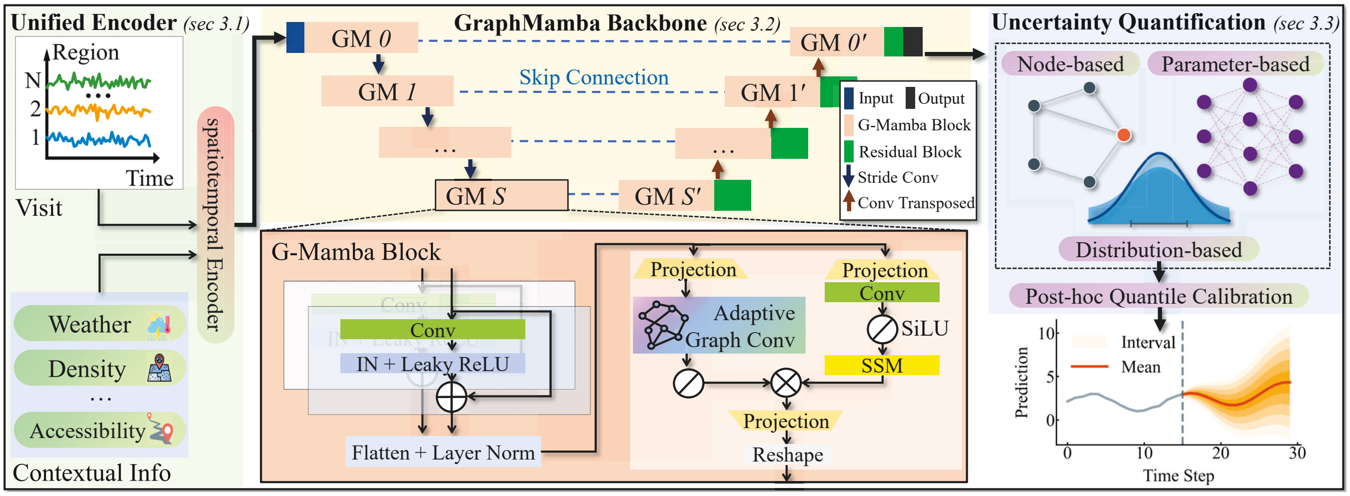


Figure 1: Framework of HealthMamba, which consists of three major components: (i) Unified Spatiotemporal Context Encoder for heterogeneous input fusion, (ii) GraphMamba Backbone for hierarchical spatiotemporal modeling with adaptive graph learning, and (iii) comprehensive uncertainty quantification (UQ) module with three types of UQ mechanisms and post-hoc quantile calibration for reliable prediction.

Specifically, we first map inputs to a common hidden space via feature embedding layers:

$$\mathbf{H}^v_{i,t} = \phi_v\big(W_v\,\mathbf{v}_{i,t} + b_v\big) \in \mathbb{R}^{d_{\text{hid}}}, \quad (5)$$

$$\mathbf{H}^d_{i} = \phi_d\big(W_d\,\mathbf{D}_{i} + b_d\big) \in \mathbb{R}^{d_{\text{hid}}}, \quad (6)$$

$$\mathbf{H}^e_{i,t} = \phi_e\big(W_e\,\mathbf{E}_{i,t} + b_e\big) \in \mathbb{R}^{d_{\text{hid}}}, \quad (7)$$

where ϕ_* denotes a nonlinear function (e.g., SiLU/GELU) with dropout. The initial fused embedding at each (i, t) is:

$$\mathbf{X}_{i,t} = \mathbf{H}^v_{i,t} + \mathbf{H}^d_{i} + \mathbf{H}^e_{i,t}. \quad (8)$$

For spatial encoding, we apply graph convolution with activation function ReLU on node features $\mathbf{X}_{:,t}$ at each time t:

$$\widetilde{\mathbf{X}}_{:,t} = \mathrm{GConv}\big(\mathbf{X}_{:,t};\mathbf{A}\big) = \sigma\Big(\sum_{j=1}^{N}\tilde{a}_{ij}\,W_g\,\mathbf{X}_{j,t} + b_g\Big). \quad (9)$$

For temporal encoding, we mix information across the T_{in} steps for each node i using a light-weight temporal mixer combining depthwise 1D convolution and channel mixing:

$$\mathbf{U}_{i,:} = \mathrm{DepthConv1D}\big(\widetilde{\mathbf{X}}_{i,:}\big) + \widetilde{\mathbf{X}}_{i,:} \in \mathbb{R}^{T_{\text{in}}\times d_{\text{hid}}}, \quad (10)$$

$$\widehat{\mathbf{X}}_{i,t} = \mathrm{MLP}_{\text{chan}}\big(\mathrm{LayerNorm}(\mathbf{U}_{i,t})\big) + \mathbf{U}_{i,t}, \quad (11)$$

where $\mathrm{DepthConv1D}$ mixes information across the temporal axis and $\mathrm{MLP}_{\text{chan}}$ performs channel mixing. LayerNorm and residual connections stabilize training. The final projection is:

$$\mathbf{Z}_{i,t} = \mathrm{LayerNorm}\Big(W_p\,\big[\widehat{\mathbf{X}}_{i,t}\,;\,\mathbf{H}^d_{i}\big] + b_p\Big), \quad (12)$$

$$\mathbf{R}_{i,t} = \mathrm{Dropout}\big(\mathrm{SiLU}(W_o\,\mathbf{Z}_{i,t} + b_o)\big) \in \mathbb{R}^{d_{\text{model}}}. \quad (13)$$

The encoder output $\mathbf{R} \in \mathbb{R}^{N\times T_{\text{in}}\times d_{\text{model}}}$ serves as the input to the GraphMamba backbone.

