

Article

Multi-Regional Infectious Disease Transmission Forecasting Based on Graph-Structure-Enhanced Large Language Model

Siru Chen ^{1,†}, Xuhao Guo ^{1,†}, Zige Liu ^{2,†}, Zhijie Lin ² and Junying Chen ^{1,*}

¹ School of Software Engineering, South China University of Technology, Guangzhou 510006, China; 202230480205@mail.scut.edu.cn (S.C.); sexuhao.guo@mail.scut.edu.cn (X.G.)

² Respiratory Disease AI Laboratory in Epidemic Intelligence and Medical Big Data Applications, Faculty of Innovation Engineering, Macau University of Science and Technology, Macau SAR, China; zgliu@must.edu.mo (Z.L.); zhijie_jim@foxmail.com (Z.L.)

* Correspondence: jychense@scut.edu.cn

† These authors contributed equally to this work.

Abstract

Infectious disease outbreaks strain medical resources and public-health systems, making accurate multi-horizon, multi-regional forecasting essential for early warning and resource allocation. Existing approaches face a trade-off: data-driven deep models can capture nonlinear spatiotemporal patterns but often overlook transmission mechanisms, whereas mechanistic models are interpretable but limited in modeling complex regional dependencies. To address this challenge, we propose Epidemic Spatial–Temporal Large Language Model (EpiSTLLM), which is a graph-structure-enhanced large language model for multi-regional infectious disease forecasting. EpiSTLLM injects regional adjacency information into Transformer-based representation learning to capture cross-regional transmission structure and long-term temporal dependencies. A temporal-gated cross-attention module generates horizon-specific latent transmission and recovery parameters, while a latent-space SIR-inspired propagation mechanism with a residual correction branch enables stable multi-horizon forecasting without requiring fully observed compartmental states. Experiments on the FluView state-level influenza-like illness dataset and NHSN state-level influenza hospitalization dataset show that EpiSTLLM achieves the best or highly competitive performance in most evaluation settings against statistical, deep learning, graph-based, and mechanism-guided baselines across 4-, 8-, and 12-week horizons. For example, EpiSTLLM reduces MAE and RMSE values by 9.5% and 6.2% at $H = 4$ on FluView, and by 12.0% and 13.1% at $H = 8$ on NHSN compared with the strongest baselines, respectively.

Keywords: global public health security; spatiotemporal large language model; intelligent epidemic forecasting; graph structure enhancement; integrating infectious disease mechanisms with AI



Academic Editor: Victor Mitrana

Received: 31 May 2026

Revised: 9 July 2026

Accepted: 17 July 2026

Published: 20 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

MSC: 92D30; 92B20

1. Introduction

Infectious disease outbreaks continue to have profound impacts on the operation of healthcare systems, public health responses, and socio-economic activities [1,2]. Taking seasonal influenza as an example, there are approximately 291,000 to 646,000 respiratory deaths worldwide each year related to it [3]. The global spread of emerging infectious diseases

such as COVID-19 has further exposed the serious consequences of strained healthcare resources, uncontrolled regional transmission, and delayed public health responses [4,5]. To support vaccine allocation, healthcare resource scheduling, and early warning in high-risk areas, constructing an accurate and stable multi-regional infectious disease transmission forecasting model has become a key issue in intelligent public health decision making [6,7]. Compared to single-step short-term prediction, multi-step prediction can provide a longer look-ahead window for formulating intervention strategies and therefore has greater practical significance in public-health decision making [8,9]. However, as the prediction span increases, uncertainties, regional heterogeneity, and error accumulation in the transmission process significantly intensify, making long-term prediction one of the most challenging tasks in the current field of infectious disease prediction [8–10].

Multi-step and multi-regional infectious disease prediction is essentially a complex spatiotemporal dynamic modeling problem [11]. On the one hand, disease activity exhibits significant time-dependent characteristics, and its transmission trends are influenced by various factors such as seasonal variations, human contact behaviors, and public health interventions [12,13]. On the other hand, there are also spatial transmission correlations between different regions driven by geographical proximity, population mobility, and social connections. Therefore, the model not only needs to learn the long-term temporal evolution patterns within a single region but also needs to characterize the dynamically changing transmission influence relationships between regions. At the same time, real-world infectious disease surveillance data often suffer from issues such as reporting delays, statistical noise, and incomplete observations. Under limited sample size and high noise conditions, models that rely solely on historical data fitting are prone to overfitting and decreased generalizability [14]. Furthermore, prediction systems in public health scenarios not only require high accuracy but also possess a certain degree of mechanism explanatory power to provide a reliable basis for transmission risk analysis and intervention decision making. Hence, how to strike a balance between complex spatiotemporal modeling, long-term stable prediction, and transmission mechanism constraints has become a core issue in current research [15].

Early research relied primarily on statistical time-series models and epidemiological mechanistic models. Statistical models learn temporal correlations through historical observational data, while mechanism models utilize dynamic equations such as susceptible–infectious–recovered (SIR) and susceptible–exposed–infectious–recovered (SEIR) to describe the disease transmission process [16]. Although mechanism models have strong epidemiological interpretability, they often rely on idealized assumptions such as homogeneous mixing, making it difficult to depict complex nonlinear transmission patterns and regional transmission heterogeneity in real-world environments. With the development of deep learning, research has gradually shifted toward data-driven nonlinear spatiotemporal modeling. Methods based on long short-term memory (LSTM), convolutional neural network (CNN), and graph neural network (GNN) can learn complex spatiotemporal dependencies and achieve good results in short-term predictions. However, most of these models lack explicit modeling of disease transmission mechanisms, which can lead to error accumulation in long-term rolling predictions [6,8]. Additionally, the implicit patterns learned by these models often lack clear epidemiological implications, limiting the credibility and interpretability of the models in public health decision-making scenarios [17,18].

In recent years, mechanism-guided deep learning models have begun to attempt to integrate propagation dynamics with neural network modeling capabilities. One approach utilizes deep learning to dynamically estimate time-varying epidemiological parameters such as transmission rate and recovery rate; the other approach embeds mechanism constraints into neural network structures or hidden space evolution processes, thus enhancing

long-term prediction stability and model interpretability [19]. Meanwhile, large language models (LLMs) have demonstrated strong potential in sequence representation learning, context modeling, and cross-task transfer, and they are gradually being introduced into the field of time-series prediction. However, existing general spatiotemporal large models mainly focus on learning statistical correlations and have not fully considered the dynamic constraints and regional propagation mechanisms in infectious disease transmission; in addition, simply expanding multi-regional spatiotemporal data into sequences of tokens also makes it difficult to effectively model non-Euclidean spatial propagation structures. Therefore, current research still lacks a unified prediction framework that can simultaneously integrate the regional graph structure, the long-term temporal representation ability of LLMs, and constraints of the epidemiological mechanism [20,21].

To address these challenges, we propose an *Epidemic Spatial–Temporal Large Language Model (EpiSTLLM)*, which is a multi-regional, multi-horizon infectious disease forecasting framework that integrates graph-structured modeling, LLM-based spatiotemporal representations, and latent-space propagation constraints. EpiSTLLM jointly models inter-regional transmission connectivity, long-range temporal dependencies, and the evolution of disease dynamics, thereby improving the stability and interpretability of long-horizon forecasts. The main contributions of this work are summarized as follows:

- We develop a graph-structure-enhanced LLM spatiotemporal representation method by injecting regional adjacency into Transformer attention, enabling the model to learn long-range temporal context while explicitly capturing cross-regional propagation structure. We further adopt parameter-efficient low-rank adaption (LoRA) fine-tuning to improve adaptability and training efficiency.
- We design a temporal-gated cross-attention parameter generation module that dynamically produces latent propagation parameters via future-horizon queries, adaptive temporal gating, and cross-attention, enhancing robustness to horizon-dependent shifts in propagation patterns.
- We propose a latent-space SIR-inspired propagation forecasting scheme that enables long-horizon rolling prediction without fully observing mechanistic states, by coupling graph-neighborhood infection pressure with latent state transitions, improving stability while providing mechanistic interpretability.

Extensive experiments on the FluView state-level influenza-like illness dataset and the NHSN state-level influenza hospitalization dataset demonstrate that EpiSTLLM achieves the best or highly competitive performance across most evaluation settings for 4-, 8-, and 12-week-ahead forecasting with particularly strong gains in long-horizon robustness, complex spatiotemporal dynamics modeling, and inter-regional transmission trend characterization.

2. Related Work

2.1. Traditional Statistical and Mechanistic Models

Infectious disease forecasting has long relied on (i) statistical time-series models that extrapolate future incidence from historical observations and (ii) mechanistic compartmental models that describe transmission through interpretable state transitions. Mechanistic models partition a population into compartments and represent the transmission process via dynamical laws. Using the classic SIR setting as an example, the key idea is that new infections grow with effective contact between susceptible and infectious individuals, while infectious individuals leave the infectious compartment at a recovery rate. A representative core equation is

$$\frac{dI(t)}{dt} = \beta \frac{S(t)I(t)}{N} - \gamma I(t), \quad (1)$$

where β and γ correspond to transmission and recovery rates, respectively. Although such models offer epidemiological interpretability, simplified assumptions (e.g., homogeneous mixing) and unmodeled factors (latent infection, waning immunity, severity differences) often motivate extensions such as SEIR and susceptible–infectious–recovered–deceased (SIRD); these improvements increase expressiveness but can aggravate identifiability and calibration under noisy, partially observed surveillance data [22].

2.2. Data-Driven Deep Learning for Epidemic Forecasting

Deep learning methods improve forecasting accuracy by learning nonlinear temporal and spatial dependencies directly from data. Recurrent architectures (e.g., LSTM [23], gated recurrent unit (GRU) [24]) and convolutional designs have been widely used for temporal modeling, while GNNs incorporate inter-regional coupling for multi-regional prediction (e.g., Cola-GNN [1], STAN [25], EpiGNN [26]). However, purely data-driven models may ignore epidemiological mechanisms and can suffer from error accumulation in long-horizon rolling forecasts.

Mechanism-guided deep learning attempts to combine data fitting with mechanistic consistency. Instead of explicitly writing out full compartmental ordinary differential equation systems, many approaches let neural networks infer time-varying latent rates or latent states and then regularize training so that predictions remain compatible with a chosen mechanistic prior. A commonly used formulation is a composite objective:

$$\mathcal{L} = \mathcal{L}_{\text{data}} + \lambda \mathcal{L}_{\text{mech}}, \quad (2)$$

where $\mathcal{L}_{\text{data}}$ fits observed surveillance targets, $\mathcal{L}_{\text{mech}}$ penalizes deviations from mechanistic dynamics, and λ controls the strength of the constraint. Despite promising results, many hybrid methods still rely on (partially) observable compartment states or strong assumptions that may not hold for real-world indirect surveillance streams.

2.3. Graph Structure Modeling and Graph Neural Networks

Multi-regional epidemic dynamics are coupled through mobility, geographic proximity and socio-economic connections, making graphs a natural abstraction for cross-regional transmission [27,28]. GNNs learn regional embeddings by repeatedly aggregating information from neighboring regions, enabling the model to capture multi-hop spatial dependence. A generic message-passing update can be written as

$$\mathbf{h}_i^{(l+1)} = \phi \left(\mathbf{W}_{\text{self}}^{(l)} \mathbf{h}_i^{(l)} + \sum_{j \in \mathcal{N}(i)} \alpha_{ij} \mathbf{W}_{\text{neigh}}^{(l)} \mathbf{h}_j^{(l)} \right), \quad (3)$$

where $\mathbf{h}_i^{(l)}$ is the representation of region i at layer l , $\mathcal{N}(i)$ is its neighborhood, and α_{ij} specifies how strongly region j influences region i (from fixed normalization as in graph convolutional networks (GCNs) [29] or learned attention as in graph attention networks (GATs) [30]). This family of models provides a flexible way to encode inter-regional dependence but may still require additional temporal modeling and mechanistic constraints to achieve stable long-horizon forecasting.

2.4. Infectious Disease Prediction and Parameter Efficient Adaptation Technology Driven by Large Language Models

Attention is the core computation in LLMs, enabling each token representation to retrieve the most relevant context via pairwise similarity [31]. Systems advances such as FlashAttention accelerate exact attention through improved memory access [32,33], while broader surveys summarize efficient attention variants [34]. Given a sequence of length

n , linear projections form queries $\mathbf{Q} \in \mathbb{R}^{n \times d_k}$, keys $\mathbf{K} \in \mathbb{R}^{n \times d_k}$ and values $\mathbf{V} \in \mathbb{R}^{n \times d_v}$. The essential operation is scaled dot-product attention:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}. \quad (4)$$

In practice, Transformers usually employ multi-head self-attention, residual connections, layer normalization and a position-wise feed-forward sublayer to increase expressivity, while causal masking in decoder-only LLMs enforces autoregressive dependence.

Recent studies show that pretrained LLM representations can transfer to time-series forecasting by converting continuous observations into token-like embeddings and leveraging long-context sequence modeling. Representative directions include reprogramming time series into pretrained token spaces [35], prompting to elicit temporal semantics [36], patch-based tokenization for long-range dependencies [37], and zero-shot evaluation [38]. For spatiotemporal tasks, ST-LLM+ injects spatial adjacency into Transformer layers for traffic prediction [39]. Although recent studies have begun to explore the application of LLMs to infectious disease prediction [40,41], the integration of graph-structured spatial dependency relationships and epidemiological mechanism constraints into LLM-based prediction frameworks remains largely unexplored. Specifically, PandemicLLM [40] demonstrates the feasibility of LLMs in multimodal prediction, but it does not incorporate explicit spatial interaction modeling; CAPE [41] introduces a pretraining strategy based on interval models but does not utilize graph-structured regional dependency relationships. The EpiSTLLM proposed in this paper uniquely integrates graph-structured bias injection, temporal-gated cross-attention parameter generation, and a potential spatial SIR mechanism into a unified LLM framework for multi-regional infectious disease prediction.

Furthermore, the full fine-tuning of LLMs is often computationally expensive and can overfit under limited public-health data. Parameter-efficient fine-tuning addresses this by freezing most pretrained weights and learning a small number of additional parameters. A widely used approach is LoRA [42], which adds a low-rank update to a pretrained matrix and yields the forward computation:

$$\mathbf{y} = (\mathbf{W}_0 + \mathbf{W}_{\text{up}}\mathbf{W}_{\text{down}})\mathbf{x}. \quad (5)$$

Here, $\mathbf{W}_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ is the frozen pretrained weight matrix, $\mathbf{W}_{\text{down}} \in \mathbb{R}^{r \times d_{\text{in}}}$ projects to a matrix of lower rank (r), and $\mathbf{W}_{\text{up}} \in \mathbb{R}^{d_{\text{out}} \times r}$ projects back to the original dimension. This reduces trainable parameters from $d_{\text{in}} \times d_{\text{out}}$ to $r(d_{\text{in}} + d_{\text{out}})$ with $r \ll \min(d_{\text{in}}, d_{\text{out}})$ while largely preserving pretrained knowledge. Building on this idea, QLoRA combines low-bit quantization with low-rank updates to further reduce memory requirements [43].

3. The Proposed Method

3.1. Problem Formulation

Multi-regional infectious disease forecasting aims to predict epidemic activity across multiple locations over a future horizon by leveraging historical surveillance time series together with their spatial couplings. Unlike single-regional time-series prediction, this task is inherently spatiotemporal: temporal dependence within each region coexists with spatial correlations induced by inter-regional transmission, and the problem is naturally formulated as forecasting on a graph.

Consider N regions, each represented as a node, $V = \{v_1, v_2, \dots, v_N\}$. The inter-regional relationships are encoded by a graph $G = (V, E, A)$, where E denotes the set of edges capturing potential transmission links and $A \in \mathbb{R}^{N \times N}$ is the adjacency ma-

trix. At time step t , observations across regions are collected in $X^t \in \mathbb{R}^{N \times F}$, where F is the feature dimension for each node. Given a historical input sequence $X_{t-T+1:t} = \{X^{t-T+1}, X^{t-T+2}, \dots, X^t\}$, multi-horizon forecasting seeks to learn a predictor $f_\theta(\cdot)$ that outputs predictions for H future lead times across all regions, $\hat{Y}_{t+1:t+H} = f_\theta(X_{t-T+1:t}, A)$, where $\hat{Y}_{t+1:t+H} \in \mathbb{R}^{H \times N \times C}$ and C is the target dimension. Accurate multi-regional forecasts provide an evidence base for situational awareness, risk early warning and public-health resource allocation.

3.2. Overall Framework of EpiSTLLM

As shown in Figure 1, the proposed EpiSTLLM in this paper constructs a unified multi-regional epidemic prediction framework that integrates graph-structured spatiotemporal modeling, LLM representation learning, and constraints on the epidemic transmission mechanism in latent space. The framework aims to simultaneously address issues such as insufficient modeling of regional transmission correlation, difficulty in capturing long-term temporal dependencies, and weak stability in multi-horizon predictions. Unlike traditional time-series models that focus solely on local dynamics within a single region, EpiSTLLM represents the regional epidemic system as a dynamic graph structure, jointly modeling inter-regional transmission coupling relationships and long-term temporal evolution patterns in a unified latent space.

