Bridging Dynamics and Data: A Unified Diffusion Framework for Mechanistically-Informed Epidemic Forecasting

Guanghui Min [1] **Tianhao Huang** [1] **Ke Wan** [1] **Qi R. Wang** [2] **Chen Chen** [1]

Abstract

Reliable epidemic forecasting is critical for public health decision-making yet remains challenging due to data sparsity and the non-stationary nature of disease dynamics. While recent hybrid models attempt to integrate mechanistic principles with data-driven approaches, they often relegate mechanistic priors to merely auxiliary features or regularization terms. This design not only obscures the interpretability of the mechanistic contribution but also fails to inherit the capability of physical models to generalize under non-stationary dynamics, as the core architecture remains predominantly data-driven. To address these limitations, we propose `EpiDiff`, a unified framework that synergizes epidemiological domain knowledge with the generative power of diffusion models. Unlike methods that rigidly fuse features, `EpiDiff` employs a novel uncertainty-aware steering mechanism during inference. Specifically, we quantify the posterior uncertainty of mechanistic estimations and use it to dynamically modulate the diffusion process. Extensive experiments on real-world datasets demonstrate that `EpiDiff` consistently outperforms state-of-the-art baselines in accuracy and robustness, particularly under non-stationary distributions, while offering transparent insights into model reliance by explicitly visualizing when the forecast is governed by mechanistic laws versus data-driven patterns. Our code and datasets are available at `https://github.com/GuanghuiMin/EpiDiff`.

[1]Department of Computer Science, University of Virginia, Charlottesville, USA [2]Department of Civil and Environmental Engineering, Northeastern University, Boston, USA. Correspondence to: Guanghui Min <jjm8vr@virginia.edu>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

1. Introduction

In recent years, outbreaks of infectious diseases such as COVID-19 have underscored the urgent need for accurate modeling and forecasting of epidemic dynamics (Flaxman et al., 2020; Chinazzi et al., 2020; He et al., 2025; Min et al., 2025). Reliable forecasts are critical for informing policymakers about effective interventions and for guiding the allocation of limited healthcare resources (Reich et al., 2019). Existing epidemic forecasting approaches can be divided into two categories: model-based (mechanistic) and data-driven methods. Commonly used mechanistic models include patch-based and network-based compartmental frameworks (Hethcote, 2000; Pastor-Satorras et al., 2015), which rely on explicit assumptions about transmission dynamics, such as the Susceptible–Infectious–Recovered (SIR) framework (Kermack & McKendrick, 1927; Anderson, 1991), and typically use systems of differential equations with calibrated parameters to simulate future epidemic trajectories (Chowell et al., 2020; Jewell et al., 2020). The data-driven methods range from traditional statistical approaches such as autoregressive (AR) and autoregressive moving average (ARMA) models (Box et al., 2015; Brockwell & Davis, 2002), to state-space formulations such as Kalman filters (Kalman, 1960), and deep learning models (Chimmula & Zhang, 2020; Kapoor et al., 2020; Wang et al., 2020). These methods learn patterns from historical epidemiological data through statistical modeling or complex spatiotemporal architectures to generate forecasts.

In the context of emerging epidemics such as COVID-19, forecasting faces several critical challenges (Ioannidis et al., 2022): (1) surveillance data is often sparse due to the absence of long-term historical records, particularly during the initial phases of an outbreak. This scarcity could restrict the capacity of data-driven models to adequately train predictive models, often resulting in overfitting or poor generalization to unseen future states (Delussu et al., 2023). In addition, their dependence on historical patterns renders them fragile to distributional shifts induced by sudden outbreaks or behavioral changes. Such non-stationary dynamics violate the independent and identically distributed (i.i.d.) assumption, leading to out-of-distribution scenarios that undermine predictive accuracy (Kapoor et al., 2020); and (2) the disease