3.2 GraphMamba Backbone

In this part, we innovatively integrate an adaptive graph learning module into a UNet-style Mamba architecture for hierarchical spatiotemporal modeling. The input feature at scale s is $\mathbf{X}^{(s)} \in \mathbb{R}^{N\times T_{\text{in}}\times d_s}$, with $\mathbf{X}^{(0)} = \mathbf{R}$. The encoder applies a sequence of G-Mamba blocks followed by downsampling; the decoder performs symmetric upsampling followed by additional G-Mamba blocks. At the coarsest scale, we initialize the decoder by setting $\mathbf{Z}^{(S)} = \mathbf{Y}^{(S)}$. For other scales, we use the following equations.

$$\mathbf{Y}^{(s)} = \text{G-Mamba}^{(s)}\big(\mathbf{X}^{(s)}\big), \quad (14)$$

$$\mathbf{X}^{(s+1)} = \mathrm{DownConv}\big(\mathbf{Y}^{(s)}\big), \qquad s = 0,\ldots,S-1, \quad (15)$$

$$\widehat{\mathbf{Y}}^{(s)} = \mathrm{UpConv}\big(\mathbf{Z}^{(s+1)}\big), \qquad s = S-1,\ldots,0, \quad (16)$$

$$\mathbf{Z}^{(s)} = \mathrm{Fuse}\big(\mathbf{Y}^{(s)}, \widehat{\mathbf{Y}}^{(s)}\big). \quad (17)$$

Here, $\mathrm{DownConv}$ denotes stride convolution, UpConv denotes transposed convolution, and Fuse represents skip-connection fusion. Each scale maintains its own G-Mamba block without shared weights, allowing scale-specific modeling capacity.

G-Mamba Block

Each GraphMamba Block (G-Mamba) is designed to capture spatial dependencies, temporal dynamics, and channel interactions, incorporating an Adaptive Graph Learning mechanism that learns scale-specific, data-driven graph structures.

Adaptive Graph Learning. We first derive a node-level embedding by aggregating temporal information:

$$\mathbf{u}_i = \mathrm{Pooling}\big(\mathbf{X}^{(s)}_{i,:}\big). \quad (18)$$

Node affinities are learned using an attention-style mechanism:

$$\tilde{e}_{ij} = \mathrm{LeakyReLU}\big(\mathbf{a}^\top [W_u \mathbf{u}_i \,\|\, W_u \mathbf{u}_j]\big), \tag{19}$$

$$\alpha_{ij} = \frac{\exp(\tilde{e}_{ij})}{\sum_{k=1}^{N} \exp(\tilde{e}_{ik})} \in \mathbf{A}^{(\mathrm{ag})}. \tag{20}$$

The resulting adjacency is symmetrized and normalized:

$$\mathbf{A}^{(\mathrm{sym})} = \tfrac{1}{2}\big(\mathbf{A}^{(\mathrm{ag})} + (\mathbf{A}^{(\mathrm{ag})})^\top\big), \quad \widehat{\mathbf{A}} = \tilde{\mathbf{D}}^{-\frac{1}{2}} \mathbf{A}^{(\mathrm{sym})} \tilde{\mathbf{D}}^{-\frac{1}{2}}, \tag{21}$$

where $\tilde{\mathbf{D}}$ is the degree matrix of $\mathbf{A}^{(\mathrm{sym})}$. If prior adjacency $\mathbf{A}_0$ is available, we blend the two sources:

$$\mathbf{A}^* = \lambda \mathbf{A}_0 + (1-\lambda)\widehat{\mathbf{A}}, \quad \lambda \in [0,1]. \tag{22}$$

Graph-enhanced Spatial Mixing. Using the learned adjacency $\mathbf{A}^*$, the block performs graph convolution at each time step:

$$\mathbf{G}_{:,t} = \sigma\Big(\mathbf{A}^* \,(W_g \mathbf{X}^{(s)}_{:,t}) + b_g\Big). \tag{23}$$

A residual connection stabilizes learning:

$$\mathbf{G} \leftarrow \mathbf{G} + \mathbf{X}^{(s)}. \tag{24}$$

Temporal and Channel Mixing. Temporal dependencies are modeled via a State-Space Module (SSM), followed by channel mixing:

$$\mathbf{T}_{i,:} = \mathrm{SSM}\big(\mathbf{G}_{i,:}\big) + \mathbf{G}_{i,:}, \tag{25}$$

$$\mathbf{C}_{i,t} = \mathrm{MLP}_{\mathrm{chan}}\big(\mathrm{LayerNorm}(\mathbf{T}_{i,t})\big) + \mathbf{T}_{i,t}. \tag{26}$$

Projection and Residual Refinement. The block output is produced by projecting back to the original dimension with a residual connection:

$$\mathbf{O} = \mathrm{LayerNorm}\big(W_o \mathbf{C} + b_o\big) + \mathbf{X}^{(s)}. \tag{27}$$

Multi-scale Fusion and Skip Connections. During decoding, the upsampled representation at scale s is fused with the encoder's corresponding feature $\mathbf{Y}^{(s)}$ via concatenation and projection:

$$\mathbf{Z}^{(s)} = \mathrm{Fuse}(\mathbf{Y}^{(s)}, \widehat{\mathbf{Y}}^{(s)}) = \mathrm{Proj}\big([\mathbf{Y}^{(s)}; \widehat{\mathbf{Y}}^{(s)}]\big). \tag{28}$$

The fused tensor is passed through another G-Mamba block, enabling joint refinement of encoder and decoder information.