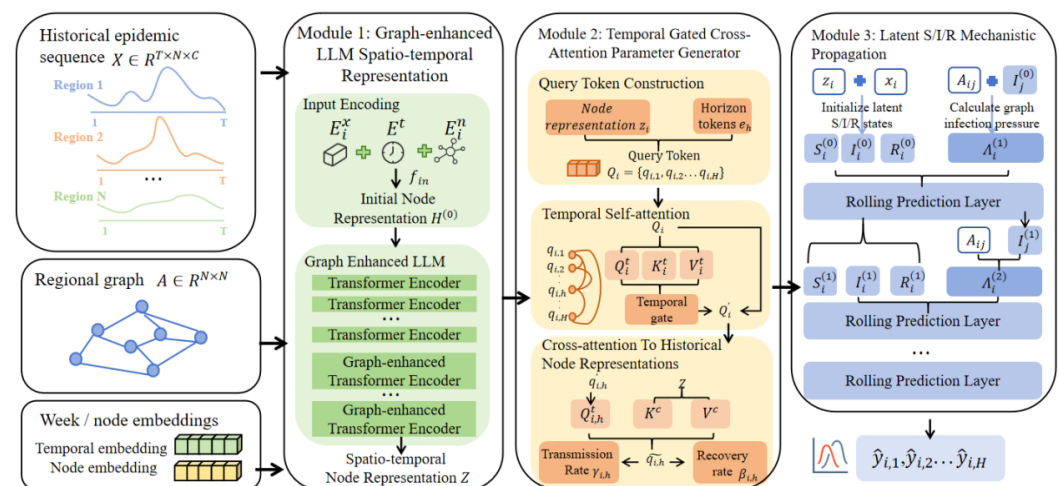


Figure 1. Overall architecture of EpiSTLLM.

Specifically, the framework first utilizes a graph-enhanced spatiotemporal representation module of an LLM to jointly encode historical epidemic observation sequences, time embeddings, and regional node embeddings. It explicitly injects regional adjacency relationships into the Transformer attention propagation process, enabling the pretrained LLM to perceive cross-regional transmission dependencies during long-term temporal context modeling, thereby learning node spatiotemporal representations with spatial transmission semantics. Subsequently, the model further designs a temporal-gated cross-attention parameter generation mechanism, constructs dynamic query representations through future prediction lead time embeddings, and utilizes a temporal-gated structure to adjust the intensity of information interaction between different prediction stages. This ensures the continuity of future trends while avoiding error coupling and information collapse in multi-horizon predictions. On this basis, the framework dynamically extracts the most relevant contextual information from historical node representations through a cross-attention mechanism, generating latent space transmission parameters and recovery parameters for different prediction lead times, achieving an adaptive modeling of transmission dynamics over time. Finally, to enhance the explanatory power of the prediction process, this

paper does not directly rely on traditional explicit SIR differential equations. Instead, it constructs a learnable S/I/R state evolution process in the neural network latent space, driving the rolling update of hidden states through the combined influence of regional graph neighborhood infection pressure and dynamic transmission parameters. This enables multi-regional rolling predictions that combine data-driven expressive power with epidemiological transmission constraints. Overall, EpiSTLLM realizes a deep coupling of graph-structure transmission correlation, long-term LLM temporal understanding, and latent space mechanism transmission deduction within a unified framework, providing a new modeling paradigm for credible multi-horizon predictions in complex non-stationary epidemic systems.

3.3. Graph-Enhanced Spatiotemporal Representation Module for Large Language Models

In multi-regional infectious disease transmission forecasting, models must not only capture temporal evolution within each region but also learn the spatial correlation structure induced by geographical adjacency, population mobility and diffusion-driven spread. Conventional time-series models typically focus on temporal dependence among historical observations while neglecting transmission coupling across regions. Meanwhile, although standard LLMs excel at modeling long-context sequences, their attention mechanisms were designed for one-dimensional natural-language streams and do not explicitly perceive non-Euclidean, graph-structured spatial dependencies. Here, we develop a graph-structure-enhanced LLM spatiotemporal representation framework that preserves the long-range dependency modeling benefits of pretrained Transformers while injecting the regional transmission graph as an explicit spatial prior into the attention propagation process, thereby enabling a unified modeling of inter-regional transmission relations and long-term temporal dynamics.

Specifically, we denote the historical surveillance sequence as $X \in \mathbb{R}^{B \times T \times N \times C}$, where B is the batch size, T is the length of the historical window, N is the number of regional nodes, and C is the feature dimension. For a regional node v_i , the model first maps its past T -week surveillance sequence through a historical-sequence encoder to obtain a node-level contextual representation E_i^x . Because respiratory diseases such as influenza often exhibit pronounced seasonality and periodicity, we further introduce a temporal embedding E^t to encode week-level positions. To improve the discrimination of static inter-regional differences, we additionally assign each region a learnable node embedding E_i^n . These components are fused to form the initial node-wise spatiotemporal representation:

$$H^{(0)} = f_{\text{in}}([E_i^x; E^t; E_i^n]), \quad (6)$$

where $f_{\text{in}}(\cdot)$ denotes the input fusion mapping. This design ensures that each node token contains not only local historical trends but also temporal periodicity and regional identity information, providing a unified input for subsequent large-model spatiotemporal representation learning.

Given the initial node-token representations, we perform spatiotemporal context modeling using a graph-enhanced Transformer encoder. Unlike a vanilla Transformer that relies purely on data-driven learning of implicit correlations, we incorporate the regional adjacency matrix A directly into the attention computation so that the model can explicitly sense spatial connectivity during contextual propagation. As illustrated in Figure 2, at

the l -th layer of the graph-enhanced Transformer encoder, the input node representations $H^{(l-1)}$ are first normalized and then projected to queries, keys and values:

$$\begin{aligned} Q^{(l)} &= H^{(l-1)} W_Q^{(l)}, \\ K^{(l)} &= H^{(l-1)} W_K^{(l)}, \\ V^{(l)} &= H^{(l-1)} W_V^{(l)}. \end{aligned} \quad (7)$$

We then construct a graph-structure bias term B_A from the adjacency matrix and add it to the attention logits:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + B_A\right)V, \quad (8)$$

where $B_A = \phi(A)$ is the bias mapped from A and $\phi(\cdot)$ is a graph-structure mapping function. This mechanism assigns higher relevance to spatially adjacent regions during attention propagation, strengthening the model's ability to capture inter-regional transmission dependence. Compared with a standard Transformer that depends only on implicit contextual correlation, this approach more robustly recovers spatial diffusion patterns in real transmission networks.

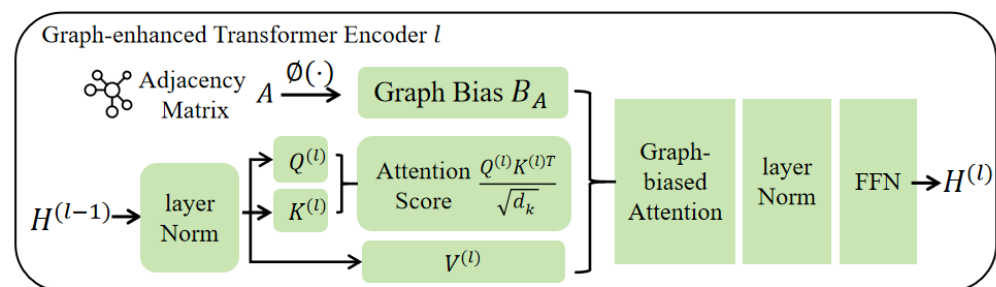


Figure 2. Schematic diagram of graph-enhanced large language model representation layer.

Because pretrained LLMs are typically large, whereas public-health surveillance datasets are often limited in size, full-parameter fine-tuning can be computationally expensive and prone to overfitting. We therefore adopt LoRA for parameter-efficient adaptation, introducing low-rank trainable updates in selected attention layers:

$$\Delta W = \mathbf{W}_{\text{up}} \mathbf{W}_{\text{down}}, \quad (9)$$

where $\mathbf{W}_{\text{down}} \in \mathbb{R}^{r \times d_{\text{in}}}$ and $\mathbf{W}_{\text{up}} \in \mathbb{R}^{d_{\text{out}} \times r}$ are the trainable low-rank matrices (with $r \ll \min(d_{\text{in}}, d_{\text{out}})$) applied to a frozen base weight $W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$. By freezing most pretrained parameters and optimizing only these low-rank increments, the model retains general sequence-modeling capacity while adapting efficiently to multi-regional infectious disease forecasting.

In summary, the proposed graph-enhanced LLM spatiotemporal representation framework unifies historical temporal dynamics, inter-regional transmission associations and the contextual representation power of pretrained Transformers. Self-attention effectively learns long-term epidemic evolution, whereas graph-structure bias injects a spatial prior so that node representations encode both temporal context and latent transmission dependence between regions. The resulting node spatiotemporal representation Z serves as the input to the subsequent temporal-gated cross-attention parameter generation module and the latent-space S/I/R mechanism-driven propagation module, providing a more stable

and mechanistically interpretable spatiotemporal basis for multi-horizon, multi-regional epidemic forecasting.

3.4. Temporal-Gated Cross-Attention Parameter Generation Module

Transmission dynamics can differ markedly across forecast lead times. Near-term forecasts are often dominated by the persistence of recent trends, whereas medium- to long-range trajectories may be driven by seasonal forcing and spatial diffusion across regions. Consequently, using a single fixed parameter set, or generating all future parameters in one shot via a simple mapping, may fail to capture lead-time-specific heterogeneity. We therefore assign each lead time an independent query token and introduce a temporally gated cross-attention mechanism that (i) models dependencies among future horizons and (ii) extracts transmission-relevant information from historical node representations to dynamically generate latent-space propagation parameters for each lead time.

Let the node-level spatiotemporal representations produced by the graph-enhanced LLM be

$$\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_N]^\top \in \mathbb{R}^{N \times d}, \quad (10)$$

where N is the number of regions, $\mathbf{z}_i \in \mathbb{R}^d$ is the historical spatiotemporal embedding of region i , and d is the representation dimension. To couple region-specific state information with a cohort-level trend, we first aggregate all node embeddings by mean pooling and map the result through a nonlinear function to obtain a global context vector:

$$\mathbf{g} = f_{\text{global}}(\text{MeanPool}(\mathbf{Z})), \quad (11)$$

where $\mathbf{g} \in \mathbb{R}^d$ summarizes the global context across regions, and f_{global} is a learnable mapping (implemented as a multilayer perceptron (MLP) with an activation function).

For each lead time $h \in \{1, \dots, H\}$, we introduce a learnable lead-time embedding $\mathbf{e}_{\text{lead}}^{(h)} \in \mathbb{R}^{d_e}$ to encode the lead-time identity, where d_e is the lead-time embedding dimension. We then concatenate the local node representation, global context and lead-time embedding (denoted by $\parallel$) and project them to a lead-time-specific query token as

$$\mathbf{q}_i^{(h)} = f_{\text{query}}\left(\left[\mathbf{z}_i \parallel \mathbf{g} \parallel \mathbf{e}_{\text{lead}}^{(h)}\right]\right), \quad (12)$$

where f_{query} is a learnable query projection and $\mathbf{q}_i^{(h)} \in \mathbb{R}^{d_q}$ is the query embedding for region i at lead time h (with dimension d_q). Repeating this procedure for $h = 1, \dots, H$ yields the query sequence $[\mathbf{q}_i^{(1)}, \mathbf{q}_i^{(2)}, \dots, \mathbf{q}_i^{(H)}]$.

Adjacent lead times are not independent: transmission parameters typically evolve smoothly across lead times. We therefore model temporal dependence among the query tokens via self-attention along the lead time dimension:

$$\tilde{\mathbf{Q}}_i = \text{SelfAttn}\left(\left[\mathbf{q}_i^{(1)}, \dots, \mathbf{q}_i^{(H)}\right]\right), \quad (13)$$

where $\tilde{\mathbf{Q}}_i = [\tilde{\mathbf{q}}_i^{(1)}, \dots, \tilde{\mathbf{q}}_i^{(H)}]$ is the temporally contextualized query sequence. In standard dot-product attention, the core operation is given in Equation (4), where $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ are the query, key and value matrices, respectively, and d_k is the key dimension.

However, near-term and longer-term propagation patterns may differ, and directly injecting the full temporal self-attention signal across all lead times may over-couple parameter representations. We therefore introduce a learnable temporal gate that adaptively controls the degree of information fusion among different lead times for each query token:

$$\hat{\mathbf{q}}_i^{(h)} = \text{LN}\left(\mathbf{q}_i^{(h)} + \alpha \cdot \tilde{\mathbf{q}}_i^{(h)}\right), \quad (14)$$

where LN denotes layer normalization, $\alpha = \sigma(w_g)$ is a global scalar gate shared across regions, lead times, and attention heads in the adopted model, $\sigma(\cdot)$ is the Sigmoid function, and w_g is a learnable parameter initialized to -1.0 (giving $\alpha \approx 0.27$ at the start of training). A smaller α keeps the gated query closer to the original lead-time-conditioned query $q_i^{(h)}$, thereby limiting information coupling among lead times. A larger α increases the contribution of temporally contextualized query information $\tilde{q}_i^{(h)}$, enabling stronger temporal-context sharing across future lead times. Thus, the adopted global gate provides a learnable trade-off between preserving lead-time-conditioned query separability and sharing temporal context across lead times. Given the gated lead-time-specific query $\hat{q}_i^{(h)}$, we then use cross-attention to retrieve transmission-relevant information from the historical node representations $\mathbf{Z}$ as

$$p_i^{(h)} = \text{CrossAttn}(\hat{q}_i^{(h)}, \mathbf{Z}, \mathbf{Z}), \quad (15)$$

where $\hat{q}_i^{(h)}$ serves as the query and $\mathbf{Z}$ serves as both keys and values, allowing each region at each horizon to attend to its own and other regions' historical spatiotemporal signals. Consistent with Equation (4), a single-head cross-attention instance can be expressed as

$$\text{CrossAttn}(q, \mathbf{Z}, \mathbf{Z}) = \text{Attn}(q\mathbf{W}_Q, \mathbf{Z}\mathbf{W}_K, \mathbf{Z}\mathbf{W}_V), \quad (16)$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$ and $\mathbf{W}_V$ are learnable projection matrices (with dimensions compatible with d_q and d).

Finally, two output heads generate the latent transmission and recovery parameters as

$$\beta_i^{(h)} = \text{softplus}(f_\beta(p_i^{(h)})), \quad \gamma_i^{(h)} = \text{softplus}(f_\gamma(p_i^{(h)})), \quad (17)$$

where f_β and f_γ are MLP-based mappings for transmission and recovery, respectively, and $\text{softplus}(x) = \log(1 + e^x)$ enforces positivity. Here, $\beta_i^{(h)}$ denotes the latent-space transmission intensity of region i at lead time h , and $\gamma_i^{(h)}$ denotes the corresponding recovery intensity. These dynamically generated parameters directly drive the subsequent latent-space S/I/R propagation module.

3.5. Propagation Prediction Module of Latent Space S/I/R Mechanism

Classical mechanism-guided hybrid models typically rely on observable compartment counts as supervision; however, influenza surveillance streams provide only indirect proxies such as the proportion of influenza-like illness or hospital-admission counts, rendering full population states unobserved [44,45]. To address this structural limitation, EpiSTLLM constructs S/I/R state representations in the neural latent space, as shown in Figure 3, thereby embedding epidemiological propagation structure into multi-horizon forecasting without requiring explicit population-level measurements. Concretely, given the graph-enhanced LLM encoding for region i , $z_i \in \mathbb{R}^d$ (where d is the embedding dimension), three independent projection functions map it to latent susceptible, infectious and removed states with non-negativity enforced to preserve physical plausibility:

$$\mathbf{S}_i^{(0)} = f_S(z_i), \quad (18)$$

$$\mathbf{I}_i^{(0)} = f_I(z_i) + \lambda \cdot \bar{x}_i^{\text{rec}}, \quad (19)$$

$$\mathbf{R}_i^{(0)} = f_R(z_i). \quad (20)$$

Here, f_S , f_I and f_R are learnable MLPs; $\bar{x}_i^{\text{rec}}$ denotes the mean observed disease activity for region i over the most recent rec weeks; and λ is a learnable scaling coefficient. Because

the infectious state is most directly tethered to recent incidence, the anchor term $\bar{x}_i^{\text{rec}}$ stabilizes the mechanistic roll-out by partially constraining the latent initial condition with near-term observations.