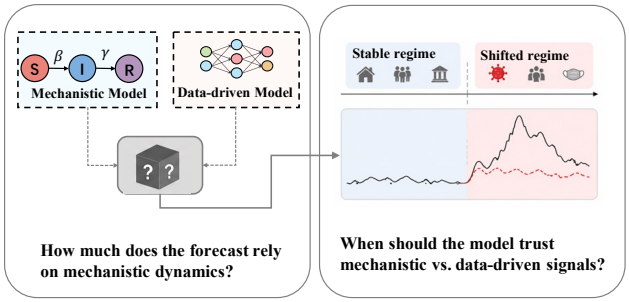


Figure 1. An illustration of Challenges of Hybrid Epidemic Forecasting Models.

spread process reflects the rapidly co-evolving interplay of individual behavior, government interventions, and disease transmission rate (Chang et al., 2021; Haug et al., 2020). This complex heterogeneity cannot be adequately captured by a single epidemic model, which inevitably suffers from structural misspecification and limited expressivity (Stocks et al., 2020; Castro et al., 2021).

One promising solution to address these challenges is to develop hybrid epidemic forecasting models that merge neural networks' predictive power with mechanistic models' fundamental principles (Liu et al., 2024). Recent approaches have begun integrating epidemiological domain knowledge with neural architectures (Song et al., 2020b), such as using compartmental model states as causal features (Wang et al., 2022) or employing SIR-inspired neural ODE formulations (Wan et al., 2025). However, existing hybrid models primarily focus on improving prediction accuracy and face two critical limitations: (1) they lack explainable mechanisms to quantify the respective contributions of mechanistic and data-driven components in the final prediction, making it difficult to explain the source and reliability of forecasts; and (2) they lack sufficient flexibility to handle the uncertainty and non-stationarity inherent in epidemic systems, where surveillance data are noisy and model parameters evolve over time (Holmdahl & Buckee, 2020; Stocks et al., 2020). The limitations are illustrated in Fig. 1. In the context of this work, we define *explainability* not as causal inference of external factors (e.g., policy effects), but as the *transparency of the inference process*.

In this paper, we address the challenges of explainability and robustness in hybrid epidemic forecasting models by developing a novel unified epidemiology-aware spatiotemporal forecasting framework based on diffusion models. We specifically choose diffusion models as the backbone data-driven approach due to their demonstrated prominence in spatiotemporal forecasting across diverse domains, including transportation (Wen et al., 2023; Zheng et al., 2023) and crime risk prediction (Wang et al., 2024). Our approach quantifies the uncertainty in mechanistic model estimates, enabling the diffusion model to be heterogeneously guided by mechanistic estimations across different spatiotemporal regions during inference. This design produces robust and explainable epidemic forecasts that leverage both mechanistic understanding and data-driven flexibility. Our key contributions are as follows:

- **Unified Hybrid Framework**: We propose `EpiDiff`, a unified epidemiology-aware spatiotemporal forecasting framework that integrates mechanistic models with diffusion-based deep learning, bridging the gap between explainability and accuracy in epidemic forecasting.
- **Uncertainty-Aware Mechanistic Guidance**: We introduce a principled mechanism to quantify the posterior uncertainty of mechanistic estimations, which dynamically modulates the influence of mechanistic estimations in the diffusion process, leading to robust and adaptive predictions across heterogeneous spatiotemporal regions.
- **Comprehensive Empirical Validation**: Extensive experiments on real-world COVID-19 and influenza datasets show that `EpiDiff` consistently outperforms state-of-the-art baselines in accuracy and robustness, while providing explainable, uncertainty-aware predictions.

2. Preliminaries

2.1. Problem Formulation

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A}, \mathbf{X})$ denote an attributed graph, where $\mathcal{V}$ is the set of nodes with $|\mathcal{V}| = N$, and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the edge set describing spatial relations. The adjacency matrix is $\mathbf{A} \in \mathbb{R}^{N \times N}$, and $\mathbf{X} \in \mathbb{R}^{N \times C}$ is the node feature matrix. Each row $\mathbf{X}_i$ contains multiple attributes of region v_i, such as epidemic indicators and auxiliary covariates.