3.3 Uncertainty-aware Prediction

We design a modular uncertainty quantification component atop the GraphMamba backbone that produces calibrated uncertainty estimates. This component comprises three complementary mechanisms, node-based, distribution-based, and parameter-based uncertainty quantification, which together capture local heteroscedasticity, predictive distributional shape, and epistemic uncertainty.

Node-based Uncertainty

Node-based uncertainty produces per-node prediction intervals expressing local uncertainty at the region level. For a prediction target $y_{i,t+\tau}$ at node i and horizon τ, we learn lower and upper quantile heads $\ell_\alpha(i,t,\tau)$ and $u_\alpha(i,t,\tau)$ for a nominal miscoverage level α (e.g., $\alpha = 0.1$ for a 90% interval). Each quantile head is implemented as a projection from the backbone's final hierarchical representation $\mathbf{Z}$. The pinball loss for a set $\mathcal{Q}$ of quantile levels is:

$$\mathcal{L}_{\mathrm{quant}} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{1}{|\mathcal{S}|} \sum_{(i,t+\tau) \in \mathcal{S}} \rho_q\big(y_{i,t+\tau} - \hat{q}_{i,t+\tau}\big), \tag{29}$$

where $\rho_q(z) = \max(qz, (q-1)z)$ is the pinball loss, $\hat{q}_{i,t+\tau}$ denotes the predicted q-quantile, and $\mathcal{S}$ is the training set. Node-based intervals are robust to misspecification of the full predictive distribution, as they directly target coverage without assuming any parametric form.

Distribution-based Uncertainty

Distribution-based uncertainty fits a parametric predictive distribution (e.g., Gaussian or Laplace) per node-time. The most common instantiation is a heteroscedastic Gaussian predictor outputting mean $\mu_{i,t+\tau}$ and variance $\sigma^2_{i,t+\tau}$:

$$p(y_{i,t+\tau} \mid \mathbf{x}) = \mathcal{N}\big(y_{i,t+\tau}\,;\, \mu_{i,t+\tau},\, \sigma^2_{i,t+\tau}\big). \tag{30}$$

This head is trained by minimizing the negative log likelihood (NLL):

$$\mathcal{L}_{\mathrm{nll}} = \frac{1}{|\mathcal{S}|} \sum_{(i,t+\tau) \in \mathcal{S}} \left(\tfrac{1}{2} \log \sigma^2_{i,t+\tau} + \tfrac{(y_{i,t+\tau} - \mu_{i,t+\tau})^2}{2\sigma^2_{i,t+\tau}}\right). \tag{31}$$

Distribution-based heads capture aleatoric uncertainty (data noise and heteroscedasticity) and allow analytic computation of prediction intervals, e.g., $\mu \pm z_{1-\alpha/2}\sigma$ for Gaussian.

Parameter-based Uncertainty

Parameter-based uncertainty estimates epistemic uncertainty arising from model parameters. We adopt the commonly-used MC Dropout, which performs M stochastic forward passes at inference with dropout enabled. Given M stochastic predictions $\{\mu^{(m)}_{i,t+\tau}\}_{m=1}^{M}$ and $\{\sigma^{2,(m)}_{i,t+\tau}\}_{m=1}^{M}$, the predictive mean and total variance decompose as:

$$\widehat{\mu}_{i,t+\tau} = \frac{1}{M} \sum_{m=1}^{M} \mu^{(m)}_{i,t+\tau}, \tag{32}$$

$$\mathrm{Var}_{\mathrm{pred}}(y) = \underbrace{\frac{1}{M} \sum_{m=1}^{M} \sigma^{2,(m)}_{i,t+\tau}}_{\text{aleatoric}} + \underbrace{\frac{1}{M} \sum_{m=1}^{M} \big(\mu^{(m)}_{i,t+\tau} - \widehat{\mu}_{i,t+\tau}\big)^2}_{\text{epistemic}}. \tag{33}$$

This decomposition yields interval estimates that combine both sources of aleatoric and epistemic uncertainty. The parameter-based loss minimizes the variance of ensemble predictions to encourage consistency, i.e., the model should produce similar predictions across different dropout masks when it is confident about a given input:

$$\mathcal{L}_{\mathrm{param}} = \frac{1}{|\mathcal{S}|} \sum_{(i,t+\tau) \in \mathcal{S}} \frac{1}{M} \sum_{m=1}^{M} \big(\mu^{(m)}_{i,t+\tau} - \widehat{\mu}_{i,t+\tau}\big)^2. \tag{34}$$

A lower value of $\mathcal{L}_{\mathrm{param}}$ indicates that the stochastic predictions are tightly clustered around their mean, reflecting lower epistemic uncertainty.