Given the latent initialization, the model performs rolling propagation across lead times $h = 1, 2, \dots, H$ by explicitly incorporating inter-regional spread through a row-normalized adjacency matrix $\tilde{\mathbf{A}}$ [46]. In particular, the neighborhood infection pressure acting on region i at lead time h is computed by aggregating the previous-lead-time infectious states of its neighbors:

$$\mathbf{P}_i^{(h)} = \sum_j \tilde{A}_{ij} \cdot \mathbf{I}_j^{(h-1)}, \quad (21)$$

where $\tilde{A}_{ij}$ is the normalized influence weight from region j to region i , and $\mathbf{I}_j^{(h-1)}$ is the latent infectious state of region j at lead time $h - 1$. One convenient normalization is

$$\tilde{\mathbf{A}} = \mathbf{D}^{-1} \mathbf{A}, \quad \mathbf{D}_{ii} = \sum_j A_{ij}, \quad (22)$$

where $\mathbf{A}$ is the (possibly self-looped) adjacency matrix and $\mathbf{D}$ is the corresponding out-degree matrix.

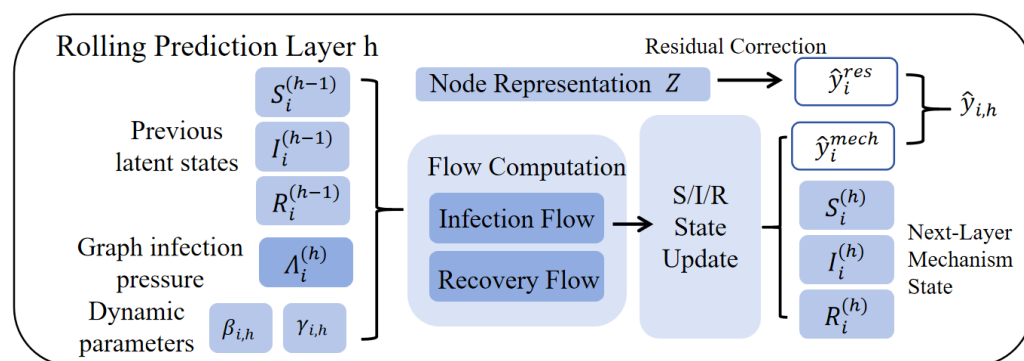


Figure 3. Framework of S/I/R latent space rolling prediction layer.

Conditioned on the infection pressure $\mathbf{P}_i^{(h)}$, the new-infection flow is jointly driven by the current latent susceptible state and neighborhood pressure, passed through a learnable infection-flow map f_{inf} , and modulated by the lead-time-specific transmission parameter $\beta_i^{(h)}$ generated by the temporally gated cross-attention module. The recovery flow is computed from the current infectious state via a learnable map f_{rec} and modulated by the recovery parameter $\gamma_i^{(h)}$:

$$\Delta \mathbf{I}_i^{(h)} = \beta_i^{(h)} \cdot f_{\text{inf}}\left(\left[\mathbf{S}_i^{(h-1)} \parallel \mathbf{P}_i^{(h)}\right]\right), \quad (23)$$

$$\Delta \mathbf{R}_i^{(h)} = \gamma_i^{(h)} \cdot f_{\text{rec}}(\mathbf{I}_i^{(h-1)}), \quad (24)$$

where $\beta_i^{(h)} > 0$ and $\gamma_i^{(h)} > 0$ denote, respectively, the latent transmission and recovery intensities for region i at lead time h , and $\parallel$ denotes concatenation. The latent S/I/R states are then updated by a mass-conserving roll-out as

$$\mathbf{S}_i^{(h)} = \mathbf{S}_i^{(h-1)} - \Delta \mathbf{I}_i^{(h)}, \quad (25)$$

$$\mathbf{I}_i^{(h)} = \mathbf{I}_i^{(h-1)} + \Delta \mathbf{I}_i^{(h)} - \Delta \mathbf{R}_i^{(h)}, \quad (26)$$

$$\mathbf{R}_i^{(h)} = \mathbf{R}_i^{(h-1)} + \Delta \mathbf{R}_i^{(h)}, \quad (27)$$

which induces a continuous mechanistic chain in latent space, providing step-to-step structural constraints for long-horizon forecasting.

Real-world influenza surveillance time series are additionally shaped by reporting delays, measurement noise and unobserved external drivers [14]; a purely mechanistic roll-out may therefore underfit complex observation patterns. We thus adopt a dual-branch design. At each lead time, a mechanistic readout maps the current infectious state and incident infections to an observable-scale prediction as

$$\hat{y}_{i,h}^{\text{mech}} = f_{\text{obs}}\left(\left[\Delta \mathbf{I}_i^{(h)} \parallel \mathbf{I}_i^{(h)}\right]\right), \quad (28)$$

where f_{obs} is a learnable MLP and $\hat{y}_{i,h}^{\text{mech}}$ is the mechanism-derived forecast for region i at lead time h . In parallel, a data-driven residual branch leverages the node representation $\mathbf{z}_i$ to capture patterns not explained by the mechanistic approximation as

$$\hat{y}_i^{\text{res}} = \mathbf{W}_2 \delta(\mathbf{W}_1 \mathbf{z}_i + \mathbf{b}_1) + \mathbf{b}_2, \quad (29)$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are learnable linear maps, $\mathbf{b}_1$ and $\mathbf{b}_2$ are bias vectors, and $\delta(\cdot)$ is a nonlinear activation. The two branches are combined additively:

$$\hat{y}_{i,h} = \hat{y}_{i,h}^{\text{mech}} + \hat{y}_{i,h}^{\text{res}}, \quad (30)$$

and a ReLU constraint enforces the non-negativity of the final prediction:

$$\hat{y}_i^{(h)} = \text{ReLU}\left(f_{\text{obs}}\left(\left[\mathbf{I}_i^{(h)} \parallel \Delta \mathbf{I}_i^{(h)}\right]\right) + f_{\text{res}}(\mathbf{z}_i)\right), \quad (31)$$

where $f_{\text{res}}(\mathbf{z}_i)$ corresponds to the residual mapping in Equation (29). Here, $\hat{y}_i^{(h)} \in \mathbb{R}^C$ denotes the final output (with C output channels, for example a single target when $C = 1$). Together, the mechanistic branch regularizes multi-step forecasts through latent S/I/R dynamics, whereas the residual branch restores expressive flexibility to accommodate complex observation processes.

3.6. Training Objective and Loss Functions

To jointly enforce numerical accuracy and stable latent transmission dynamics in multi-horizon, multi-regional forecasting, we optimize a composite objective comprising a prediction term and mechanism-aware regularizers. Let $\hat{\mathbf{Y}}, \mathbf{Y} \in \mathbb{R}^{B \times N \times H}$ denote the model outputs and observations over a mini-batch of size B , N regions and forecast horizon H , where $\hat{y}_{b,i,h}$ and $y_{b,i,h}$ are the predicted and true targets for batch element b , region i and lead time h . Because nearer-term forecasts are typically more actionable and less uncertain than longer horizons, we weight each lead time with a monotonically decaying schedule w_h and its normalized form $\tilde{w}_h$ as

$$\tilde{w}_h = \frac{w_h}{\frac{1}{H} \sum_{r=1}^H w_r}, \quad w_h = 2 - \frac{h-1}{H-1}, \quad h = 1, \dots, H. \quad (32)$$

The weighted prediction loss $\mathcal{L}_{\text{pred}}$ combines an ℓ_1 term with a weighted mean absolute percentage error (WMAPE) as

$$\mathcal{L}_{\text{pred}} = \frac{1}{BNH} \sum_{b=1}^B \sum_{i=1}^N \sum_{h=1}^H \tilde{w}_h |y_{b,i,h} - \hat{y}_{b,i,h}| + \lambda_w \text{WMAPE}_w(\hat{\mathbf{Y}}, \mathbf{Y}), \quad (33)$$

where

$$\text{WMAPE}_w(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{\sum_{b=1}^B \sum_{i=1}^N \sum_{h=1}^H \tilde{w}_h |y_{b,i,h} - \hat{y}_{b,i,h}|}{\sum_{b=1}^B \sum_{i=1}^N \sum_{h=1}^H \tilde{w}_h |y_{b,i,h}| + \epsilon}. \quad (34)$$

$\epsilon > 0$ prevents division by zero, and λ_w controls the contribution of the percentage-error term.

Beyond output fidelity, we regularize the latent mechanistic states to avoid pathological drift: letting $S_b^{(h)}, I_b^{(h)}, R_b^{(h)} \in \mathbb{R}^N$ denote the susceptible, infectious and removed state vectors at lead time h , we penalize deviations from conservation of total mass relative to the initial state to construct a mass-conservation loss $\mathcal{L}_{\text{mass}}$ as

$$\mathcal{L}_{\text{mass}} = \frac{1}{BH} \sum_{b=1}^B \sum_{h=1}^H \left\| \left(S_b^{(h)} + I_b^{(h)} + R_b^{(h)} \right) - \left(S_b^{(0)} + I_b^{(0)} + R_b^{(0)} \right) \right\|_1. \quad (35)$$

In the current implementation, the rollout quality is mainly ensured by a mass-preserving state-update structure rather than being learned solely from an additional mass-conservation loss. Concretely, at each lead time h , we update the latent states as

$$\begin{aligned} S^{(h)} &= S^{(h-1)} - \Delta I^{(h)}, \\ I^{(h)} &= I^{(h-1)} + \Delta I^{(h)} - \Delta R^{(h)}, \\ R^{(h)} &= R^{(h-1)} + \Delta R^{(h)}. \end{aligned} \quad (36)$$

Therefore, up to numerical errors and rounding, $S^{(h)} + I^{(h)} + R^{(h)}$ remains consistent with the initial total mass. We further conduct SIR mass-conservation diagnostics and mass-loss analysis to verify this behavior in Section 5.7.

Finally, to discourage abrupt fluctuations across adjacent lead times in the dynamically generated transmission and recovery parameters, we impose an ℓ_1 smoothness penalty to form a parameter-smoothness loss $\mathcal{L}_{\text{param}}$ as

$$\mathcal{L}_{\text{param}} = \sum_{h=1}^{H-1} (\|\beta_{h+1} - \beta_h\|_1 + \|\gamma_{h+1} - \gamma_h\|_1), \quad (37)$$

where β_h and γ_h denote the latent (region-wise) transmission and recovery parameter vectors at lead time h .

The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_m \mathcal{L}_{\text{mass}} + \lambda_p \mathcal{L}_{\text{param}}, \quad (38)$$

with λ_m and λ_p balancing mechanism regularization against predictive fit.

4. Experimental Setup

4.1. Datasets

To assess the effectiveness of the proposed model for multi-regional influenza transmission forecasting, we conducted experiments on two U.S. state-level surveillance datasets, i.e., the FluView dataset and the National Healthcare Safety Network (NHSN) influenza hospitalization dataset. The FluView dataset is based on the U.S. Centers for Disease Control (CDC) FluView surveillance data, accessed through the Delphi Epidata FluView endpoint, and uses the weekly population-weighted percentage of influenza-like illness (ILI) as the prediction target across 50 States and the District of Columbia, covering epidemiological week 40 of 2013 to week 40 of 2023 (523 weekly time steps) [47,48]. The hospitalization dataset was obtained from CDC's National Healthcare Safety Network (NHSN) Weekly

Hospital Respiratory Data. It reports weekly counts of laboratory-confirmed influenza hospital admissions at the same geographic resolution, spanning week 5 of 2022 to week 16 of 2026 (220 weekly time steps) [49].

We applied a leakage-free preprocessing pipeline to both datasets before model training. As summarized in Algorithm 1, we first harmonized geographic identifiers and converted dates to epidemiological-week indices, retaining only the 51 state-level nodes and removing invalid or incomplete records; duplicate entries within a (state, week) pair were aggregated by taking the mean for FluView and the sum for NHSN dataset. We then constructed a complete weekly panel and an observation mask, and we imputed missing values using a hierarchical strategy (forward fill when available; otherwise week-wise median across observed regions; otherwise an expanding global median; otherwise zero), with negative NHSN values clipped to zero. To prevent temporal leakage in evaluation across multiple lead times, we formed sliding windows and split them chronologically into training, validation and test sets with an embargo gap of $(H - 1)$ windows between splits. Finally, we estimated normalization statistics (μ, σ) using only training inputs and applied them to all splits, and the evaluation metrics were computed only at target positions corresponding to originally observed values.

4.2. Evaluation Metrics

We quantified predictive performance using three complementary metrics: mean absolute error (MAE), root mean squared error (RMSE) and mean absolute percentage error (MAPE). Let N denote the number of spatial units (states/regions) and H the forecast horizon (maximum number of weeks ahead). For region $i \in \{1, \dots, N\}$ at lead time $h \in \{1, \dots, H\}$, $y_i^{(h)}$ is the ground-truth target value and $\hat{y}_i^{(h)}$ is the corresponding model forecasting result. MAE measures the average absolute deviation as

$$\text{MAE} = \frac{1}{N \cdot H} \sum_{i=1}^N \sum_{h=1}^H \left| \hat{y}_i^{(h)} - y_i^{(h)} \right|, \quad (39)$$

whereas RMSE up-weights larger errors as

$$\text{RMSE} = \sqrt{\frac{1}{N \cdot H} \sum_{i=1}^N \sum_{h=1}^H \left(\hat{y}_i^{(h)} - y_i^{(h)} \right)^2}. \quad (40)$$

To assess relative error and stability across scales, we also report MAPE as

$$\text{MAPE} = \frac{1}{N \cdot H} \sum_{i=1}^N \sum_{h=1}^H \frac{\left| \hat{y}_i^{(h)} - y_i^{(h)} \right|}{\left| y_i^{(h)} \right| + \epsilon}, \quad (41)$$

where ϵ is a small constant added to avoid division by zero. For all three metrics, lower values indicate better predictive accuracy.

4.3. Implementation Details

We used a 24-week input window and forecast horizons of 4, 8 and 12 weeks, i.e., $H \in \{4, 8, 12\}$, in the experiments. Samples were split chronologically into training, validation and test sets with a ratio of 7:2:1. In the latent compartmental propagation module, the hidden-state dimension was set to 16. As shown in Algorithm 2, EpiSTLLM is trained in an end-to-end manner through mini-batch optimization, where graph-enhanced temporal representations and latent S/I/R dynamics are jointly learned. The composite objective combines prediction accuracy, mass conservation, and epidemiological parameter

regularization. Ranger with gradient clipping is adopted for stable optimization, while only LoRA adapters and unfrozen task-specific modules are updated during training. For the language-model backbone, we adopted GPT-2 base (~124 M parameters) with 6 fusion layers; the temporal gate parameter was initialized to -1.0 , and all experiments used a fixed random seed of 6666 to ensure reproducibility. Models were trained with batch size 32, a base learning rate of 5×10^{-4} and 400 epochs. Owing to the sensitivity of LLM adaptation, our model and GPT-augmented baselines were optimized with Ranger, whereas other conventional deep-learning baselines used Adam; weight decay was set to 1×10^{-4} and the gradient-norm clipping threshold to 5.