Definition 2.1 (Spatio-Temporal Graph)**.** A spatio-temporal graph $\{\mathcal{G}^t\}_{t=1}^T$ is a sequence of graphs indexed by time t, where each $\mathcal{G}^t = (\mathcal{V}, \mathcal{E}, \mathbf{A}, \mathbf{X}^t)$ and $\mathbf{X}^t \in \mathbb{R}^{N \times C}$ denotes node features at time t. The adjacency $\mathbf{A}$ encodes spatial dependencies, while temporal features capture disease dynamics and related covariates across T steps.

We focus on predicting a subset of node features related to epidemic outcomes. Let $\mathbf{Y}^t \in \mathbb{R}^{N \times C'}$ ($C' < C$, in our paper we set $C' = 1$) denote the epidemic features (e.g., cases) extracted from $\mathbf{X}^t$.

Problem 1 (Spatio-Temporal Epidemic Forecasting)**.** *Given a sliding window of κ historical steps $\mathcal{G}^{t-\kappa:t}$, the goal is to predict future epidemic features $\mathbf{Y}^{t+1:t+h}$, where h is the prediction horizon.*

Unlike standard spatio-temporal forecasting, epidemic features follow mechanistic compartmental dynamics with heterogeneous parameters across nodes and time. To capture this, we formulate the following task.

Problem 2 (Epidemic Forecasting with Dynamic Mechanistic Guidance)**.** *Given $\mathcal{G}^{t-\kappa:t}$, assume each epidemic feature $y_i^t \in \mathbf{Y}^t$ of node v_i follows an epidemic model Φ*

with parameters Θ_i^t that vary by node and time. The task is to predict $\mathbf{Y}^{t+1:t+h}$ while ensuring consistency with the mechanistic dynamics governed by Φ.

2.2. SIR Dynamics

As a representative mechanistic model, we adopt the susceptible–infected–recovered (SIR) system (Kermack & McKendrick, 1927; Keeling & Rohani, 2008). In this formulation, the total population N is partitioned into three time-dependent compartments: S denotes the number of susceptible individuals, I the infected, and R the recovered. The evolution of these states is described by the system as follows:

$$\frac{dS}{dt} = -\beta SI, \quad \frac{dI}{dt} = \beta SI - \gamma I, \quad \frac{dR}{dt} = \gamma I, \quad (1)$$

where β is the transmission rate and γ is the recovery rate. This formulation links interpretable parameters (β, γ) to observable new cases. We adopt the SIR system as a *canonical representative* to serve as the mechanistic prior for guiding data-driven inference in our framework. It is important to note that our framework is model-agnostic regarding the choice of mechanistic dynamics; while we employ the SIR model for clarity and universality, the proposed uncertainty-aware steering mechanism can be seamlessly extended to more complex compartmental models (e.g., SEIR, SIRD) or age-stratified structures to capture finer-grained epidemiological characteristics.

3. `EpiDiff`: The Proposed Method

In our framework, we focus on predicting new cases as a representative epidemic feature with the guidance of SIR dynamics. We first introduce a metric that quantifies mechanistic uncertainty based on the mechanistic estimation error, then present a parameterized and efficient method for estimating mechanistic model parameters. Next, we detail the architecture and training objectives of our graph diffusion backbone. Finally, we present how mechanistic estimation and uncertainty characterization are used to provide heterogeneous guidance during diffusion model inference. The overall framework is illustrated in Fig. 2.

3.1. Uncertainty Quantification of Mechanistic Estimations

In an epidemic system with a given infection mechanism (e.g., SIR, SIS), epidemic dynamics are governed by deterministic differential equations (Kermack & McKendrick, 1927). Over short horizons, the model parameters (e.g., infection and recovery rates) can be regarded as stable. However, because real-world observations are noisy and incomplete, estimating these parameters is inherently uncertain. This estimation error can be summarized by a covariance matrix over the parameters, which quantifies the difficulty of accurate inference. When such parameter uncertainty is propagated through the mechanistic model, it induces variability in the predicted epidemic trajectories. We refer to this induced predictive variability as *mechanistic uncertainty*.