Integrated Training Objective

We jointly train all the above uncertainty heads using:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{quant}} + \mathcal{L}_{\text{nll}} + \mathcal{L}_{\text{param}} + \mathcal{L}_{\text{calib}}, \tag{35}$$

where $\mathcal{L}_{\text{quant}}$, $\mathcal{L}_{\text{nll}}$, and $\mathcal{L}_{\text{param}}$ correspond to node-based, distribution-based, and parameter-based uncertainty losses, respectively. The calibration loss $\mathcal{L}_{\text{calib}}$ encourages standardized residuals to match a standard normal distribution. Over a training minibatch $\mathcal{B}$, we compute

$$\mathcal{L}_{\text{calib}} = \Big(\tfrac{1}{|\mathcal{B}|}\sum_{(i,t+\tau)\in\mathcal{B}} r_{i,t+\tau}\Big)^2 + \Big(\tfrac{1}{|\mathcal{B}|}\sum_{(i,t+\tau)\in\mathcal{B}} r_{i,t+\tau}^2 - 1\Big)^2, \tag{36}$$

which enforces the standardized residuals to have zero mean and unit variance, thereby improving probabilistic calibration. Here, the standardized residual is defined as

$$r_{i,t+\tau} = \frac{y_{i,t+\tau} - \mu_{i,t+\tau}}{\sigma_{i,t+\tau} + \varepsilon}, \tag{37}$$

where $\varepsilon > 0$ is a small constant for numerical stability.

Post-hoc Quantile Calibration

To improve the empirical coverage of the trained quantile intervals, we apply a post-hoc quantile calibration on a held-out calibration set. Let (ℓ_i, u_i) be the predicted interval and y_i be the ground truth, we compute the empirical coverage gap:

$$\Delta_{\text{cov}} = (1-\alpha) - \frac{1}{|\mathcal{C}|}\sum_{i\in\mathcal{C}} \mathbf{1}[\ell_i \le y_i \le u_i], \tag{38}$$

where $\mathcal{C}$ is the calibration set and α is the target miscoverage level. We then compute the adjustment margin:

$$c = \text{Quantile}_{1-\alpha}\big\{\max\{\ell_i - y_i,\, y_i - u_i,\, 0\}\big\}_{i\in\mathcal{C}}. \tag{39}$$

The calibrated interval for a new prediction becomes

$$\big[\ell^{\text{cal}}, u^{\text{cal}}\big] = \big[\ell - c,\, u + c\big], \tag{40}$$

which adjusts the quantile regression intervals to achieve the target coverage level.

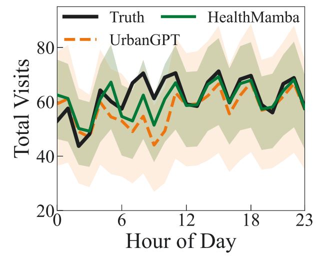
4 Evaluation

In this section, we conduct a comprehensive experimental evaluation of our proposed HealthMamba. Specifically, we aim to address the following five research questions:

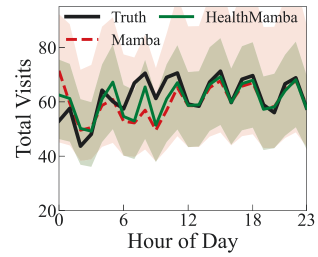
- **RQ 1**: How does HealthMamba perform compared to state-of-the-art baselines?
- **RQ 2**: Is HealthMamba effective for visit prediction of different types of healthcare facilities?
- **RQ 3**: Are all components in HealthMamba effective?
- **RQ 4**: How does HealthMamba perform under abnormal scenarios such as public emergencies (COVID-19)?

4.1 Evaluation Setup

Datasets. We utilize healthcare facility visit data from four major U.S. states (California, New York, Texas, and Florida), provided by Dewey. The dataset captures daily visit counts to four categories (i.e., Ambulatory Health Care Services, Hospitals, Nursing and Residential Care Facilities, and Social Assistance) of healthcare facilities at the county level from January 2019 to April 2022, spanning 40 months.



(a) UrbanGPT vs. HealthMamba



(b) Mamba vs. HealthMamba

Figure 2: Prediction results on the California dataset aggregated in 24 hours, with best baselines selected for visualization.

Baselines. We compare HealthMamba with 5 categories of 13 state-of-the-art baselines: (1) GNN-based: DCRNN [Li *et al.*, 2017], STGCN [Yu *et al.*, 2018], AGCRN [Bai *et al.*, 2020], DGCRN [Li *et al.*, 2023], UQGNN [Yu *et al.*, 2025]; (2) Attention-based: DSTAGNN [Lan *et al.*, 2022], ASTGCN [Zhu *et al.*, 2021]; (3) Transformer-based: GluonTS [Alexandrov *et al.*, 2020], PatchTST [Nie, 2022]; (4) LLM-based: ST-LLM [Liu *et al.*, 2024], UrbanGPT [Li *et al.*, 2024b]; (5) Mamba-based: Mamba [Gu and Dao, 2024], U-Mamba [Ma *et al.*, 2024].

Metrics. We utilize two widely used metrics, including Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), to evaluate the performance of deterministic prediction, and three other commonly used metrics, including Mean Prediction Interval Width (MPIW), Interval Score (IS), and Coverage (COV), to evaluate the performance of uncertainty quantification. It is worth noting that our model outputs prediction intervals, and the deterministic metrics are computed based on the mean of the upper and lower bounds.

4.2 Overall Performance Comparison (RQ 1)

An overall comparison of our HealthMamba and other baselines is presented in Table 1. We found that our HealthMamba reduces MAE by approximately 6.0% compared to the best baseline Mamba based on the average of all four datasets. In addition, HealthMamba also demonstrates superior performance on prediction reliability, with a 3.5% improvement in IS and reaching the target coverage. As shown in Figure 2, the green shadow of HealthMamba is more compact while still covering most observations, indicating more reliable and precise estimates compared to baselines. Our experiments show that HealthMamba consistently achieves the best performance for both short-term (next-day) and medium-term (7-day-ahead) prediction.