Algorithm 1: Leakage-free data cleaning and standardization

Require: Raw records R with fields (date, state, value); dataset type $d \in \{\text{FluView}, \text{NHSN}\}$; input length T ; forecast horizon H

Ensure : Leakage-free train/val/test windows $\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{val}}, \mathcal{D}_{\text{test}}$ and normalization statistics (μ, σ) estimated from training inputs only

- 1 Keep only the 51 state-level nodes (50 States + District of Columbia)
- 2 Parse date to epidemiological week index t ; cast value to numeric Remove records with invalid or missing (state, t , value)
- 3 **foreach** (state, t) group **do**
 - 4 **if** $d = \text{FluView}$ **then**
 - 5 value $\leftarrow$ mean(value)
 - 6 **else**
 - 7 value $\leftarrow$ sum(value)
- 8 Construct a complete weekly panel $Y[t, \text{state}]$ over all states and weeks Record an observation mask $M[t, \text{state}]$ (1 if originally observed, else 0)
- 9 **foreach** missing entry $Y[t, i]$ **do**
 - 10 **if** region i has an available historical value before t **then**
 - 11 $Y[t, i] \leftarrow$ forward fill($Y[\leq t, i]$)
 - 12 **else if** other regions have observations at week t **then**
 - 13 $Y[t, i] \leftarrow$ median($\{Y[t, j] \mid M[t, j] = 1\}$)
 - 14 **else if** any historical observations exist up to week t **then**
 - 15 $Y[t, i] \leftarrow$ median($\{Y[\tau, j] \mid \tau \leq t, M[\tau, j] = 1\}$)
 - 16 **else**
 - 17 $Y[t, i] \leftarrow 0$
- 18 **if** $d = \text{NHSN}$ **then**
 - 19 $Y \leftarrow \max(Y, 0)$ // clip negatives
- 20 Construct sliding-window samples for each start week s : $X_s \leftarrow Y[s : s + T]$, $Y_s \leftarrow Y[s + T : s + T + H]$, $M_s^x \leftarrow M[s : s + T]$, $M_s^y \leftarrow M[s + T : s + T + H]$; Split windows chronologically into $\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{val}}, \mathcal{D}_{\text{test}}$ with an embargo gap $\text{gap} = H - 1$ between splits; Estimate normalization statistics from training inputs only: $\mu \leftarrow \text{mean}(X_{\text{train}})$, $\sigma \leftarrow \text{std}(X_{\text{train}})$; Standardize all splits using μ and σ ; During evaluation, compute MAE/RMSE/MAPE only on target positions with $M_s^y = 1$
- return** $\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{val}}, \mathcal{D}_{\text{test}}, \mu, \sigma$

Algorithm 2: Training Procedure of EpiSTLLM

Input : Training set $\mathcal{D}_{\text{train}}$, validation set $\mathcal{D}_{\text{val}}$, forecast horizon H , epochs E , batch size B , learning rate η , weight decay ω , gradient clipping threshold τ , loss weights λ_m and λ_p

Output: Best parameters θ^*

// Initialization

- 1 Initialize model parameters θ ; load the pretrained GPT-2 backbone and freeze its weights; initialize LoRA adapters and other trainable modules; initialize the Ranger optimizer on trainable parameters with (η, ω) $\mathcal{L}_{\text{best}} \leftarrow +\infty$
- 2 **for** epoch = 1, ..., E **do**
- 3 **foreach** mini-batch $(\mathbf{X}, \mathbf{Y}, \mathcal{G}) \sim \mathcal{D}_{\text{train}}$ **do**
- 4 // Data loading and forward propagation
- 4 Load inputs $\mathbf{X}$, targets $\mathbf{Y}$ and graph $\mathcal{G}$
- 4 $\hat{\mathbf{Y}}, \{\beta_h, \gamma_h\}_{h=1}^H, \{\mathbf{S}^{(h)}, \mathbf{I}^{(h)}, \mathbf{R}^{(h)}\}_{h=0}^H \leftarrow \text{EpiSTLLM}(\mathbf{X}, \mathcal{G})$
- 4 // Loss computation
- 5 Compute prediction loss $\mathcal{L}_{\text{pred}}(\hat{\mathbf{Y}}, \mathbf{Y})$ (Equation (33))
- 6 Compute mass-conservation regularizer $\mathcal{L}_{\text{mass}}(\{\mathbf{S}^{(h)}, \mathbf{I}^{(h)}, \mathbf{R}^{(h)}\}_{h=0}^H)$ (Equation (35))
- 7 Compute parameter-smoothness regularizer $\mathcal{L}_{\text{param}}(\{\beta_h, \gamma_h\}_{h=1}^H)$ (Equation (37))
- 8 Compute $\mathcal{L} \leftarrow \mathcal{L}_{\text{pred}} + \lambda_m \mathcal{L}_{\text{mass}} + \lambda_p \mathcal{L}_{\text{param}}$ (Equation (38))
- 4 // Backpropagation and parameter update
- 9 Backpropagate $\nabla_{\theta} \mathcal{L}$ (updating only LoRA parameters and other non-frozen modules) Clip gradients to satisfy $\|\nabla\| \leq \tau$ Ranger optimizer step
- 4 // Validation and checkpointing
- 10 Evaluate validation loss $\mathcal{L}_{\text{val}}$ on $\mathcal{D}_{\text{val}}$ **if** $\mathcal{L}_{\text{val}} < \mathcal{L}_{\text{best}}$ **then**
- 11 $\mathcal{L}_{\text{best}} \leftarrow \mathcal{L}_{\text{val}}$; save checkpoint $\theta^* \leftarrow \theta$
- 12 **return** θ^*

For the proposed mechanistic multi-step model, we additionally reweighted the MAE term across lead times using a linearly decaying schedule from 2.0 to 1.0 as

$$w_h = 2.0 - \frac{h-1}{H-1}, \quad h = 1, \dots, H, \quad (42)$$

where h indexes the forecast lead time and H is the horizon length. We further introduced two physics-inspired regularizers: a mass-conservation loss $\mathcal{L}_{\text{mass}}$ with $\lambda_m = 0.01$ and a parameter-smoothness loss $\mathcal{L}_{\text{param}}$ with $\lambda_p = 0.01$.

For all models, we retained the checkpoint with the lowest validation loss for final evaluation on the test set. All models were implemented in PyTorch 2.3.0 and trained and evaluated on a single NVIDIA RTX 4090 GPU (CUDA). Unless otherwise stated, baseline hyperparameters followed the recommendations in the original papers.

4.4. Baseline Models

To comprehensively evaluate EpiSTLLM, we benchmark against 14 baselines spanning complementary modeling paradigms, which were arranged from simple heuristics to mechanism-guided deep models. Classical statistical baselines include persistence (using the most recent observation as the forecasting result for future lead times), autoregressive

(AR) forecasting [50], and vector autoregression (VAR) [51], which together expose the limits of purely linear assumptions when modeling the strongly nonlinear and heterogeneous dynamics of influenza spread.

Sequence neural networks (LSTM [52] and GRU [53]) provide competitive temporal baselines by learning nonlinear dependencies from historical observations, yet they do not explicitly represent inter-regional propagation.

Spatiotemporal graph neural networks, including STGCN [54], DCRNN [55], Cola-GNN [1] and EpiGNN [26], introduce graph structure, attention, or diffusion-style operators to capture cross-region coupling and to quantify the benefit of explicit spatial modeling in multi-regional forecasting.

We further consider LLM-enhanced graph variants: GCNGPT [56] couples a graph convolutional encoder with a pretrained GPT-2 backbone, whereas GATGPT [57] replaces GCN with graph attention to test whether straightforward GNN-LLM fusion can bridge spatial modeling and long-range sequence representation.

Finally, we include classical compartmental mechanism models (SIR [58] and SEIR [59]) that describe macroscopic $S \rightarrow I \rightarrow R$ and $S \rightarrow E \rightarrow I \rightarrow R$ progressions, as well as a representative hybrid approach, CausalGNN [45], which embeds an SIRD-based causal mechanism module into a dynamic graph-attention network. This suite enables a comprehensive comparison across purely data-driven, explicitly spatial, mechanistic, and mechanism-guided learning paradigms.

5. Results and Discussion

5.1. Comparison of Influenza Forecasting Performance

We compared our approach with a diverse set of baselines on two U.S. state-level influenza-related datasets, reporting multi-step forecasting performance averaged over all lead times within each forecast horizon. We deliberately evaluate on both FluView ILI and NHSN influenza hospitalizations because they pose distinct modeling challenges: the FluView dataset more directly reflects ILI activity and underlying transmission trends, where mechanistic and graph-based models remain highly competitive, whereas NHSN hospitalizations exhibit larger magnitudes and are more susceptible to observation noise, which can limit purely mechanistic fits and accentuate the value of strong contextual sequence representations. As summarized in Tables 1 and 2, EpiSTLLM delivers the best or consistently competitive forecasting performance across most horizons, with the largest gains typically observed in terms of MAE and RMSE, indicating more reliable control of absolute error and peak deviation across heterogeneous surveillance signals.

For ILI forecasting on the FluView dataset, VAR remains the weakest baseline across horizons, indicating that linear multivariate dynamics alone cannot capture the nonlinear and heterogeneous state-level ILI patterns. Persistence and AR models are competitive at short horizon ($H = 4$) but degrade noticeably as the horizon increases. Pure sequence models (LSTM/GRU) generally improve over statistical models by learning nonlinear temporal dependencies, yet their lack of explicit spatial propagation limits robustness at longer horizons. Graph-based approaches reduce error by modeling inter-regional coupling, but their performance varies; notably, EpiSTLLM achieves the best MAE/RMSE performance at all horizons (e.g., MAE/RMSE as 0.5621/0.9858 at $H = 4$ and 0.9068/1.4321 at $H = 12$), while Cola-GNN attains the lowest MAPE (0.3891) at $H = 12$.

For influenza hospitalization forecasting on the NHSN dataset, mechanism-guided learning is particularly helpful: CausalGNN achieves the lowest MAE/MAPE values (63.302/0.5975) at $H = 4$, while EpiSTLLM delivers the best RMSE (174.065) at $H = 4$ and consistently outperforms all baselines at longer horizons ($H = 8$ and $H = 12$), reaching MAE/RMSE/MAPE values as 78.681/217.971/0.6715 and 85.688/244.514/0.6663,

respectively. In contrast, the standalone SIR/SEIR model exhibits substantially larger error (e.g., $MAE > 135$ at $H = 12$), suggesting that mechanistic constraints alone are insufficient for the higher-variance hospitalization information without stronger data-driven representations.

The above observed advantage is attributable to the complementary roles of three design contributions: graph-enhanced LLM representations that strengthen contextual modeling while encoding inter-regional association; a temporal-gated cross-attention mechanism that adapts propagation behavior to the forecast lead time; and a latent SIR-inspired propagation component that regularizes medium- to long-range dynamics.

Table 1. Forecasting results on the FluView dataset. H denotes the forecasting horizon (weeks). Lower values indicate better performance in terms of MAE, RMSE and MAPE. Best performance is highlighted in **bold**.

Model	MAE	$H = 4$		MAE	$H = 8$		MAE	$H = 12$	
		RMSE	MAPE		RMSE	MAPE		RMSE	MAPE
Persistence	0.6251	1.1069	0.2899	0.9003	1.5247	0.4334	1.1009	1.7958	0.5453
AR	0.6740	1.1452	0.3097	0.8823	1.4525	0.4126	1.0305	1.6774	0.4948
VAR	1.0606	1.5582	0.6321	1.1783	1.7380	0.7029	1.3041	1.9365	0.7381
LSTM	0.6213	1.0514	0.2976	0.9219	1.4993	0.4621	1.0395	1.7426	0.4874
GRU	0.6554	1.1149	0.3169	0.8586	1.4262	0.4185	1.0142	1.6805	0.4752
STGCN	0.8211	1.3191	0.4138	1.0323	1.6126	0.4869	1.5405	2.2780	0.6645
DCRNN	0.6355	1.0883	0.2644	0.9212	1.5249	0.3874	0.9956	1.6631	0.3971
Cola-GNN	0.6882	1.2015	0.2763	0.8653	1.4416	0.3986	1.0211	1.7273	0.3891
EpiGNN	0.7510	1.2539	0.3094	0.8673	1.4365	0.3560	0.9755	1.5654	0.4280
GCNGPT	1.0531	1.5995	0.4797	1.1899	1.9709	0.5531	1.2316	2.0382	0.5508
GATGPT	1.0640	1.6283	0.4913	1.1971	1.9795	0.5528	1.2195	2.0194	0.5469
SIR	0.6274	1.0763	0.2883	0.8421	1.4245	0.3716	0.9731	1.6426	0.4167
SEIR	0.6308	1.0782	0.2891	0.8563	1.4287	0.3853	0.9868	1.6458	0.4280
CausalGNN	0.6559	1.1812	0.2625	0.8180	1.3239	0.4074	0.9517	1.5112	0.4984
EpiSTLLM (ours)	0.5621	0.9858	0.2575	0.7483	1.2912	0.3250	0.9068	1.4321	0.4409

Table 2. Forecasting results on the NHSN dataset. H denotes the forecasting horizon (weeks). Lower values indicate better performance in terms of MAE, RMSE and MAPE. Best performance is highlighted in **bold**.

Model	MAE	$H = 4$		MAE	$H = 8$		MAE	$H = 12$	
		RMSE	MAPE		RMSE	MAPE		RMSE	MAPE
Persistence	95.703	254.719	0.7979	134.319	340.206	1.1037	158.009	389.009	1.3071
AR	93.290	250.335	0.9610	116.520	306.413	1.3030	134.038	340.787	1.5817
VAR	140.982	298.059	6.8024	156.621	333.462	7.7612	168.974	354.894	9.7624
LSTM	77.558	208.474	0.7766	105.942	291.820	0.9664	121.956	328.903	0.9859
GRU	78.304	217.133	0.8126	106.825	294.212	0.8523	115.436	309.674	1.0169
STGCN	98.265	248.985	1.9861	102.846	297.221	0.9834	112.945	272.003	2.8774
DCRNN	64.562	185.246	0.8853	109.362	316.396	1.0277	116.942	323.190	1.2400
Cola-GNN	71.119	189.787	0.6989	101.367	267.350	1.9149	123.339	331.091	0.8346
EpiGNN	86.185	240.877	2.6175	102.572	299.122	1.6922	121.365	341.340	1.4294
GCNGPT	88.948	253.866	1.1701	89.365	250.831	1.0708	97.523	276.183	0.9907
GATGPT	89.738	254.636	1.1929	91.265	257.727	1.0019	101.109	284.541	1.0464
SIR	95.101	254.532	0.7509	118.691	312.086	0.8345	136.569	352.162	0.9062
SEIR	94.481	256.093	0.6752	120.859	319.643	0.7542	135.685	350.752	0.8085
CausalGNN	63.302	193.593	0.5975	106.095	314.562	0.7941	107.487	305.184	1.0104
EpiSTLLM (ours)	64.229	174.065	0.6615	78.681	217.971	0.6715	85.688	244.514	0.6663

In addition, to examine trend-following behavior, we visualized NHSN forecasts at $H = 12$ for different lead times (Lead $h \in \{1, 6, 12\}$), as shown in Figure 4. At Lead = 1, most models are able to follow the short-term outbreak trajectory, whereas their fidelity differs substantially in capturing peak timing, amplitude, and local fluctuations. In contrast,

EpiSTLLM closely follows the overall rise–peak–decline pattern in both the states of Montana (MT) and Wyoming (WY) while avoiding some of the excessive oscillations or delayed responses observed in several competing baselines. At Lead = 6, the forecasting task becomes more challenging and baseline degradation is more evident particularly around rapidly increasing or post-peak periods; EpiSTLLM still provides a comparatively stable trajectory, preserving the dominant direction of change and maintaining a reasonable balance between global epidemic trends and local temporal variability. At Lead = 12, although errors increase for all models and precise peak reconstruction remains difficult, EpiSTLLM continues to capture the main long-horizon epidemic evolution with less severe drift than many baselines, demonstrating improved robustness in long-range spatiotemporal forecasting. These qualitative results are consistent with the complementary contributions of graph-enhanced LLM representations, horizon-adaptive parameter generation based on temporal-gated cross-attention, and latent S/I/R-constrained propagation.

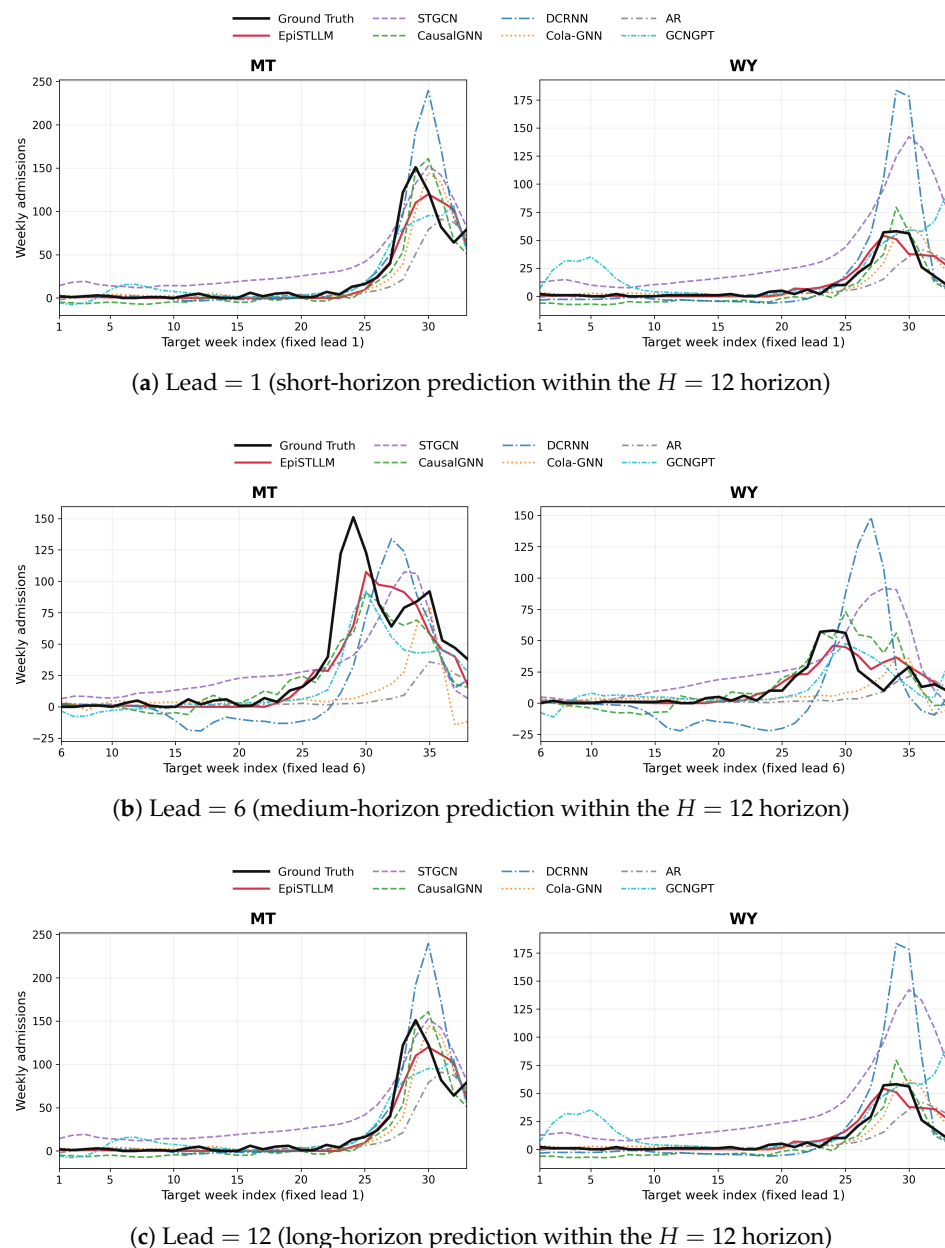


Figure 4. Comparison of model forecasts on the NHSN dataset at $H = 12$ for different lead times in the states of Montana (MT) and Wyoming (WY). The x-axis of each plot starts from the Lead value, indicating the forecasting (a) 1 week ahead, (b) 6 weeks ahead, and (c) 12 weeks ahead.