Definition 3.1 (Mechanistic Uncertainty)**.** Let θ denote the parameter vector of a mechanistic epidemic model Φ, with point estimate $\hat{\theta}$ and covariance Σ obtained from parameter inference. For node v_i and time t, the mechanistic uncertainty is defined as the predictive variance

$$u_{i,t}^{\mathrm{M}} = \mathrm{Var}_{\theta \sim \mathcal{N}(\hat{\theta}, \Sigma)}[\Phi(\theta)_{i,t}], \quad (2)$$

where $\Phi(\theta)_{i,t}$ denotes the model's forecast for node v_i at time t. Thus, mechanistic uncertainty measures how parameter estimation error translates into variability of epidemic forecasts.

The most common approaches for quantifying uncertainty in parameter estimation include Approximate Bayesian Computation (ABC) (Marjoram et al., 2003) and Markov chain Monte Carlo (MCMC) techniques such as Hamiltonian Monte Carlo (HMC) (Neal et al., 2011; Brooks et al., 2011). These methods approximate the posterior distribution but are computationally intensive, particularly for high-dimensional or time-sensitive forecasting tasks.

To efficiently approximate the posterior distribution of the parameters for each node, we adopt a Gaussian parameterization of the posterior. Specifically, we work in the transformed parameter space $\theta_i = (\log R_{0,i}, \log(1/\gamma_i))$, where $R_{0,i} = \beta_i/\gamma_i$ is the basic reproduction number. This log-transformation ensures positivity of the original epidemiological parameters and provides a numerically stable representation. Moreover, $R_{0,i}$ and $1/\gamma_i$ (the mean infectious period) are epidemiologically interpretable parameters that have been extensively studied in public health and clinical research. This choice of parameterization allows us to directly incorporate domain knowledge into the prior distribution $p(\theta_i)$ and ensures the plausibility of the parameter estimates.

We define the negative log-posterior of node i as

$$\mathcal{J}_i(\theta_i) = -\log p(y_i^{t-\kappa:t-1} \mid \theta_i) - \log p(\theta_i), \quad (3)$$

where $y_i^{t-\kappa:t-1}$ denotes the epidemic features of node i observed in the time window. The posterior estimate is obtained by minimizing $\mathcal{J}_i(\theta_i)$, which can be efficiently solved using quasi-Newton methods such as L-BFGS-B (Byrd et al., 1995). We then compute the Hessian of $\mathcal{J}_i$ at $\hat{\theta}_i$, i.e., $\mathbf{H}_i = \nabla^2_{\theta_i} \mathcal{J}_i(\theta_i)|_{\theta_i = \hat{\theta}_i}$, via finite-difference approximation. A second-order Taylor expansion around $\hat{\theta}_i$ yields the Laplace approximation:

$$p(\theta_i \mid y_i^{t-\kappa:t-1}) \approx \mathcal{N}(\hat{\theta}_i, \mathbf{H}_i^{-1}), \quad (4)$$