4.3 Type-specific Visit Prediction (RQ 2)

We further show the prediction results on different healthcare facility categories in Table 2. We found that HealthMamba consistently outperforms Mamba and UrbanGPT across all categories. Notably, Mamba fails to achieve target coverage

		California					New York				
Category	**Method**	MAE↓	RMSE↓	MPIW↓	IS↓	COV	MAE↓	RMSE↓	MPIW↓	IS↓	COV
GNN	DCRNN	30.537	98.941	76.215	229.083	✗	10.957	35.371	55.028	65.715	✓
	STGCN	32.708	95.427	120.883	203.678	✗	8.967	34.218	36.705	55.320	✗
	AGCRN	17.494	56.480	46.071	130.347	✗	6.453	21.913	17.106	47.504	✗
	DGCRN	28.408	84.729	96.559	208.084	✗	17.526	57.587	19.001	273.164	✗
	UQGNN	10.235	36.520	35.680	77.672	✓	4.585	15.820	15.230	30.426	✓
Attention	DSTAGNN	20.675	55.261	63.458	131.816	✗	8.234	29.719	32.791	49.481	✗
	ASTGCN	47.249	178.047	50.549	749.407	✗	16.679	60.544	15.101	267.103	✗
Transformer	GluonTS	16.945	44.336	87.273	107.044	✗	4.900	16.220	26.784	33.798	✓
	PatchTST	12.381	36.301	59.704	80.008	✗	5.555	20.426	29.517	37.260	✓
LLM	ST-LLM	14.391	47.347	37.231	111.233	✗	5.565	20.800	16.181	38.178	✗
	UrbanGPT	9.734	38.925	47.906	83.962	✓	4.870	17.685	20.866	31.100	✓
Mamba	Mamba	9.467	35.825	33.306	77.232	✓	4.176	**14.152**	14.174	30.608	✗
	U-Mamba	10.052	34.180	37.561	78.868	✗	4.292	14.993	16.798	29.958	✗
Ours	**HealthMamba**	**9.150**	**33.401**	**30.970**	**74.953**	✓	**3.926**	14.395	**14.110**	**29.361**	✓

		Texas					Florida				
Category	**Method**	MAE↓	RMSE↓	MPIW↓	IS↓	COV	MAE↓	RMSE↓	MPIW↓	IS↓	COV
GNN	DCRNN	25.612	178.462	34.506	344.785	✗	185.410	418.957	403.593	856.315	✗
	STGCN	22.176	164.037	75.763	165.212	✗	168.203	385.663	489.048	786.315	✗
	AGCRN	8.632	54.573	21.883	93.392	✗	114.045	277.833	468.500	643.579	✗
	DGCRN	32.224	201.640	55.081	547.595	✓	175.580	405.767	436.702	927.128	✗
	UQGNN	7.125	49.350	**23.580**	60.850	✓	72.350	175.620	259.559	465.230	✓
Attention	DSTAGNN	13.805	91.072	49.969	95.638	✗	150.137	355.974	447.363	654.617	✗
	ASTGCN	31.285	197.239	36.954	580.777	✗	170.262	418.748	357.474	871.215	✗
Transformer	GluonTS	7.502	50.466	37.472	61.071	✓	120.852	282.507	539.905	700.880	✗
	PatchTST	28.904	191.871	25.174	450.373	✗	167.780	370.929	**253.671**	845.201	✗
LLM	ST-LLM	8.921	60.350	17.686	87.761	✓	99.501	224.992	340.592	616.386	✗
	UrbanGPT	11.024	67.689	48.997	66.529	✓	58.459	158.116	307.648	436.149	✓
Mamba	Mamba	6.676	47.318	24.595	61.466	✗	85.793	209.168	386.492	567.448	✗
	U-Mamba	7.283	50.011	23.887	60.232	✗	85.256	221.271	315.136	624.394	✗
Ours	**HealthMamba**	**6.275**	**46.566**	24.015	**58.934**	✓	**54.951**	**156.535**	255.230	**420.884**	✓

Table 1: Comparison with state-of-the-art baselines on four datasets with a horizon of 3 days. ↓ indicates lower is better. The best results are in **bold** and the second-best are underlined. ✓ indicates target coverage (≥90%) achieved, while ✗ indicates failure.

in the Nursing and Social categories, while UrbanGPT exhibits substantially wider prediction intervals despite achieving coverage. In contrast, HealthMamba maintains valid and compact confidence intervals across all categories, demonstrating robustness to facility-type heterogeneity.

4.4 Ablation Study (RQ 3)

To evaluate the contribution of each key component in HealthMamba, we conduct an ablation study by comparing the full model against six variants: (1) **w/o STCE**, which removes the contextual information; (2) **w/o G-Mamba**, which replaces the GraphMamba with conventional Mamba; (3) **w/o Node-based** UQ module, (4) **w/o Distribution-based** UQ module, and (5) **w/o Parameter-based** UQ module; and (6) **w/o UQ**, which uses only quantile regression.

Table 3 validates our design choices. Architecturally, removing GraphMamba or STCE significantly degrades MAE (up to 15.782), highlighting the crucial role of spatiotemporal and contextual modeling. Furthermore, omitting any individual UQ module (Node-, Parameter-, or Distribution-based) increases MAE and fails the coverage constraint (COV = ✗). Finally, completely removing UQ fails to provide valid prediction intervals entirely (missing IS and COV), demonstrating the necessity of our synergistic UQ framework.