Building on the visualization at forecast horizon $H = 12$ (Figure 4), we further focus on the most challenging setting, Lead = 12, to assess stability and trend preservation under long forecasting horizon (Figure 5). Under the data leakage-free evaluation protocol, the forecasting difficulty increases substantially at this lead time, and all models exhibit varying degrees of peak-amplitude error, temporal misalignment, or local fluctuation distortion. Several temporal or graph-based baselines can capture parts of the epidemic evolution, but they often suffer from over-smoothed trajectories, delayed responses, or attenuated peak magnitudes, especially during rapid epidemic ascent and post-peak decay. EpiSTLLM does not perfectly reconstruct all peaks and still underestimates or overestimates the magnitude in some states; however, it generally provides a more stable and competitive long-horizon trajectory across multiple regions, preserving the main rise–peak–decline pattern and the direction of change more consistently than most baselines in challenging cases such as the cases in California (CA), Tennessee (TN), New Mexico (NM) and Montana (MT). These qualitative observations are consistent with the horizon-wise error profile and the graph-structure sensitivity analysis, suggesting that the benefit of EpiSTLLM should be interpreted as improved robustness under appropriate spatial priors rather than uniform dominance across all regions. The results further indicate that graph-enhanced LLM representations can help encode inter-regional coupling and long-range temporal context, while temporal-gated cross-attention supports horizon-adaptive parameterization and latent S/I/R-constrained propagation provides an additional regularizing effect for long-term epidemic dynamics.

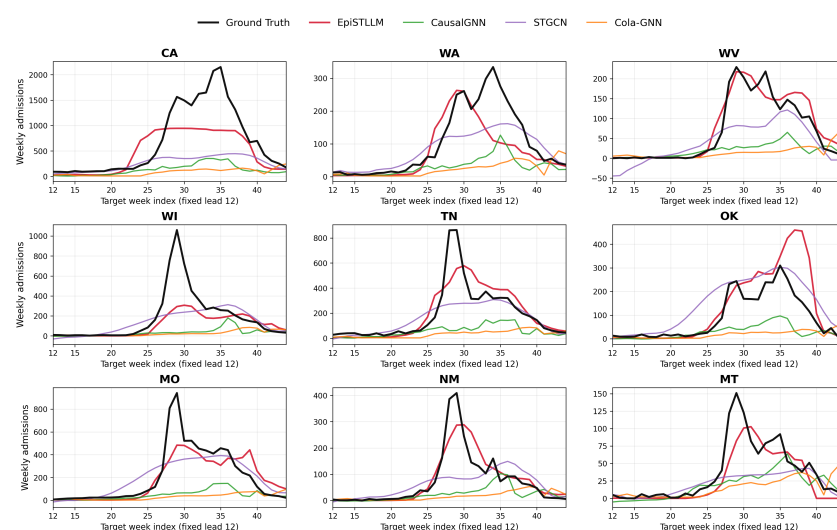


Figure 5. Comparison of long-horizon (Lead = 12) NHSN influenza forecasting across multiple U.S. states. (The state abbreviations in the subplots are as follows: CA denotes California, WA denotes Washington, WV denotes West Virginia, WI denotes Wisconsin, TN denotes Tennessee, OK denotes Oklahoma, MO denotes Missouri, NM denotes New Mexico, and MT denotes Montana.) The x-axis of each plot starts from the Lead value, indicating forecasting 12 weeks ahead.

5.2. Performance Analysis Across Different Forecasting Horizons

As shown in Figure 6, the per-horizon MAE on the leakage-free test sets generally increases as the forecasting horizon H extends from 1 to 12 weeks ($H \in \{1, \dots, 12\}$), reflecting the accumulated uncertainty of multi-horizon epidemic forecasting. On the FluView dataset, EpiSTLLM does not always achieve the lowest MAE at shorter horizons, but its error growth becomes more stable as the horizon increases, and it remains among the best-performing models at medium and long horizons. This indicates that EpiSTLLM is less prone to rapid horizon-wise error accumulation than several temporal, graph-based, and mechanism-guided baselines. On the NHSN dataset, this long-horizon advantage

is more pronounced: after the initial horizons, EpiSTLLM maintains a relatively flat and consistently low MAE curve, whereas the competing methods generally show continued error growth as the prediction horizon extends. In particular, the SIR baseline exhibits increasing errors at longer horizons, suggesting that a fixed mechanistic structure alone may not be sufficient to maintain stable long-term forecasts under non-stationary admission dynamics. In contrast, EpiSTLLM does not rely solely on a fixed transition rule; instead, it combines temporal-gated horizon interactions and latent S/I/R-constrained propagation to dynamically generate propagation parameters for different forecast lead times. This more flexible mechanism-aware design may explain the flatter MAE trajectory observed at longer horizons, especially on the NHSN dataset. Thus, the main advantage of the proposed EpiSTLLM lies in providing stronger robustness against long-horizon error accumulation, especially when forecasting uncertainty increases over multiple future weeks.

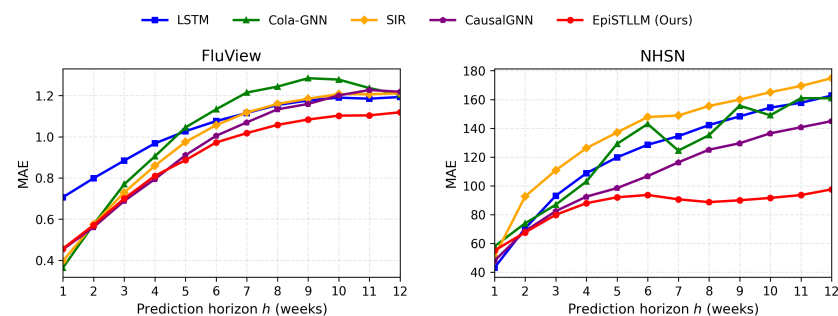


Figure 6. Comparison of MAE across different forecasting horizons on FluView and NHSN datasets.

5.3. Model Ablation Analysis

To quantify the contribution of each component in EpiSTLLM, we conducted an extended ablation study with nine variants: (1) Random init GPT-2, which is the same GPT-2 architecture trained from random initialization to disentangle architectural effects from pretrained initialization; (2) Vanilla Transformer, replacing the GPT-2 backbone with a randomly initialized Transformer encoder to test whether gains arise from generic self-attention rather than a pretrained LLM backbone; (3) Without LLM, removing the GPT-2/LLM-style representation branch while retaining the graph information and the mechanistic propagation pathway; (4) Frozen GPT-2 (no LoRA), keeping the GPT-2 pretrained weights but freezing the backbone and disabling LoRA, training only task-specific layers; (5) No LLM graph, removing graph bias in the LLM representation stage by replacing the adjacency with the identity matrix while preserving the true regional graph in the latent S/I/R mechanism; (6) No S/I/R graph, preserving graph bias in the LLM representation stage but replacing the propagation graph in the latent S/I/R mechanism with the identity matrix; (7) Identity, replacing the graphs in both the LLM representation branch and the latent S/I/R propagation branch with the identity matrix to remove cross-region message passing; (8) Without S/I/R mechanism, removing the latent-space S/I/R mechanistic module while retaining the data-driven prediction branch and the LLM-style representations; and (9) MLP parameter generator (MLP Param Gen), replacing the temporal-gated cross-attention parameter generator with an MLP to assess whether horizon-adaptive parameter generation improves over a static mapping.

As shown in Tables 3 and 4, removing the LLM branch (i.e., Without LLM) significantly degrades NHSN long-horizon performance (e.g., MAPE increases from 0.6663 to 1.0273 at $H = 12$) and substantially worsens FluView errors. Pretraining and lightweight adaptation are also important: Random init GPT-2 is strongest at short-horizon FluView forecasting (MAE/RMSE/MAPE as 0.5563/0.9708/0.2486 at $H = 4$) but degrades sharply at longer horizons, whereas the full EpiSTLLM model achieves the best MAPE (0.3250) at $H = 8$ on

FluView and the best overall performance (MAE/RMSE as 78.681/217.971) at $H = 8$ on NHSN. Graph usage shows horizon- and module-specific effects: removing graph coupling in the mechanism layer (i.e., No S/I/R graph) yields the best MAE/RMSE (82.776/233.953) performance at $H = 12$ on NHSN, while disabling graph-biased attention in the LLM (i.e., No LLM graph) achieves the best RMSE (1.3979) performance at $H = 12$ on FluView; however, both variants sacrifice accuracy at other horizons, suggesting that jointly using graph cues in representation learning and mechanistic propagation is generally beneficial but not uniformly optimal across datasets. The Identity variant further confirms the value of cross-region coupling: when both graphs are replaced by the identity matrix, performance becomes less stable at longer horizons (e.g., RMSE increases to 1.4447 at $H = 12$ on FluView), even though it can remain competitive at short horizon on NHSN (MAE/RMSE as 63.678/173.484 at $H = 4$). Finally, replacing the temporal-gated cross-attention with an MLP parameter generator generally weakens performance (e.g., MAE/RMSE increases from 78.681/217.971 to 81.264/222.994 at $H = 8$ on NHSN), indicating that lead-time-adaptive parameter generation contributes to robust forecasting across multiple lead times beyond a static mapping; removing the mechanism module (i.e., Without S/I/R) also consistently harms long-horizon forecasting (e.g., MAE rises from 85.688 to 91.424 at $H = 12$ on NHSN), supporting the role of latent epidemiological constraints as structural regularization against error accumulation.

Table 3. Ablation study on the FluView dataset. Each row removes or replaces a single component of the full EpiSTLLM model. Lower values indicate better performance; best performance is highlighted in **bold**.

Variant	$H = 4$			$H = 8$			$H = 12$		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
Random init GPT-2	0.5563	0.9708	0.2486	0.8007	1.3115	0.3873	1.0187	1.5957	0.4757
Vanilla Transformer	0.7784	1.1385	0.3883	0.9256	1.5628	0.3803	1.2006	1.8057	0.5465
w/o LLM	0.6841	1.1527	0.3139	0.9400	1.5288	0.4358	1.0565	1.6945	0.4976
Frozen GPT-2 (no LoRA)	0.5921	1.0105	0.2808	0.7827	1.2815	0.3697	0.8657	1.4484	0.3967
No LLM graph	0.5675	1.0056	0.2589	0.7780	1.3210	0.3373	0.8958	1.3979	0.4588
No S/I/R graph	0.5829	1.0106	0.2633	0.7813	1.3270	0.3348	0.8730	1.4166	0.4144
Identity	0.5696	1.0040	0.2592	0.7407	1.2678	0.3324	0.9243	1.4447	0.4484
w/o S/I/R	0.5912	1.0279	0.2729	0.7349	1.2526	0.3403	0.9396	1.5017	0.4586
MLP Param Gen	0.6093	1.0155	0.2758	0.7669	1.2708	0.3617	0.8754	1.4313	0.4184
Full (EpiSTLLM)	0.5621	0.9858	0.2575	0.7483	1.2912	0.3250	0.9068	1.4321	0.4409

Table 4. Ablation study on the NHSN dataset. Each row removes or replaces a single component of the full EpiSTLLM model. Lower values indicate better performance; best performance is highlighted in **bold**.

Variant	$H = 4$			$H = 8$			$H = 12$		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
Random init GPT-2	67.878	200.794	0.7680	87.078	257.464	0.7268	96.783	276.197	0.7520
Vanilla Transformer	65.021	189.352	0.6550	80.132	233.011	0.6524	86.214	254.745	0.7063
w/o LLM	74.602	213.935	0.6269	100.145	286.957	0.7903	110.783	307.802	1.0273
Frozen GPT-2 (no LoRA)	64.996	186.065	0.6273	82.956	240.046	0.7390	83.899	236.108	0.7776
No LLM graph	67.395	184.756	0.6698	79.185	222.151	0.7090	86.391	245.234	0.6717
No S/I/R graph	66.225	179.566	0.6661	79.800	218.380	0.7483	82.776	233.953	0.6758
Identity	63.678	173.484	0.6172	78.765	218.556	0.7237	87.781	248.549	0.7013
w/o S/I/R	64.177	174.454	0.6493	83.339	231.131	0.7611	91.424	252.240	0.7772
MLP Param Gen	65.941	186.598	0.6334	81.264	222.994	0.7197	89.698	237.007	0.8187
Full (EpiSTLLM)	64.229	174.065	0.6615	78.681	217.971	0.6715	85.688	244.514	0.6663

5.4. Performance Analysis with Different Graph Structures

We adopted a state-border adjacency graph (Border) [60] in the proposed EpiSTLLM model, which is a purely geographic adjacency graph constructed according to spatial contiguity. To address the concern that a purely geographic adjacency graph may not adequately capture real transmission pathways, we replaced the Border graph with four alternative graph structures to assess how different spatial priors affect the predictive performance. The compared four graph structures include Identity, Distance k NN [60] ($k = 4$), Correlation top- k [61] ($k = 4$), and Gravity top- k [62] ($k = 4$). Here, Identity is a no-graph baseline that retains only self-connections and therefore removes explicit inter-regional message passing; Distance k NN ($k = 4$) connects each state to its four geographically nearest states; Correlation top- k ($k = 4$) connects each state to the four states with the highest historical time-series correlation computed on the training split, thereby avoiding leakage from validation or test data; and Gravity top- k ($k = 4$) constructs a gravity proxy based on population size and geographic distance and selects, for each state, the four neighbors with the highest gravity scores.

As presented in Tables 5 and 6, the results indicate that performance differences across graph structures are generally limited, and no single graph is consistently superior across the FluView and NHSN datasets, across forecast horizons, and across evaluation metrics, suggesting that the model's predictive capability does not hinge on any particular graph structure. For example, on the FluView dataset, Correlation top- k ($k = 4$) is slightly better in MAE at $H = 4$, whereas Gravity top- k ($k = 4$) performs better in MAE and RMSE at $H = 12$; on the NHSN dataset, Border performs best in MAE and MAPE metrics at $H = 8$ and $H = 12$. Notably, Identity yields results close to those obtained with explicit graphs in several settings, and it is even marginally better than the Border graph on some metrics, but it performs much worse than the Border graph at long-horizon $H = 12$. In addition, the complex graphs constructed from correlation or gravity proxies provide improvements in some longer-horizon settings but do not consistently outperform the simple Border graph. Meanwhile, the complex graph constructions requires additional data sources and extra preprocessing steps, whereas the Border graph only depends on readily available state-level geographic contiguity and is therefore easier to reproduce, and it is also less prone to dataset-specific noise. Therefore, we adopted the Border graph in the proposed EpiSTLLM model, as it is a sparse, low-data-demand, yet interpretable approximation to spatial coupling.

Table 5. Performance analysis with different graph structures on FluView dataset. Lower values indicate better performance in terms of MAE, RMSE and MAPE. Best performance in each column is highlighted in **bold**.

Graph Structure	$H = 4$			$H = 8$			$H = 12$		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
Border	0.5621	0.9858	0.2575	0.7483	1.2912	0.3250	0.9068	1.4321	0.4409
Identity	0.5696	1.0040	0.2592	0.7407	1.2678	0.3324	0.9243	1.4447	0.4484
Distance k NN ($k = 4$)	0.5679	1.0019	0.2574	0.7800	1.3137	0.3421	0.8741	1.4373	0.4043
Correlation top- k ($k = 4$)	0.5608	1.0010	0.2558	0.7851	1.3242	0.3421	0.8900	1.4342	0.4342
Gravity top- k ($k = 4$)	0.5828	1.0005	0.2673	0.7629	1.3018	0.3324	0.8688	1.4201	0.4123

Table 6. Performance analysis with different graph structures on NHSN dataset. Lower values indicate better performance in terms of MAE, RMSE and MAPE. Best performance in each column is highlighted in **bold**.