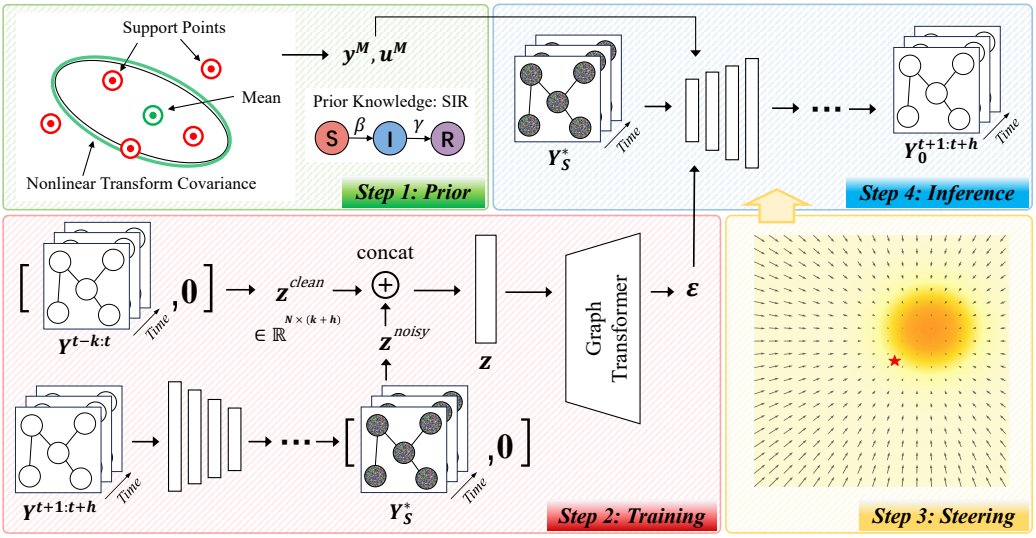


Figure 2. Overview of the EpiDiff pipeline. **Step 1: Prior**. We first fit a mechanistic SIR model to the historical epidemic trajectory and generate sigma-point-based posterior estimations of the future, including both the mean trajectory $\mathbf{y}^M$ and its uncertainty $\mathbf{u}^M$. **Step 2: Training**. We train a spatio-temporal diffusion backbone using observed epidemic data, which learns to generate future trajectories conditioned on past windows and graph structure. **Step 3: Steering**. During inference, we inject mechanistic prior information: $\mathbf{y}^M$ acts as a soft target and $\mathbf{u}^M$ modulates the guidance strength, enabling adaptive, uncertainty-aware generation. **Step 4: Inference**. The diffusion model generates future trajectories from noise, conditioned on history, graph structure, and dynamically steered by the mechanistic estimation.

where $\mathbf{H}_i^{-1}$ serves as the local covariance estimate of the parameters. We now propagate parameter uncertainty through the nonlinear epidemic dynamics. Let Φ denote the state-transition operator (i.e. (1) for SIR dynamics), $y_i^{t+1} = \Phi(y_i^t, \theta_i)$, where $\mathbf{y}_t^i$ denotes the epidemic state of node i at time t. The predictive distribution is formally

$$p(y_i^{t+1}) = \int \Phi(y_i^t, \theta_i)\, p(\theta_i \mid y_i^{t-\kappa:t-1})\, d\theta_i, \tag{5}$$

which is generally intractable. To approximate this integral, we follow the standard deterministic construction used in nonlinear transforms, where a set of $2d+1$ support points $\{\theta_i^{(m)}\}_{m=0}^{2d}$ with associated weights $\{w^{(m)}\}$ are chosen such that

$$\begin{aligned} \hat{\theta}_i &= \sum_{m=0}^{2d} w^{(m)} \theta_i^{(m)}, \\ \mathbf{H}_i^{-1} &= \sum_{m=0}^{2d} w^{(m)} (\theta_i^{(m)} - \hat{\theta}_i)(\theta_i^{(m)} - \hat{\theta}_i)^\top. \end{aligned} \tag{6}$$

In the SIR dynamics considered here, the parameter dimension is $d = 2$, so only $2d + 1 = 5$ support points are required per node. The construction algorithm is provided in Appendix B.1.

Theorem 3.2 (Second-order accuracy of propagation)**.** *Let $g : \mathbb{R}^d \to \mathbb{R}$ be any twice continuously differentiable function. For a Gaussian random vector $\theta \sim \mathcal{N}(\hat{\theta}, \Sigma)$, the deterministic propagation scheme described above exactly reproduces the first and second moments of $g(\theta)$ up to $\mathcal{O}(\|\theta - \hat{\theta}\|^3)$.*