4.5 Prediction under Abnormal Situations (RQ 4)

A key challenge in healthcare visit prediction is handling abnormal scenarios (e.g., emergencies, extreme weather) that cause sudden pattern deviations. We evaluate HealthMamba on two representative events: the COVID-19 lockdown in LA County (March 2020) causing a sudden drop in visits, and Hurricane Hanna in Texas (July 2020) triggering a surge. As shown in Figure 3, under these situations, HealthMamba successfully maintained accurate predictions, and ground truth values are also within the prediction intervals in most cases, whereas the best baseline struggled to adapt to these abrupt

Category	Model	MAE↓	RMSE↓	MPIW↓	IS↓	COV
Hospitals	Mamba	12.384	42.156	38.720	95.832	✓
	UrbanGPT	12.058	45.623	52.418	101.254	✓
	HealthMamba	**11.628**	**39.524**	**35.105**	**91.250**	✓
Ambulatory	Mamba	8.152	31.478	29.640	68.425	✓
	UrbanGPT	8.536	34.127	42.815	73.680	✓
	HealthMamba	**7.683**	**29.352**	**27.218**	**65.127**	✓
Nursing	Mamba	7.945	28.634	31.852	62.148	✗
	UrbanGPT	8.213	32.576	45.037	70.425	✓
	HealthMamba	**7.486**	**26.815**	**29.476**	**59.268**	✓
Social	Mamba	9.387	40.032	33.012	82.523	✗
	UrbanGPT	10.127	43.374	51.352	89.489	✓
	HealthMamba	**8.803**	**37.913**	**31.081**	**78.167**	✓

Table 2: Fine-grained prediction on four healthcare facility types.

Variants	MAE↓	RMSE↓	MPIW↓	IS↓	COV
w/o STCE	13.610	48.535	39.807	108.110	✗
w/o G-Mamba	15.782	49.092	42.625	117.382	✗
w/o Node-based	9.682	34.925	32.540	80.115	✗
w/o Distribution-based	10.154	36.218	34.112	85.330	✗
w/o Parameter-based	10.573	37.665	35.890	90.224	✗
w/o UQ (Total)	11.240	39.150	37.520	-	-
HealthMamba	**9.150**	**33.401**	**30.970**	**74.953**	✓

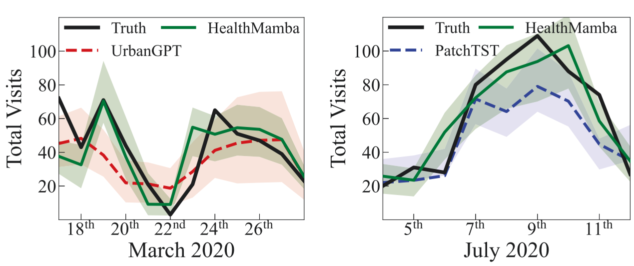
Table 3: Ablation study on the California dataset.

changes. These results highlight HealthMamba's superior adaptability to both demand surges and drops under abnormal scenarios.

5 Related Work

5.1 Healthcare Facility Visit Prediction

Healthcare facility visit prediction has attracted much interest from both academia and the government, as it is important for health resource allocation and public emergency responses. Traditional methods rely on statistical models such as ARIMA [Alghamdi *et al.*, 2019; Orhan and Kurutkan, 2025] and regression-based methods [Pun *et al.*, 2019; Avinash *et al.*, 2025; Shen *et al.*, 2025] for daily or weekly patient count prediction, but overlook temporal dynamics and spatial dependencies. Recent works have explored different approaches to address this problem, including GNN-based methods [Yu *et al.*, 2018; Li *et al.*, 2024a; Shen *et al.*, 2026], attention-based methods [Lan *et al.*, 2022], Transformer-based methods [Nie, 2022], and LLM-based methods [Li *et al.*, 2024b; Li *et al.*, 2026; Li *et al.*, 2025]. However, most existing works focus on aggregated prediction without considering different types of healthcare facilities, such as hospitals and nursing facilities, which is more important for practical decision support. In addition, existing solutions usually consider this problem a time-series prediction problem, without capturing spatial dependencies, which can potentially be used to improve prediction performance.



(a) California: COVID-19. (b) Texas: Hurricane Hanna.

Figure 3: Prediction results under extreme events.

5.2 Uncertainty-aware Spatiotemporal Prediction

Uncertainty quantification has attracted much interest from the spatiotemporal prediction community because it lays the foundation for reliable and safe decision-making [Jiang *et al.*, 2025b; Yu *et al.*, 2026; Cheng *et al.*, 2025]. DeepSTUQ [Qian *et al.*, 2023] employs dual sub-networks to separately quantify aleatoric and epistemic uncertainty at each node. DiffSTG [Wen *et al.*, 2023] generalizes denoising diffusion probabilistic models to spatiotemporal graphs to capture intrinsic uncertainties. STZINB-GNN [Zhuang *et al.*, 2022] models sparse spatiotemporal data through zero-inflated negative binomial distributions. CF-GNN [Huang *et al.*, 2024] extends conformal prediction to graph-based models for guaranteed coverage, providing distribution-free prediction intervals. UQGNN [Yu *et al.*, 2025] integrates graph neural networks with node-level uncertainty estimation for spatiotemporal prediction. Different from existing works [Yang *et al.*, 2023; Yang *et al.*, 2024], we design a comprehensive uncertainty quantification module that integrates distribution-based, parameter-based, and node-based uncertainty quantification mechanisms, combined with post-hoc quantile calibration to produce well-calibrated prediction intervals for uncertainty-aware spatiotemporal prediction.