Graph Structure	$H = 4$			$H = 8$			$H = 12$		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
Border	64.229	174.065	0.6615	78.681	217.971	0.6715	85.688	244.514	0.6663
Identity	63.678	173.484	0.6172	78.765	218.556	0.7237	87.781	248.549	0.7013
Distance k NN ($k = 4$)	63.405	175.107	0.6520	79.306	215.296	0.7435	87.258	243.261	0.6923
Correlation top- k ($k = 4$)	65.791	177.291	0.7072	80.734	224.055	0.7005	86.349	243.590	0.6909
Gravity top- k ($k = 4$)	66.371	185.293	0.6499	80.359	225.529	0.7396	91.092	253.566	0.7007

5.5. Analysis of the Temporal-Gated Cross-Attention Parameter Generation Module

The temporal-gated cross-attention parameter generator is a key design in EpiSTLLM, dynamically producing lead-specific latent propagation parameters to accommodate horizon-dependent changes in transmission dynamics. We compare four parameter generation strategies in this experiment: an MLP parameter generator, a cross-attention generator that uses lead-time tokens to query historical node representations, a temporal cross-attention (TCA) variant that additionally applies self-attention among future lead-time tokens, and the proposed temporal cross-attention with a learnable gate initialized at -1.0 weighting (TCA + Gate(-1.0)). As shown in Tables 7 and 8, attention-based generators are generally more robust than a plain MLP generator, especially on the NHSN dataset where the MLP generator only achieves the best performance in terms of RMSE metric at $H = 12$. TCA + Gate(-1.0) achieves the best MAE/RMSE (64.229/174.065) performance at $H = 4$, the best RMSE/MAPE (217.971/0.6715) performance at $H = 8$, and the best MAE (85.688) performance at $H = 12$. On the FluView dataset, TCA + Gate(-1.0) yields the lowest MAE/RMSE/MAPE (0.5621/0.9858/0.2575) values at $H = 4$, and the lowest MAPE (0.3250) value at $H = 8$, while the MLP generator achieves the best scores (MAE/RMSE/MAPE as 0.8754/1.4313/0.4184) at $H = 12$, suggesting that the optimal lead-adaptive parameterization can vary with dataset noise characteristics and forecast range. As we aim to design a forecasting model which basically performs well across all horizons, although the MLP generator exhibits some advantage at certain horizons, the TCA + Gate(-1.0) generator is more suitable for the proposed model.

Table 7. Effect of different parameter generators on the FluView dataset. Lower values indicate better performance; best performance is highlighted in **bold**.

Generator	$H = 4$			$H = 8$			$H = 12$		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
MLP Param Gen	0.6093	1.0155	0.2758	0.7669	1.2708	0.3617	0.8754	1.4313	0.4184
Cross-Attention	0.5855	0.9880	0.2759	0.8187	1.3252	0.3809	0.9542	1.5324	0.4544
TCA	0.5669	0.9957	0.2586	0.7459	1.2824	0.3274	0.9134	1.4661	0.4216
TCA + Gate(-1.0)	0.5621	0.9858	0.2575	0.7483	1.2912	0.3250	0.9068	1.4321	0.4409

Table 8. Effect of different parameter generators on the NHSN dataset. Lower values indicate better performance; best performance is highlighted in **bold**.

Generator	$H = 4$			$H = 8$			$H = 12$		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
MLP Param Gen	65.941	186.598	0.6334	81.264	222.994	0.7197	89.698	237.007	0.8187
Cross-Attention	64.577	177.363	0.6675	81.268	225.820	0.7179	88.199	240.738	0.7709
TCA	64.983	177.972	0.6289	78.339	219.013	0.6970	87.320	248.286	0.6640
TCA + Gate(-1.0)	64.229	174.065	0.6615	78.681	217.971	0.6715	85.688	244.514	0.6663

To further examine how gate granularity and its initialization affect predictive performance, we compared five configurations: TCA without gating, TCA + a global scalar gate with learnable parameter w_g initialized to -1.0 (i.e., TCA + Gate(-1.0)) or -2.0 , TCA + a horizon-specific gate, and TCA + a head-specific gate. The global scalar gate uses a single shared learnable parameter w_g (plus its associated bias, totally two parameters) across all horizons and all attention heads. The horizon-specific gate assigns an independent scalar gate $w_g^{(h)}$ to each of the forecast lead times so that the model can apply a different blending ratio between the original query and the temporally attended query at each horizon. Each gate consists of a weight and a bias, yielding $2H$ extra parameters in total, where H is the forecast horizon. The head-specific gate assigns an independent scalar gate $w_g^{(m)}$ to each of the M attention heads, enabling different heads to specialize in different levels of temporal coupling. This similarly introduces $2M$ extra parameters.

The results are summarized in Table 9. On the FluView dataset, at $H = 4$, the head-specific gate leads in MAE (0.5615) and RMSE (0.9772), but it lags behind the global scalar gate (-1.0) in MAPE (0.2588 vs. 0.2575); the horizon-specific gate likewise improves RMSE (0.9812) but underperforms in both MAE and MAPE relative to the global scalar gate. At $H = 8$, the head-specific gate achieves the best MAE (0.7449) and MAPE (0.3232), yet the global scalar gate (-1.0) obtains a better RMSE (1.2912 vs. 1.3035) and a competitive MAE. At $H = 12$, the horizon-specific gate leads in MAE (0.8930) and RMSE (1.4181), while the head-specific gate degrades on all three metrics relative to the global scalar gate. On the NHSN dataset, at $H = 4$, the head-specific gate achieves the best MAE (64.044) but produces the worst MAPE (0.6717) among all gated variants; the horizon-specific gate achieves the best MAPE (0.6130) but the worst MAE (66.400) and RMSE (182.458). At $H = 8$, neither fine-grained variant improves over the global scalar gate on any of the three metrics: the global scalar gate (-1.0) achieves the best RMSE (217.971) and MAPE (0.6715), and the best MAE belongs to TCA without gating (78.339) rather than to any fine-grained gate. At $H = 12$, the head-specific gate leads in MAE (84.790) and RMSE (241.086), while the horizon-specific gate leads in MAPE (0.6420), but no single fine-grained variant simultaneously dominates on all three metrics.

Therefore, our choice of the global scalar gate is based primarily on the following three considerations. First, from a parameter-efficiency perspective, achieving comparable or better overall performance with only two parameters against variants that use up to 24 parameters (horizon-specific gate at $H = 12$, or head-specific gate with $M = 12$ heads) indicates that the horizon- and head-level differences in transmission dynamics do not require independent gating to be captured effectively; and the global mechanism appears sufficient to implicitly encode this variation. Second, the primary concern is not which variant wins on a single metric in isolation but which one maintains stable three-metric performance across all evaluation conditions. The inconsistency of fine-grained gates, such as the horizon-specific gate degrading on both MAE and RMSE at $H = 8$ on NHSN despite its best MAPE, confirms that their added complexity introduces instability rather than reliable improvement. Third, following the principle of parsimony, when two designs achieve comparable overall performance, the simpler one should be preferred; pursuing marginal gains on individual metrics at the cost of cross-setting consistency is not a sound basis for model design.

Table 9. Effect of different temporal gate designs and initializations. Results are reported on the FluView and NHSN datasets. H denotes the forecast horizon, and M denotes the number of attention heads. Lower values indicate better performance; best performance is highlighted in **bold**.

Dataset	H Value	Models	Extra Gate Params	MAE	RMSE	MAPE
FluView	$H = 4$	TCA without gate	0	0.5669	0.9957	0.2586
		TCA + global scalar gate (−1.0)	2	0.5621	0.9858	0.2575
		TCA + global scalar gate (−2.0)	2	0.5687	0.9940	0.2603
		TCA + horizon-specific gate	$2H$	0.5629	0.9812	0.2586
		TCA + head-specific gate	$2M$	0.5615	0.9772	0.2588
	$H = 8$	TCA without gate	0	0.7459	1.2824	0.3274
		TCA + global scalar gate (−1.0)	2	0.7483	1.2912	0.3250
		TCA + global scalar gate (−2.0)	2	0.7486	1.2913	0.3267
		TCA + horizon-specific gate	$2H$	0.7832	1.3452	0.3359
		TCA + head-specific gate	$2M$	0.7449	1.3035	0.3232
	$H = 12$	TCA without gate	0	0.9134	1.4661	0.4216
		TCA + global scalar gate (−1.0)	2	0.9068	1.4321	0.4409
		TCA + global scalar gate (−2.0)	2	0.9376	1.4308	0.4838
		TCA + horizon-specific gate	$2H$	0.8930	1.4181	0.4370
		TCA + head-specific gate	$2M$	0.9305	1.4586	0.4467
NHSN	$H = 4$	TCA without gate	0	64.983	177.972	0.6289
		TCA + global scalar gate (−1.0)	2	64.229	174.065	0.6615
		TCA + global scalar gate (−2.0)	2	64.657	177.322	0.6318
		TCA + horizon-specific gate	$2H$	66.400	182.458	0.6130
		TCA + head-specific gate	$2M$	64.044	176.443	0.6717
	$H = 8$	TCA without gate	0	78.339	219.013	0.6970
		TCA + global scalar gate (−1.0)	2	78.681	217.971	0.6715
		TCA + global scalar gate (−2.0)	2	80.129	228.033	0.6911
		TCA + horizon-specific gate	$2H$	79.259	225.395	0.7137
		TCA + head-specific gate	$2M$	79.107	219.575	0.7156
	$H = 12$	TCA without gate	0	87.320	248.286	0.6640
		TCA + global scalar gate (−1.0)	2	85.688	244.514	0.6663
		TCA + global scalar gate (−2.0)	2	85.347	242.853	0.6742
		TCA + horizon-specific gate	$2H$	85.718	243.576	0.6420
		TCA + head-specific gate	$2M$	84.790	241.086	0.6457

5.6. Cross-Regional and Seasonal Analysis of Potential Transmission and Recovery Parameters

To further assess the interpretability of the parameters generated by the temporal-gated cross-attention module, we performed cross-regional and seasonal analyses of the region-specific and lead-specific latent transmission parameters $\beta_i^{(h)}$ and latent recovery parameters $\gamma_i^{(h)}$ output during testing under the $H = 12$ setting. $\beta_i^{(h)}$ and $\gamma_i^{(h)}$ quantify transmission and recovery strengths learned by the neural network in a latent space, and these are therefore not equivalent to biological parameters that can be directly estimated from classical epidemiological investigations. Accordingly, our analyses aim to examine their internal consistency with the transmission process and their epidemiological plausibility. As shown in Figure 7, we first aggregated and averaged the parameter values associated with the same region and epidemiological week across different test samples, and we visualized their spatiotemporal patterns across the 50 U.S. States and District of Columbia using heatmaps. The results show that on both the FluView and NHSN datasets, $\beta_i^{(h)}$ exhibits pronounced regional heterogeneity and seasonal variation. During periods of elevated influenza activity (approximately from the 40th week of the current year to the 10th week of the following year), the latent transmission strength increases broadly in several high-burden states, whereas during the low-incidence phase in summer and early autumn, $\beta_i^{(h)}$ decreases to relatively low levels in most regions. By contrast, $\gamma_i^{(h)}$ shows dynamics that are comparatively complementary to $\beta_i^{(h)}$ with higher values more

frequently occurring during the post-peak decline. This suggests that as the pressure of new infections progressively weakens, the relative contribution of recovery or removal processes to latent-state evolution correspondingly increases.

Further, we aggregated the parameters from all regions by epidemiological week and compared them with the normalized observed disease intensity. On FluView, the seasonal peak of β is primarily located in the winter–spring period of heightened influenza activity, whereas higher values of γ typically lag behind the peak of β . This temporal relationship is consistent with the canonical progression of influenza epidemics, in which transmission intensifies, cases accumulate, recovery accelerates and incidence subsequently declines. A similar seasonal pattern is also observed on NHSN. In particular, β increases markedly during the winter phase of rapidly rising hospitalization burden and remains low in summer, whereas γ is relatively enhanced during the post-peak waning phase.

The quantitative analyses in Figure 8 further support these interpretations. On the FluView test set, the Spearman correlation between β and future disease activity is 0.364 ($p < 0.001$), indicating that higher latent transmission strength is typically associated with higher subsequent disease activity. Meanwhile, the Spearman correlation between γ and the declining trend of disease activity is 0.325 ($p < 0.001$), suggesting that higher latent recovery strength is more likely to occur when disease activity begins to recede. On the NHSN test set, the Spearman correlation between β and future disease burden is 0.588 ($p < 0.001$), whereas the correlation between γ and the declining trend is 0.181 ($p < 0.001$). Collectively, these results indicate that the proposed parameter-generation module does not merely produce unstructured intermediate variables for minimizing predictive loss but instead learns coherent latent dynamical representations that align regional transmission differences with seasonal influenza activity and the epidemic waning process, thereby providing additional in-model interpretability for the mechanism-guided forecasting of EpiSTLLM.

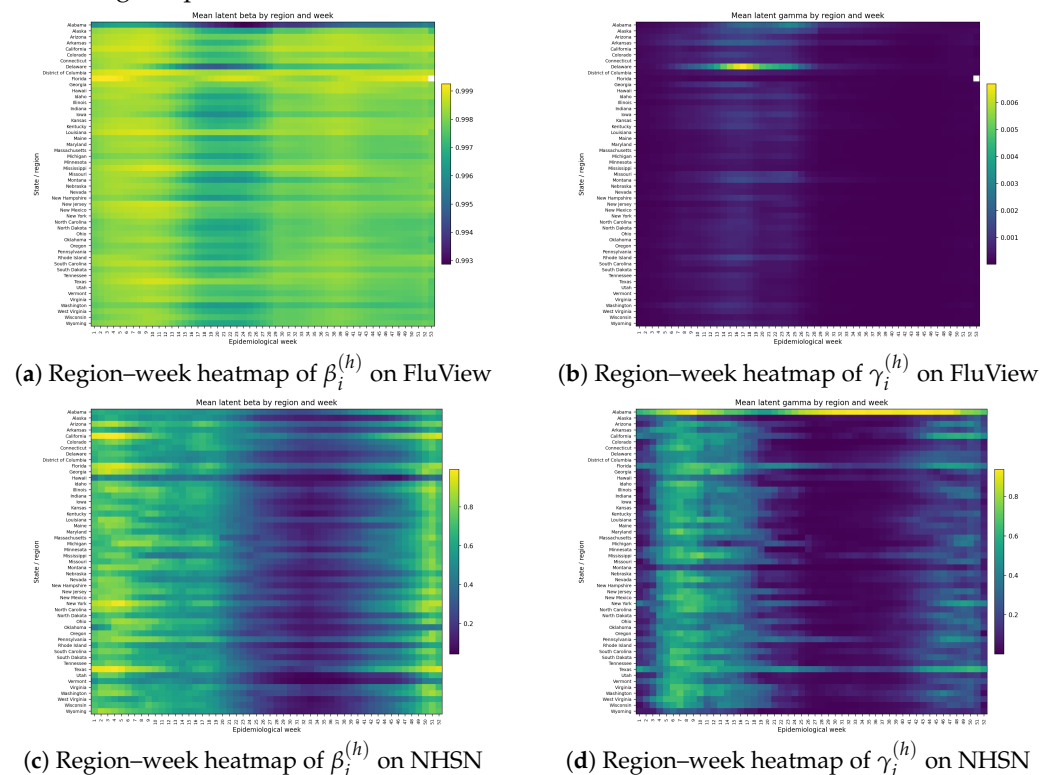


Figure 7. Region-week heatmaps of the learned transmission-rate parameters $\beta_i^{(h)}$ and recovery-rate parameters $\gamma_i^{(h)}$ on FluView and NHSN datasets.

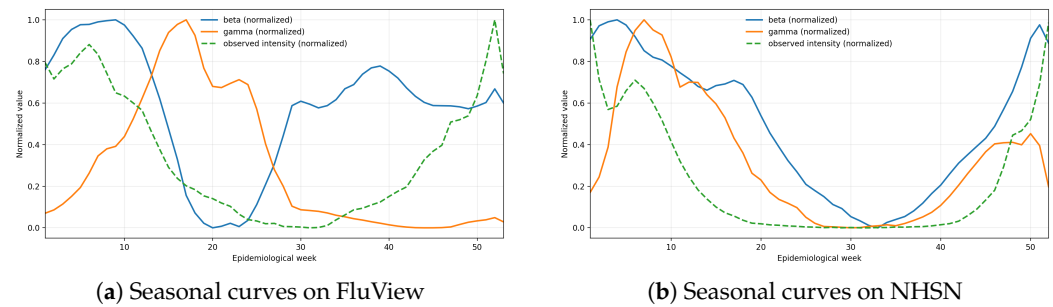


Figure 8. Seasonal curves of the learned parameters and disease intensity. (a) shows the normalized β , γ , and disease intensity on FluView; (b) shows the corresponding curves on NHSN.