Theorem 3.2 establishes that the deterministic support point construction above provides an accurate second-order approximation to the predictive mean and variance under nonlinear dynamics. The full proof is provided in Appendix A.1. Propagating each support point through the dynamics yields $y_i^{t+1,(m)} = \Phi(y_i^t, \theta_i^{(m)})$. The predictive mean and uncertainty are then approximated as

$$y_{i,t+1}^{\mathrm{M}} \triangleq \sum_{m=0}^{2d} w^{(m)} y_i^{t+1,(m)}, \tag{7}$$

$$u_{i,t+1}^{\mathrm{M}} \triangleq \sum_{m=0}^{2d} w^{(m)} (y_i^{t+1,(m)} - y_{i,t+1}^{\mathrm{M}})^2. \tag{8}$$

3.2. Diffusion Backbone with Spatiotemporal Graph Transformer

In Section 3.1, we provided mechanistic point estimates together with principled uncertainty quantification. However, such estimates have two critical limitations: (i) they cannot capture long-time global spatiotemporal dependencies across the network, and (ii) they are inflexible in accommodating rich contextual information such as mobility patterns, vaccination records, or other exogenous covariates. To overcome these limitations, we introduce a diffusion-based data-driven backbone with a spatiotemporal graph transformer, which learns predictive distributions directly from historical data and integrates diverse contextual signals within a unified generative framework.

During training, we follow the standard denoising diffusion

framework (Ho et al., 2020; Song et al., 2020a). To formalize the conditional denoising process, let $\mathbf{Y}^{t-\kappa:t}$ denote the historical observations and $\mathbf{Y}^{t+1:t+h}$ the clean future trajectory. At a sampled diffusion step s, the future is corrupted as

$$\mathbf{Y}_s^* = \sqrt{\bar{\alpha}_s}\,\mathbf{Y}^{t+1:t+h} + \sqrt{1-\bar{\alpha}_s}\,\varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I), \quad (9)$$

where $\bar{\alpha}_s$ is determined by the noise schedule. To avoid information leakage, we construct two complementary streams: a clean stream $\mathbf{Z}_{\text{clean}} = [\mathbf{Y}^{t-\kappa:t}; \mathbf{0}_{N\times\kappa}]^\top$ and a noisy stream $\mathbf{Z}_{\text{noisy}} = [\mathbf{0}_{N\times h}; \mathbf{Y}_s^*]^\top$. These are concatenated along the temporal axis as $\mathbf{Z} = \text{Concat}(\mathbf{Z}_{\text{clean}}, \mathbf{Z}_{\text{noisy}})$, which is then projected into the latent dimension and fed into the spatiotemporal graph transformer.

The transformer enforces structural priors through a masked attention mechanism. Let $\mathbf{A} \in \{0,1\}^{N\times N}$ be the adjacency matrix (with self-loops added). We lift $\mathbf{A}$ to the spatiotemporal domain by blockwise repetition, $\tilde{\mathbf{A}} = \mathbf{A} \otimes \mathbf{I}_{2T} \in \{0,1\}^{(N\cdot 2T)\times(N\cdot 2T)}$, where $T = \kappa + h$ and convert it into an additive bias matrix

$$\mathbf{B}_{ij} = \begin{cases} 0, & \tilde{\mathbf{A}}_{ij} = 1, \\ -\infty, & \tilde{\mathbf{A}}_{ij} = 0, \end{cases} \quad (10)$$

so that attention scores between non-adjacent nodes are suppressed. This guarantees that message passing is restricted to valid spatial edges and self-loops across all temporal slices. The diffusion step s is encoded through a sinusoidal time embedding followed by adaptive layer normalization, providing step-dependent modulation of the hidden states.

After concatenation, the hidden sequence has length $2T$. We apply a linear layer along the temporal dimension that maps $\mathbb{R}^{2T} \to \mathbb{R}^T$, so that the resulting noise estimate $\hat{\varepsilon}$ is temporally aligned with the prediction horizon. Since only the prediction horizon should be supervised, we minimize

$$\mathcal{L} = \|\hat{\varepsilon}^{t+1:t+h} - \varepsilon\|_2^2, \quad (11)$$

where $\hat{\varepsilon}^{t+1:t+h}$ denotes the last h steps of $\hat{\varepsilon}$ aligned with the noisy target $\mathbf{Y}^*$. This objective trains the backbone to denoise corrupted futures while capturing long-range spatiotemporal dependencies under the structural guidance of $\mathcal{G}^{t-\kappa:t}$.