6 Conclusion

In this paper, we propose HealthMamba, an uncertainty-aware spatiotemporal framework for accurate and reliable healthcare facility visit prediction. There are three key components in HealthMamba: (i) a Unified Spatiotemporal Context Encoder that integrates both heterogeneous static and dynamic contextual information, (ii) a novel graph state space model called GraphMamba for spatiotemporal modeling, and (iii) a comprehensive UQ module integrating three different types of UQ mechanisms and a post-hoc quantile calibration for higher prediction reliability. Extensive experiments on four real-world datasets from California, New York, Texas, and Florida demonstrate that HealthMamba outperforms 13 state-of-the-art baselines, improving prediction accuracy by 6.0% and uncertainty quantification by 3.5%. Our uncertainty-aware design also enables robust predictions under public emergencies and extreme weather events, supporting trustworthy decision-making under abnormal scenarios.

Acknowledgments

We sincerely thank all anonymous reviewers for their insightful comments and valuable suggestions. This work is partially supported by the National Science Foundation under Grant 2411152 and the FSU Institute for Successful Longevity (ISL) Planning Grant.

Ethics Statement

All authors comply with the IJCAI Code of Ethics and take full responsibility for this research. We use only de-identified data without human subjects or sensitive information, posing no foreseeable harm. There are no conflicts of interest or external sponsorships.

References

[Alexandrov *et al.*, 2020] Alexander Alexandrov, Konstantinos Benidis, Michael Bohlke-Schneider, Valentin Flunkert, Jan Gasthaus, Tim Januschowski, Danielle C Maddix, Syama Rangapuram, David Salinas, Jasper Schulz, et al. Gluonts: Probabilistic and neural time series modeling in python. *Journal of Machine Learning Research*, 21(116):1–6, 2020.

[Alghamdi *et al.*, 2019] Taghreed Alghamdi, Khalid Elgazzar, Magdi Bayoumi, Taysseer Sharaf, and Sumit Shah. Forecasting traffic congestion using arima modeling. In *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC)*, pages 1227–1232, 2019.

[Avinash *et al.*, 2025] G Avinash, Hariom Pachori, Avinash Sharma, and SukhDev Mishra. Time series forecasting of bed occupancy in mental health facilities in india using machine learning. *Scientific Reports*, 15(1):2686, 2025.

[Bai *et al.*, 2020] Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. Adaptive graph convolutional recurrent network for traffic forecasting. *Advances in neural information processing systems*, 33:17804–17815, 2020.

[Cheng *et al.*, 2025] Xueqi Cheng, Catherine Yang, Yuying Zhao, Yu Wang, Hamid Karimi, and Tyler Derr. Bts: A comprehensive benchmark for tie strength prediction. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5345–5354, 2025.

[Gu and Dao, 2024] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*, 2024.

[Huang *et al.*, 2024] Kexin Huang, Ying Jin, Emmanuel Candes, and Jure Leskovec. Uncertainty quantification over graph with conformalized graph neural networks. *Advances in Neural Information Processing Systems*, 36, 2024.

[Jiang *et al.*, 2025a] Lin Jiang, Yu Yang, and Guang Wang. Hcride: harmonizing passenger fairness and driver preference for human-centered ride-hailing. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 10289–10297, 2025.

[Jiang *et al.*, 2025b] Lin Jiang, Dahai Yu, Rongchao Xu, Tian Tang, and Guang Wang. Uncertainty-aware predict-then-optimize framework for equitable post-disaster power restoration. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 9719–9727, 2025.

[Lan *et al.*, 2022] Shiyong Lan, Yitong Ma, Weikang Huang, Wenwu Wang, Hongyu Yang, and Pyang Li. Dstagnn: Dynamic spatial-temporal aware graph neural network for traffic flow forecasting. In *International conference on machine learning*, pages 11906–11917. PMLR, 2022.

[Li *et al.*, 2017] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*, 2017.

[Li *et al.*, 2023] Fuxian Li, Jie Feng, Huan Yan, Guangyin Jin, Fan Yang, Funing Sun, Depeng Jin, and Yong Li. Dynamic graph convolutional recurrent network for traffic prediction: Benchmark and solution. *ACM Transactions on Knowledge Discovery from Data*, 17(1):1–21, 2023.

[Li *et al.*, 2024a] Lincan Li, Kaixiang Yang, Jichao Bi, and Fengji Luo. Sts-ccl: Spatial-temporal synchronous contextual contrastive learning for urban traffic forecasting. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6705–6709, 2024.

[Li *et al.*, 2024b] Zhonghang Li, Lianghao Xia, Jiabin Tang, Yong Xu, Lei Shi, Long Xia, Dawei Yin, and Chao Huang. Urbangpt: Spatio-temporal large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5351–5362, 2024.

[Li *et al.*, 2025] Lincan Li, Eren Erman Ozguven, Yue Zhao, Guang Wang, Yiqun Xie, and Yushun Dong. Typhoformer: Language-augmented transformer for accurate typhoon track forecasting. In *Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '25, page 1174–1177, New York, NY, USA, 2025. Association for Computing Machinery.

[Li *et al.*, 2026] Lincan Li, Zheng Chen, and Yushun Dong. Llm as clinical graph structure refiner: Enhancing representation learning in eeg seizure diagnosis, 2026.

[Liu *et al.*, 2024] Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. St-llm: Large language models are effective temporal learners. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024.