5.7. S/I/R Mass-Conservation Diagnostics and Mass-Loss Analysis

To further validate mass conservation in the latent S/I/R propagation process and its effect on predictive performance, we first diagnosed the latent mass drift on the test set using the full model. The results show that under the $H = 12$ setting on the NHSN and FluView datasets, the mean mass drift of the full model is 8.11×10^{-7} and 7.09×10^{-8} , respectively, and the relative drift is approximately 1×10^{-8} in both cases, indicating that the latent S/I/R rollout approximately preserves mass during inference (Table 10). Furthermore, we compared three settings: the full model ($\lambda_{\text{mass}} = 0.01, \lambda_{\text{param}} = 0.01$), removing the mass loss while retaining the parameter-smoothing term ($\lambda_{\text{mass}} = 0, \lambda_{\text{param}} = 0.01$), and removing both the mass loss and the parameter-smoothing term ($\lambda_{\text{mass}} = 0, \lambda_{\text{param}} = 0$). The ablation results show that removing the mass loss does not cause systematic degradation in MAE or mass drift (Tables 11 and 12): The full model attains lower errors in some forecasting settings, such as $H = 4$ and $H = 8$ on FluView and $H = 12$ on NHSN; however, in several settings, models with $\lambda_{\text{mass}} = 0$ or without mechanistic regularizers achieve comparable or slightly lower MAE. Meanwhile, the mean mass drift under all ablation settings largely remains between 1×10^{-8} and 1×10^{-6} with only a modest increase at $H = 12$ on NHSN without mechanistic regularizers.

Table 10. Latent mass drift statistics on the test set ($H = 12$).

Dataset	Horizon	Split	Mean Mass drift	Max Mass Drift	Relative Drift
NHSN	$H = 12$	Test	8.11×10^{-7}	3.05×10^{-5}	2.76×10^{-8}
FluView	$H = 12$	Test	7.09×10^{-8}	1.43×10^{-6}	2.19×10^{-8}

Table 11. Mass-loss ablation: forecasting MAE. “No mech. reg.” stands for no mechanistic regularizers. The parameter pairs ($\lambda_{\text{mass}}/\lambda_{\text{param}}$) for Baseline, $\lambda_{\text{mass}} = 0$, and No mech. reg. are (0.01/0.01), (0/0.01), and (0/0), respectively.

Set	Horizon	Baseline MAE	$\lambda_{\text{mass}} = 0$ MAE	No Mech. Reg. MAE
rule FluView	$H = 4$	0.5621	0.5692	0.5696
FluView	$H = 8$	0.7483	0.7570	0.7789
FluView	$H = 12$	0.9068	0.8935	0.8963
NHSN	$H = 4$	64.229	65.502	63.680
NHSN	$H = 8$	78.681	77.777	79.889
NHSN	$H = 12$	85.688	86.726	85.956

Table 12. Mass-loss ablation: mean latent mass drift. “No mech. reg.” stands for no mechanistic regularizers. The parameter pairs $(\lambda_{\text{mass}}/\lambda_{\text{param}})$ for Baseline, $\lambda_{\text{mass}}=0$, and No mech. reg. are $(0.01/0.01)$, $(0/0.01)$, and $(0/0)$, respectively.

Dataset	Horizon	Baseline	$\lambda_{\text{mass}} = 0$	No Mech. Reg.
FluView	$H = 4$	$\sim 7 \times 10^{-8}$	5.03×10^{-8}	4.82×10^{-8}
FluView	$H = 8$	$\sim 7 \times 10^{-8}$	7.85×10^{-8}	8.11×10^{-8}
FluView	$H = 12$	$\sim 7 \times 10^{-8}$	6.88×10^{-8}	7.12×10^{-8}
NHSN	$H = 4$	$\sim 4 \times 10^{-7}$	3.71×10^{-7}	3.64×10^{-7}
NHSN	$H = 8$	$\sim 7 \times 10^{-7}$	7.78×10^{-7}	7.25×10^{-7}
NHSN	$H = 12$	$\sim 8 \times 10^{-7}$	8.28×10^{-7}	1.06×10^{-6}

This suggests that in the current implementation, the mass-conservation loss is not the primary source of performance gains; rather, it functions more as a safety constraint to prevent the accumulation of numerical drift over long-horizon rollouts, whereas approximate mass conservation mainly arises from the structured design of the S/I/R state-update process itself. These results further indicate that the long-horizon stability of EpiSTLLM does not rely on a single regularizer but instead emerges from the synergy among graph-enhanced spatiotemporal representations, horizon-adaptive parameter generation, and constraints imposed by the latent propagation mechanism.

5.8. Prediction Interval and Uncertainty Calibration

To further assess the reliability of the model’s forecasts, we complement point predictions with a post hoc prediction-interval analysis based on split-conformal prediction. Specifically, for a validation sample with point prediction $\hat{y}_i$ and ground truth y_i , the absolute residual is

$$r_i = |y_i - \hat{y}_i|. \quad (43)$$

Given a target coverage of $(1 - \alpha)$, we compute the empirical quantile q_α from the residual set $\{r_i\}$ and construct the prediction interval around a test prediction $\hat{y}$ as

$$\mathcal{C}_\alpha(\hat{y}) = [\hat{y} - q_\alpha, \hat{y} + q_\alpha], \quad (44)$$

where q_α specifies an error bound determined from validation residuals. This interval can be interpreted as adding a safety margin on both sides of the point prediction, which is estimated from the validation error distribution. Accordingly, the proposed model remains a deterministic point forecaster, and the additional intervals serve only as post hoc uncertainty calibration based on validation residuals, providing a complementary assessment of forecast uncertainty.

To evaluate interval quality, we use three metrics: prediction interval coverage probability (PICP), mean prediction interval width (MPIW) and the Winkler score. Let the test set contain N samples; for sample i , denote the lower and upper bounds by L_i and U_i , and the ground truth by y_i . PICP measures the fraction of observations covered by the interval as

$$\text{PICP} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(L_i \leq y_i \leq U_i), \quad (45)$$

where $\mathbb{I}(\cdot)$ is an indicator function that equals 1 if the observation lies within the interval and 0 otherwise. A PICP closer to the target coverage indicates better empirical calibration. For example, when the target coverage is 90%, the ideal PICP is close to 90%; a markedly

lower PICP indicates under-coverage, whereas a markedly higher PICP suggests that the interval may be overly conservative. MPIW measures the average interval width as

$$\text{MPIW} = \frac{1}{N} \sum_{i=1}^N (U_i - L_i). \quad (46)$$

This metric reflects interval tightness: a smaller MPIW indicates narrower, more informative intervals. However, MPIW must be interpreted jointly with PICP, because overly narrow intervals can fail to cover true observations and lead to under-coverage. The Winkler score jointly evaluates interval width and coverage; for a target coverage of $(1 - \alpha)$; it is defined as

$$\text{WS} = \frac{1}{N} \sum_{i=1}^N \left[(U_i - L_i) + \frac{2}{\alpha} (L_i - y_i) \mathbb{I}(y_i < L_i) + \frac{2}{\alpha} (y_i - U_i) \mathbb{I}(y_i > U_i) \right]. \quad (47)$$

The first term $(U_i - L_i)$ represents the interval width, whereas the second and third terms penalize cases where the observation falls below the lower bound or above the upper bound. Thus, a lower Winkler score indicates a better balance between coverage and tightness.

Standard split-conformal prediction calibrates with a single global residual quantile and therefore typically yields approximately equal-width intervals. However, influenza surveillance data can be heteroscedastic, such that prediction errors vary with epidemic intensity or the magnitude of the forecast. To further examine this issue, we introduce a normalized conformal variant. This approach first calibrates using relative residuals as

$$s_i = \frac{|y_i - \hat{y}_i|}{|\hat{y}_i| + \epsilon}, \quad (48)$$

where ϵ is a small positive constant to avoid numerical instability when the denominator is small. We then compute the empirical quantile q_α^{norm} from relative residuals and construct an amplitude-adaptive interval at test time:

$$\mathcal{C}_\alpha^{\text{norm}}(\hat{y}) = [\hat{y} - q_\alpha^{\text{norm}}(|\hat{y}| + \epsilon), \hat{y} + q_\alpha^{\text{norm}}(|\hat{y}| + \epsilon)]. \quad (49)$$

Compared with absolute-residual conformal prediction, normalized conformal allows larger predictions to receive wider intervals and smaller predictions to receive narrower intervals, thereby providing a more flexible characterization of uncertainty across different epidemic intensities. Empirically, as demonstrated in Table 13, on NHSN, split-conformal intervals at a 90% target coverage achieve or exceed the desired coverage. Specifically, PICP is 90.3%, 95.4% and 96.7% for $H = 4$, $H = 8$ and $H = 12$, respectively. The $H = 4$ coverage is close to the target, indicating good empirical calibration, whereas the higher PICP at $H = 8$ and $H = 12$ suggests that standard split-conformal becomes conservative at longer horizons, covering more observations at the cost of wider intervals. Consistently, MPIW increases from 230.6 to 780.3 with horizon increases, and the Winkler score increases from 762.2 to 1105.0, indicating substantially higher uncertainty at longer leads. By contrast, on FluView, split-conformal exhibits marked under-coverage at a 90% target: PICP is 73.6%, 72.2% and 71.1% for $H = 4$, $H = 8$ and $H = 12$, respectively. This indicates that a single global residual quantile is insufficient to capture heteroscedastic errors and long-horizon uncertainty on FluView. Although MPIW increases from 1.21 to 2.01 with horizon increases, coverage remains below the target, implying that the intervals are not wide enough to accommodate true variability; we therefore discuss these conformal results as a limitation of the current post hoc uncertainty estimation.

Table 13. Conformal prediction interval results. We report empirical coverage (PICP), mean prediction interval width (MPIW), and the Winkler score under a target coverage of 90%.

Dataset (Horizon)	Target Coverage	PICP	MPIW	Winkler Score
NHSN ($H = 4$)	90%	90.3%	230.6	762.2
NHSN ($H = 8$)	90%	95.4%	512.1	897.5
NHSN ($H = 12$)	90%	96.7%	780.3	1105.0
FluView ($H = 4$)	90%	73.6%	1.21	5.70
FluView ($H = 8$)	90%	72.2%	1.56	7.50
FluView ($H = 12$)	90%	71.1%	2.01	8.20

Furthermore, normalized conformal improves interval quality in several settings. Relative to absolute-residual conformal, it substantially reduces the Winkler score in multiple experiments (Table 14). For example, for NHSN at $H = 12$, the Winkler score decreases from 1105.0 to 684.5 (38% improvement); for FluView at $H = 4$ with 95% coverage, it decreases from 8.87 to 6.12 (31% improvement); and for FluView at $H = 8$ with 90% coverage, it decreases from 7.50 to 6.57 (12% improvement). These results indicate that adapting interval width to prediction magnitude can, to some extent, improve the trade-off between coverage and tightness, particularly when the error scale varies with epidemic intensity.

Table 14. Normalized conformal improvement. Winkler scores are reported before and after normalizing interval width by the prediction magnitude. “abs” denotes the absolute (unnormalized) Winkler score and “norm” denotes the score after normalization. Lower values indicate better calibration.

Dataset (Horizon)	Coverage	Winkler (abs → norm)	Improvement
NHSN ($H = 12$)	90%	1105.0 → 684.5	−38%
FluView ($H = 4$)	95%	8.87 → 6.12	−31%
FluView ($H = 8$)	90%	7.50 → 6.57	−12%

5.9. Computational Efficiency and Scalability

We further report the training time, parameter scale and GPU memory consumption of EpiSTLLM across datasets and forecast horizons (Tables 15 and 16). All experiments were conducted on a single NVIDIA RTX 4090 GPU (24 GB VRAM). The model uses GPT-2 base as the time-series representation backbone with approximately 94.2 M total parameters; with LoRA-based parameter-efficient fine-tuning, only 14.7 M parameters are trainable (15.6% of the total), substantially reducing training overhead. On NHSN, the per-epoch training time is 0.233 s and 0.310 s for $H = 4$ and $H = 12$, with total training times of approximately 189 s and 228 s, respectively; on FluView, the per-epoch training time is 0.548 s and 0.808 s for $H = 4$ and $H = 12$, with total training times of approximately 250 s and 355 s, respectively. Peak GPU memory usage increases with the forecast horizon yet remains only about 3089 MB at $H = 12$, well below the 24 GB capacity of the RTX 4090. These results indicate that despite incorporating graph-enhanced representations, LLM-based representation learning and a latent S/I/R dynamic propagation module, EpiSTLLM retains a low practical training cost with a total training time of approximately 3–6 min for 400 epochs. Given that the current study considers 51 state-level nodes, the model is computationally feasible at this scale; moreover, LoRA fine-tuning and the modular graph-propagation design provide an implementation basis for scaling to larger regional networks. At inference time, a single forward pass yields multi-step forecasts, enabling near-real-time public-health prediction and routine surveillance use cases.

Table 15. Computational efficiency on NHSN. Total parameters are fixed by the backbone; only LoRA and task-specific modules are trained.

Setting	Total Params	Trainable Params	s/Epoch	Total Time	Peak GPU Mem
NHSN ($H = 4$)	94.2 M	14.7 M (15.6%)	0.233	189 s (~ 3.2 min)	1914 MB
NHSN ($H = 12$)	94.2 M	14.7 M (15.6%)	0.310	228 s (~ 3.8 min)	3089 MB

Table 16. Computational efficiency on FluView. Total parameters are fixed by the backbone; only LoRA and task-specific modules are trained.

Setting	Total Params	Trainable Params	s/Epoch	Total Time	Peak GPU Mem
FluView ($H = 4$)	94.2 M	14.7 M (15.6%)	0.548	250 s (~ 4.2 min)	1914 MB
FluView ($H = 12$)	94.2 M	14.7 M (15.6%)	0.808	355 s (~ 5.9 min)	3089 MB

6. Conclusions

In this paper, we address multi-regional influenza transmission forecasting with an emphasis on long-horizon multi-step prediction, explicit spatial coupling and mechanism-aware interpretability. We propose EpiSTLLM, which is a unified framework that (i) represents regions as graph nodes and injects boundary-based adjacency as a spatial prior, (ii) leverages a graph-enhanced pretrained LLM backbone (efficiently adapted via LoRA) to encode long-range temporal context in regional surveillance trajectories, and (iii) introduces a temporal-gated cross-attention parameter generator to produce lead-specific propagation parameters that govern a latent S/I/R-inspired mechanism module, which is complemented by a residual correction branch. This design captures the end-to-end flow from graph-informed representation learning to horizon-adaptive parameterization and constrained latent propagation, as summarized in the framework schematic, and it enables robust forecasting across heterogeneous surveillance signals. Experiments on FluView and NHSN with 4-, 8- and 12-week horizons demonstrate consistent improvements over strong statistical, deep learning, graph-based, mechanistic and hybrid baselines, supporting the effectiveness of integrating graph structure, LLM sequence representations and mechanism-inspired constraints for stable multi-step epidemic prediction.

Despite these gains, several limitations remain. First, the regional graph is constructed from state-boundary adjacency, which captures local geographic proximity but may miss long-range transmission routes driven by travel and commuting, thereby limiting the fidelity with which edge weights reflect real-world coupling strength. Second, the latent S/I/R states are mechanism-inspired internal variables rather than directly observed epidemiological compartments, so their interpretability is primarily model-centric. Third, performance can be sensitive to certain hyperparameters, particularly the initialization of the temporal gate, which modulates information exchange across forecast leads. Fourth, the deployment of AI-based epidemic forecasts entails implicit societal risks; predictions may be misinterpreted as deterministic conclusions, potentially triggering unnecessary public anxiety, resource misapplication or regional stigmatization if disseminated without appropriate epidemiological context and uncertainty boundaries. Finally, our empirical evaluation focuses on U.S. state-level influenza surveillance and has not yet been validated across other regions, spatial scales, or disease types.

Future work will integrate multi-source heterogeneous signals (e.g., mobility, demographics and intervention indicators) to build more realistic transmission graphs, develop more adaptive gating strategies, and improve the identifiability and interpretability of latent states. Extending the framework to additional pathogens and diverse geographies will further clarify its generalizability and utility for public-health decision support.

Author Contributions: Conceptualization, X.G., Z.L. (Zige Liu) and J.C.; methodology, S.C., X.G. and J.C.; software, S.C.; validation, S.C., X.G., Z.L. (Zige Liu), Z.L. (Zhijie Lin) and J.C.; formal analysis, S.C., X.G., Z.L. (Zige Liu) and J.C.; investigation, X.G., Z.L. (Zige Liu), Z.L. (Zhijie Lin) and J.C.; resources, J.C.; data curation, S.C., X.G. and Z.L. (Zhijie Lin); writing—original draft preparation, S.C. and X.G.; writing—review and editing, S.C., X.G., Z.L. (Zige Liu) and J.C.; visualization, S.C. and X.G.; supervision, J.C.; project administration, Z.L. (Zige Liu) and J.C.; funding acquisition, J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Guangdong Basic and Applied Basic Research Foundation under Grant No. 2024A1515030128 and 2023B1515120078, the Science and Technology Development Fund of Macau SAR under Grant No. 0010/2025/RIC and 0002/2024/RDP, and the Shenzhen Science and Technology Program under Grant No. KJZD20240903104200001.

Data Availability Statement: The datasets used and analyzed in this paper are available from CDC FluView at <https://www.cdc.gov/fluview/index.html> (accessed on 30 May 2026) and CDC National Healthcare Safety Network at <https://www.cdc.gov/nhsn/index.html> (accessed on 30 May 2026). The processed data supporting the findings of this paper are included in the article. Further inquiries can be directed to the corresponding author.