3.3. Steering Diffusion Inference with Uncertain Mechanistic Estimation

While the diffusion backbone provides a strong data-driven framework for capturing global spatiotemporal dependencies and contextual signals, such purely data-driven modeling remains essentially a fit to past history and offers limited guarantees of explainability and robustness. This motivates us to further incorporate mechanistic estimations into the inference process, so that the final forecasts are not only accurate but also epidemiologically consistent.

We incorporate estimates from Section 3.1 as a Gaussian prior distribution directly in the denoising process. Our inference procedure is based on deterministic DDIM sampling (Song et al., 2020a), which enables efficient generation with far fewer steps than ancestral DDPM (Ho et al., 2020). In the standard DDIM update, the clean sample $\mathbf{Y}_0^{t+1:t+h}$ (i.e. the clean prediction for future window $\mathbf{Y}^{t+1:t+h}$) at step s is reconstructed from the noisy state $\mathbf{y}_s$ and the predicted noise $\hat{\varepsilon}_\psi$ (from Section 3.2, ψ is the trained parameters for diffusion backbone) as

$$\mathbf{Y}_0^{t+1:t+h} = \frac{1}{\sqrt{\bar{\alpha}_s}}(\mathbf{Y}_s^{t+1:t+h} - \sqrt{1-\bar{\alpha}_s}\hat{\varepsilon}_\psi), \quad (12)$$

where $\bar{\alpha}_s$ is determined by the noise schedule. Here the first term represents the denoised estimate of the clean trajectory, while the second term injects the residual noise predicted by the backbone.

We define the diagonal covariance matrix $\mathbf{\Sigma}_{\text{M}} = \text{diag}(\mathbf{u}_{\text{vec}}^{\text{M}}) \in \mathbb{R}^{Nh\times Nh}$. From a Bayesian perspective, we formulate the inference process as sampling from a posterior distribution that harmonizes data-driven patterns with physical consistency. We treat the pre-trained diffusion backbone as providing a *data-driven prior* $p_\psi(\mathbf{y}_0 \mid \cdot)$, while the mechanistic estimation provides a *physical likelihood* $p(\mathbf{y}^{\text{M}} \mid \mathbf{y}_0)$. Assuming the mechanistic estimate is a noisy observation of true state following a Gaussian distribution $\mathcal{N}(\mathbf{y}_0, \mathbf{\Sigma}_{\text{M}}/\tau)$, the posterior is derived via Bayes' rule:

$$\underbrace{p_{\text{posterior}}(\mathbf{y}_0)}_{\text{Target Distribution}} \propto \underbrace{p_\psi\big(\mathbf{y}_0 \mid \mathbf{Y}^{t-\kappa:t}, \mathcal{G}^{t-\kappa:t}\big)}_{\text{Data-Driven Prior}} \times \underbrace{\exp\Big(-\tfrac{\tau}{2}\,(\mathbf{y}_0 - \mathbf{y}^{\text{M}})^\top \mathbf{\Sigma}_{\text{M}}^{-1}(\mathbf{y}_0 - \mathbf{y}^{\text{M}})\Big)}_{\text{Physical Likelihood}}. \quad (13)$$

This formulation aligns with the *guidance* framework in diffusion models (Dhariwal & Nichol, 2021; Bansal et al., 2023), where the mechanistic prior acts as an energy function to steer the generation. Here, τ serves as the *guidance scale* (analogous to the inverse temperature), which modulates the sharpness of the physical likelihood. Intuitively, our method performs uncertainty-aware posterior sampling: when the mechanistic estimation is confident (small $\mathbf{u}^{\text{M}}$, i.e., high precision), the likelihood term dominates, strongly steering the forecast toward $\mathbf{Y}^{\text{M}}$; conversely, when uncertainty is high, the prior influence prevails, allowing the model to rely on spatiotemporal dependencies learned from data. The full inference algorithm, implemented via score-based gradient guidance, is detailed in Appendix B.3.