[Ma *et al.*, 2024] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024.

[Marcusson *et al.*, 2020] Jan Marcusson, Magnus Nord, Huan-Ji Dong, and Johan Lyth. Clinically useful prediction of hospital admissions in an older population. *BMC geriatrics*, 20(1):95, 2020.

[Nie, 2022] Y Nie. A time series is worth 64words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

[Orhan and Kurutkan, 2025] Fatih Orhan and Mehmet Nurullah Kurutkan. Predicting total healthcare demand using machine learning: separate and combined analysis of predisposing, enabling, and need factors. *BMC Health Services Research*, 25(1):366, 2025.

[Piccialli *et al.*, 2020] Francesco Piccialli, Salvatore Cuomo, Danilo Crisci, Edoardo Prezioso, and Gang Mei. A deep learning approach for facility patient attendance prediction based on medical booking data. *Scientific Reports*, 10(1):14623, 2020.

[Pun *et al.*, 2019] Lilian Pun, Pengxiang Zhao, and Xintao Liu. A multiple regression approach for traffic flow estimation. *IEEE Access*, 7:35998–36009, 2019.

[Qian *et al.*, 2023] Weizhu Qian, Dalin Zhang, Yan Zhao, Kai Zheng, and James JQ Yu. Uncertainty quantification for traffic forecasting: A unified approach. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 992–1004. IEEE, 2023.

[Shen *et al.*, 2025] Bolin Shen, Eren Ozguven, Yue Zhao, Guang Wang, Yiqun Xie, and Yushun Dong. Learning from the storm: A multivariate machine learning approach to predicting hurricane-induced economic losses. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Spatial Intelligence for Smart and Connected Communities*, pages 1–4, 2025.

[Shen *et al.*, 2026] Bolin Shen, Md Shamim Seraj, Zhan Cheng, Shayok Chakraborty, and Yushun Dong. Cited: A decision boundary-aware signature for gnns towards model extraction defense. *arXiv preprint arXiv:2602.20418*, 2026.

[Wen *et al.*, 2023] Haomin Wen, Youfang Lin, Yutong Xia, Huaiyu Wan, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. Diffstg: Probabilistic spatio-temporal graph forecasting with denoising diffusion models. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, pages 1–12, 2023.

[Xiao and Liu, 2025] Canran Xiao and Yongmei Liu. A multifrequency data fusion deep learning model for carbon price prediction. *Journal of Forecasting*, 44(2):436–458, 2025.

[Xu *et al.*, 2025] Rongchao Xu, Zhiqing Hong, and Guang Wang. Autostdiff: Autoregressive spatio-temporal denoising diffusion model for asynchronous trajectory generation. In *Proceedings of the 2025 SIAM International Conference on Data Mining (SDM)*, pages 538–547. SIAM, 2025.

[Xu *et al.*, 2026a] Rongchao Xu, Kunlin Cai, Lin Jiang, Zhiqing Hong, Yuan Tian, and Guang Wang. Geogen: A two-stage coarse-to-fine framework for fine-grained synthetic location-based social network trajectory generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 1373–1381, 2026.

[Xu *et al.*, 2026b] Rongchao Xu, Lin Jiang, Dahai Yu, Ximiao Li, and Guang Wang. Synhat: A two-stage coarse-to-fine diffusion framework for synthesizing human activity traces. *arXiv preprint arXiv:2604.14705*, 2026.

[Yang *et al.*, 2023] Peiyu Yang, Naveed Akhtar, Zeyi Wen, Mubarak Shah, and Ajmal Saeed Mian. Re-calibrating feature attributions for model interpretation. In *International Conference on Learning Representations*, 2023.

[Yang *et al.*, 2024] Peiyu Yang, Naveed Akhtar, Mubarak Shah, and Ajmal Mian. Regulating model reliance on non-robust features by smoothing input marginal density. In *European Conference on Computer Vision*, pages 329–347. Springer, 2024.

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

[Yu *et al.*, 2025] Dahai Yu, Dingyi Zhuang, Lin Jiang, Rongchao Xu, Xinyue Ye, Yuheng Bu, Shenhao Wang, and Guang Wang. Uqgnn: Uncertainty quantification of graph neural networks for multivariate spatiotemporal prediction. In *Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems*, pages 52–65, 2025.

[Yu *et al.*, 2026] Dahai Yu, Rongchao Xu, Dingyi Zhuang, Yuheng Bu, Shenhao Wang, and Guang Wang. Trustenergy: A unified framework for accurate and reliable user-level energy usage prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(46):39558–39566, Mar. 2026.

[Zhong *et al.*, 2026] Yibo Zhong, Haoxiang Jiang, Lincan Li, Ryumei Nakada, Tianci Liu, Linjun Zhang, Huaxiu Yao, and Haoyu Wang. Peanut: Parameter-efficient adaptation with weight-aware neural tweakers. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026.

[Zhu *et al.*, 2021] Jiawei Zhu, Qiongjie Wang, Chao Tao, Hanhan Deng, Ling Zhao, and Haifeng Li. Ast-gcn: Attribute-augmented spatiotemporal graph convolutional network for traffic forecasting. *Ieee Access*, 9:35973–35983, 2021.

[Zhuang *et al.*, 2022] Dingyi Zhuang, Shenhao Wang, Haris Koutsopoulos, and Jinhua Zhao. Uncertainty quantification of sparse travel demand prediction with spatial-temporal graph neural networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 4639–4647, New York, NY, USA, 2022. Association for Computing Machinery.