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

References

- Deng, S.; Wang, S.; Rangwala, H.; Wang, L.; Ning, Y. Cola-GNN: Cross-location attention based graph neural networks for long-term ILI prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 245–254. <https://doi.org/10.1145/3340531.3411975>.
- Bloch, E.M.; Sullivan, D.; Casadevall, A.; Shoham, S.; Tobian, A.A.R.; Gebo, K.A. Applying lessons of COVID-19 and other emerging infectious diseases to future outbreaks. *Mbio* **2024**, *15*, e01109-24. <https://doi.org/10.1128/mbio.01109-24>.
- Pak, A.; Adegboye, O.A.; Adekunle, A.I.; Rahman, K.M.; McBryde, E.S.; Eisen, D.P. Economic consequences of the COVID-19 outbreak: The need for epidemic preparedness. *Front. Public Health* **2020**, *8*, 241. <https://doi.org/10.3389/fpubh.2020.00241>.
- Iuliano, A.D.; Roguski, K.M.; Chang, H.H.; Muscatello, D.J.; Palekar, R.; Tempia, S.; Cohen, C.; Gran, J.M.; Schanzer, D.; Cowling, B.J.; et al. Estimates of global seasonal influenza-associated respiratory mortality: A modelling study. *Lancet* **2018**, *391*, 1285–1300. [https://doi.org/10.1016/S0140-6736\(17\)33293-2](https://doi.org/10.1016/S0140-6736(17)33293-2).
- Lee, J.M.; Jansen, R.; Sanderson, K.E.; Guerra, F.O.; Keller-Olaman, S.; Murti, M.; O'Sullivan, T.L.; Law, M.P.; Schwartz, B.; Bourns, L.E.; et al. Public health emergency preparedness for infectious disease emergencies: A scoping review of recent evidence. *BMC Public Health* **2023**, *23*, 420. <https://doi.org/10.1186/s12889-023-15313-7>.
- Johansson, M.A.; Apfeldorf, K.M.; Dobson, S.; Devita, J.; Buczak, A.L.; Baugher, B.; Moniz, L.J.; Bagley, T.; Babin, S.M.; Guven, E.; et al. An open challenge to advance probabilistic forecasting for dengue epidemics. *Proc. Natl. Acad. Sci.* **2019**, *116*, 24268–24274. <https://doi.org/10.1073/pnas.1909865116>.
- IHME COVID-19 Forecasting Team. Modeling COVID-19 scenarios for the United States. *Nat. Med.* **2021**, *27*, 94–105. <https://doi.org/10.1038/s41591-020-1132-9>.
- Scarpino, S.V.; Petri, G. On the predictability of infectious disease outbreaks. *Nat. Commun.* **2019**, *10*, 898. <https://doi.org/10.1038/s41467-019-08616-0>.
- Friedman, J.; Liu, P.; Troeger, C.E.; Carter, A.; Reiner, R.C.; Barber, R.M.; Collins, J.K.; Dandona, L.; Dandona, R.; Emerson, J.W.; et al. Predictive performance of international COVID-19 mortality forecasting models. *Nat. Commun.* **2021**, *12*, 2609. <https://doi.org/10.1038/s41467-021-22457-w>.
- Pei, S.; Kandula, S.; Yang, W. Optimizing respiratory virus surveillance networks using uncertainty propagation. *Nat. Commun.* **2021**, *12*, 222. <https://doi.org/10.1038/s41467-020-20399-3>.
- Kraemer, M.U.; Yang, C.H.; Gutierrez, B.; Wu, C.H.; Klein, B.; Pigott, D.M.; Open COVID-19 Data Working Group; Du Plessis, L.; Faria, N.R.; Li, R.; et al. The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* **2020**, *368*, 493–497. <https://doi.org/10.1126/science.abb4218>.
- Zhang, J.; Tan, S.; Peng, C.; Xu, X.; Wang, M.; Lu, W.; Wu, Y.; Sai, B.; Cai, M.; Kummer, A.G.; et al. Heterogeneous changes in mobility in response to the SARS-CoV-2 Omicron BA. 2 outbreak in Shanghai. *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2306710120. <https://doi.org/10.1073/pnas.2306710120>.
- Zhang, X.; Luo, W.; Zhu, J. Top-down and bottom-up lockdown: Evidence from COVID-19 prevention and control in China. *J. Chin. Political Sci.* **2021**, *26*, 189–211. <https://doi.org/10.1007/s11366-020-09711-6>.

14. Bastos, L.S.; Economou, T.; Gomes, M.F.; Villela, D.A.; Coelho, F.C.; Cruz, O.G.; Stoner, O.; Bailey, T.; Codeço, C.T. A modelling approach for correcting reporting delays in disease surveillance data. *Stat. Med.* **2019**, *38*, 4363–4377. <https://doi.org/10.1002/sim.8303>.
15. Shea, K.; Runge, M.C.; Pannell, D.; Probert, W.J.; Li, S.L.; Tildesley, M.; Ferrari, M. Harnessing multiple models for outbreak management. *Science* **2020**, *368*, 577–579. <https://doi.org/10.1126/science.abb9934>.
16. Wu, J.T.; Leung, K.; Leung, G.M. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: A modelling study. *Lancet* **2020**, *395*, 689–697. [https://doi.org/10.1016/S0140-6736\(20\)30260-9](https://doi.org/10.1016/S0140-6736(20)30260-9).
17. Viboud, C.; Vespignani, A. Improving epidemic surveillance and response: Big data is dead, long live big data. *Lancet Digit. Health* **2020**, *2*, e218–e220. [https://doi.org/10.1016/S2589-7500\(20\)30059-5](https://doi.org/10.1016/S2589-7500(20)30059-5).
18. The Lancet Digital Health. Artificial intelligence for COVID-19: Saviour or saboteur? *Lancet Digit. Health* **2021**, *3*, e1. [https://doi.org/10.1016/S2589-7500\(20\)30295-8](https://doi.org/10.1016/S2589-7500(20)30295-8).
19. Murphy, T.E.; Laurence, E.; Allard, A. Deep learning of contagion dynamics on complex networks. *Nat. Commun.* **2021**, *12*, 4720. <https://doi.org/10.1038/s41467-021-24732-2>.
20. Ye, Y.; Pandey, A.; Bawden, C.; Sumsuzzman, D.M.; Rajput, R.; Shoukat, A.; Singer, B.H.; Moghadas, S.M.; Galvani, A.P. Integrating artificial intelligence with mechanistic epidemiological modeling: A scoping review of opportunities and challenges. *Nat. Commun.* **2025**, *16*, 55461. <https://doi.org/10.1038/s41467-024-55461-x>.
21. Kraemer, M.U.G.; Tsui, J.L.; Chang, S.Y.; Lytras, S.; Khurana, M.P.; Vanderslott, S.; Bajaj, S.; Scheidwasser, N.; Curran-Sebastian, J.L.; Semenova, E.; et al. Artificial intelligence for infectious disease epidemiology. *Nature* **2025**, *638*, 623–635. <https://doi.org/10.1038/s41586-024-08564-w>.
22. Massonis, G.; Villaverde, A.F.; Banga, J.R. Structural identifiability and observability of compartmental models of the COVID-19 pandemic *Annu. Rev. Control* **2021**, *51*, 441–459. <https://doi.org/10.1016/j.arcontrol.2020.12.001>.
23. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>.
24. Cho, K.; Van Merriënboer, B.; Gulçehre, Ç.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2014; pp. 1724–1734. <https://doi.org/10.3115/v1/D14-1179>.
25. Gao, J.; Sharma, R.; Qian, C.; Glass, L.M.; Spaeder, J.; Romberg, J.; Sun, J.; Xiao, C. STAN: Spatio-temporal attention network for pandemic prediction using real-world evidence. *J. Am. Med. Inform. Assoc.* **2021**, *28*, 733–743. <https://doi.org/10.1093/jamia/ocaa322>.
26. Xie, F.; Zhang, Z.; Li, L.; Zhou, B.; Tan, Y. EpiGNN: Exploring spatial transmission with graph neural network for regional epidemic forecasting. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*; Springer: Cham, Switzerland, 2022; pp. 469–485. https://doi.org/10.1007/978-3-031-26422-1_29.
27. Wang, X.; Jin, Z. Multi-region infectious disease prediction modeling based on spatio-temporal graph neural network and the dynamic model. *PLoS Comput. Biol.* **2025**, *21*, e1012738. <https://doi.org/10.1371/journal.pcbi.1012738>.
28. Cao, Q.; Jiang, R.; Yang, C.; Fan, Z.; Song, X.; Shibasaki, R. Metapopulation graph neural networks: Deep metapopulation epidemic modeling with human mobility. *arXiv* **2023**, arXiv:2306.14857. https://doi.org/10.1007/978-3-031-26422-1_28.
29. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Toulon, France, 24–26 April 2017. <https://doi.org/10.48550/arXiv.1609.02907>.
30. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Vancouver, BC, Canada, 30 April–3 May 2018. <https://doi.org/10.48550/arXiv.1710.10903>.
31. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2017; pp. 5998–6008. <https://doi.org/10.48550/arXiv.1706.03762>.
32. Dao, T.; Fu, D.Y.; Ermon, S.; Rudra, A.; Ré, C. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2022; Volume 35, pp. 16344–16359. <https://doi.org/10.48550/arXiv.2205.14135>.
33. Dao, T. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Vienna, Austria, 7–11 May 2024. <https://doi.org/10.48550/arXiv.2307.08691>.
34. Tay, Y.; Dehghani, M.; Bahri, D.; Metzler, D. Efficient Transformers: A survey. *ACM Comput. Surv.* **2023**, *55*, 1–28. <https://doi.org/10.1145/3530811>.

35. Zhou, T.; Niu, P.; Wang, X.; Sun, L.; Jin, R. One fits all: Power general time series analysis by pretrained LM. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2023; pp. 43322–43355. <https://doi.org/10.48550/arXiv.2302.11939>.
36. Jin, M.; Wang, S.; Ma, L.; Chu, Z.; Zhang, J.Y.; Shi, X.; Chen, P.Y.; Liang, Y.; Li, Y.F.; Pan, S.; et al. Time-LLM: Time series forecasting by reprogramming large language models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, Vienna, Austria, 7–11 May 2024. <https://doi.org/10.48550/arXiv.2310.01728>.
37. Nie, Y.; Nguyen, N.H.; Sinthong, P.; Kalagnanam, J. A time series is worth 64 words: Long-term forecasting with Transformers. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Kigali, Rwanda, 1–5 May 2023. <https://doi.org/10.48550/arXiv.2211.14730>.
38. Gruver, N.; Finzi, M.; Qiu, S.; Wilson, A.G. Large language models are zero-shot time series forecasters. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2023; pp. 16520–16542. <https://doi.org/10.48550/arXiv.2310.07820>.
39. Liu, C.; Hettige, K.H.; Xu, Q.; Long, C.; Xiang, S.; Cong, G.; Li, Z.; Zhao, R. ST-LLM+: Graph enhanced spatio-temporal large language models for traffic prediction. *IEEE Trans. Knowl. Data Eng.* **2025**, *37*, 4846–4859. <https://doi.org/10.1109/TKDE.2025.3567890>.
40. Du, H.; Zhao, Y.; Zhao, J.; Xu, S.; Lin, X.; Chen, Y.; Gardner, L.M.; Yang, H.F. Advancing real-time infectious disease forecasting using large language models. *Nat. Comput. Sci.* **2025**, *5*, 467–480. <https://doi.org/10.1038/s43588-025-00798-6>.
41. Liu, Z.; Ni, J.; Wang, B.; Lau, M.S.; Jin, W. Pre-training epidemic time series forecasters with compartmental prototypes. *arXiv* **2025**, arXiv:2502.03393. <https://doi.org/10.48550/arXiv.2502.03393>.
42. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Virtual, 25–29 April 2022. <https://doi.org/10.48550/arXiv.2106.09685>.
43. Dettmers, T.; Pagnoni, A.; Holtzman, A.; Zettlemoyer, L. QLoRA: Efficient finetuning of quantized LLMs. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2023; Volume 36, pp. 10088–10115. <https://doi.org/10.48550/arXiv.2305.14314>.
44. Rodríguez, A.; Cui, J.; Ramakrishnan, N.; Adhikari, B.; Prakash, B.A. EINN: Epidemiologically-informed neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Washington, DC, USA, 2023; pp. 14453–14460. <https://doi.org/10.1609/aaai.v37i12.26690>.
45. Wang, L.; Adiga, A.; Chen, J.; Sadilek, A.; Venkatramanan, S.; Marathe, M. CausalGNN: Causal-based graph neural networks for spatio-temporal epidemic forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Washington, DC, USA, 2022; pp. 12191–12199. <https://doi.org/10.1609/aaai.v36i11.21479>.
46. Brockmann, D.; Helbing, D. The hidden geometry of complex, network-driven contagion phenomena. *Science* **2013**, *342*, 1337–1342. <https://doi.org/10.1126/science.1245200>.
47. Centers for Disease Control and Prevention. FluView. Online Dataset. 2026. Available online: <https://www.cdc.gov/fluview/index.html> (accessed on 30 May 2026)).
48. Carnegie Mellon University Delphi Research Group. Delphi Epidata API: FluView (ILINet). Online API Documentation. 2026. Available online: <https://cmu-delphi.github.io/delphi-epidata/api/fluview.html> (accessed on 30 May 2026).
49. Centers for Disease Control and Prevention. Weekly Hospital Respiratory Data (HRD) Metrics by Jurisdiction, National Healthcare Safety Network (NHSN). Online Dataset. 2026. Available online: https://data.cdc.gov/Public-Health-Surveillance/Weekly-Hospital-Respiratory-Data-HRD-Metrics-by-Ju/ua7e-t2fy/about_data (accessed on 30 May 2026).
50. Box, G.E.; Jenkins, G.M.; Reinsel, G.C.; Ljung, G.M. *Time Series Analysis: Forecasting and Control*; John Wiley & Sons: Hoboken, NJ, USA, 2015. <https://doi.org/10.1002/9781118619193>.
51. Kilian, L. New introduction to multiple time series analysis, by helmut lütkepohl, springer, 2005. *Econom. Theory* **2006**, *22*, 961–967. <https://doi.org/10.1017/S0266466606000442>.
52. Graves, A. *Supervised Sequence Labelling with Recurrent Neural Networks*; Studies in Computational Intelligence; Springer: Berlin/Heidelberg, Germany, 2012; Volume 385. <https://doi.org/10.1007/978-3-642-24797-2>.
53. Cho, K.; Van Merriënboer, B.; Gulçehre, Ç.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2014; pp. 1724–1734. <https://doi.org/10.3115/v1/d14-1179>.
54. Yu, B.; Yin, H.; Zhu, Z. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv* **2017**, arXiv:1709.04875. <https://doi.org/10.24963/ijcai.2018/505>.
55. Li, Y.; Yu, R.; Shahabi, C.; Liu, Y. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv* **2017**, arXiv:1707.01926. <https://doi.org/10.48550/arxiv.1707.01926>.

56. Liu, C.; Yang, S.; Xu, Q.; Li, Z.; Long, C.; Li, Z.; Zhao, R. Spatial-temporal large language model for traffic prediction. In *Proceedings of the 2024 25th IEEE International Conference on Mobile Data Management (MDM)*; IEEE: New York, NY, USA, 2024; pp. 31–40. <https://doi.org/10.1109/mdm61037.2024.00025>.
57. Chen, Y.; Wang, X.; Xu, G. Gatgpt: A pre-trained large language model with graph attention network for spatiotemporal imputation. *arXiv* **2023**, arXiv:2311.14332. <https://doi.org/10.48550/arXiv.2311.14332>.
58. Harko, T.; Lobo, F.S.; Mak, M.K. Exact analytical solutions of the Susceptible-Infected-Recovered (SIR) epidemic model and of the SIR model with equal death and birth rates. *Appl. Math. Comput.* **2014**, *236*, 184–194. <https://doi.org/10.1016/j.amc.2014.03.030>.
59. Heng, K.; Althaus, C.L. The approximately universal shapes of epidemic curves in the Susceptible-Exposed-Infectious-Recovered (SEIR) model. *Sci. Rep.* **2020**, *10*, 19365. <https://doi.org/10.1038/s41598-020-76563-8>.
60. Bivand, R.S.; Pebesma, E.; Gomez-Rubio, V. *Applied Spatial Data Analysis with R*, 2nd ed.; Springer: New York, NY, USA, 2013. <https://doi.org/10.1007/978-1-4614-7618-4>.
61. Gu, J.; Jia, Z.; Cai, T.; Song, X.; Mahmood, A. Dynamic Correlation Adjacency-Matrix-Based Graph Neural Networks for Traffic Flow Prediction. *Sensors* **2023**, *23*, 2897. <https://doi.org/10.3390/s23062897>.
62. Xia, Y.; Bjornstad, O.N.; Grenfell, B.T. Measles Metapopulation Dynamics: A Gravity Model for Epidemiological Coupling and Dynamics. *Am. Nat.* **2004**, *164*, 267–281. <https://doi.org/10.1086/422341>.

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