Although the likelihood is defined over the clean future trajectory y_0, guidance is applied within each DDIM step through a reconstruct–steer–re-encode procedure. At diffu-

Table 1. Performance comparison on ***covid-US***, ***covid-JP***, and ***influenza-US*** in RMSE and MAE ($\mathcal{R}$ and $\mathcal{P}$, lower is better). Best results are **in bold** and runner-ups are underlined.

Methods	*covid-US*				*covid-JP*				*influenza-US*			
	$h=7$		$h=14$		$h=7$		$h=14$		$h=7$		$h=14$	
	$\mathcal{R}\downarrow$	$\mathcal{P}\downarrow$	$\mathcal{R}\downarrow$	$\mathcal{P}\downarrow$	$\mathcal{R}\downarrow$	$\mathcal{P}\downarrow$	$\mathcal{R}\downarrow$	$\mathcal{P}\downarrow$	$\mathcal{R}\downarrow$	$\mathcal{P}\downarrow$	$\mathcal{R}\downarrow$	$\mathcal{P}\downarrow$
PatchSIR	1791±0	891±0	2077±0	886±0	313±0	98±0	332±0	106±0	323±0	82±0	454±0	138±0
AR	1791±16	835±16	1943±23	865±7	457±57	159±11	476±11	158±4	4579±2429	2040±1814	2328±236	1172±84
ARMA	1832±29	843±2	1906±18	852±9	460±32	156±8	496±6	166±3	4423±1971	2289±1271	2314±150	1174±68
LSTM	1797±59	867±48	1956±49	962±52	408±10	135±3	398±19	135±7	353±29	119±21	354±42	133±7
GRU	1654±81	771±52	1949±38	974±30	364±12	120±3	383±8	130±3	377±41	142±13	498±303	192±75
STGCN	2131±12	877±23	2446±51	1107±12	315±83	96±23	326±24	103±9	414±93	160±42	588±128	248±56
ColaGNN	2740±758	1175±339	3055±356	1403±181	213±20	63±4	305±47	88±11	**250**±17	85±9	**347**±228	129±27
STGODE	1763±96	719±34	2716±107	1010±40	269±10	79±2	309±23	96±7	643±194	261±94	1415±460	570±200
DiffSTG	1236±28	443±11	1284±20	444±14	246±28	81±9	423±35	147±13	344±29	151±69	520±44	254±54
STAN	1364±196	1150±174	1265±98	1046±90	222±87	127±40	**224**±88	128±41	4599±42	4306±32	3750±79	3270±79
EARTH	2145±324	603±103	2664±272	794±79	303±73	100±13	394±102	134±21	539±172	190±47	2027±1160	629±302
EpiColaGNN	1825±84	745±58	2653±22	1235±9	282±33	83±11	327±19	99±6	1311±67	693±36	1514±64	807±36
`EpiDiff`	**1096**±5	**374**±1	**1232**±5	**422**±16	**188**±7	**62**±2	247±2	**86**±1	304±1	**82**±0.4	423±1	**125**±0.1

sion step s, we first reconstruct the clean estimate $\hat{y}_0$ from the noisy state y_s, apply the mechanistic guidance gradient to $\hat{y}_0$, and then map the steered estimate back to the corresponding noise level for the next step. Thus, the mechanistic likelihood never directly constrains the noisy sample; it acts on the reconstructed clean trajectory, following standard diffusion guidance practice.

4. Experiments

In this section, we comprehensively evaluate `EpiDiff` by answering the following questions:

- **Q1: Performance.** Does `EpiDiff` outperform state-of-the-art epidemic forecasting methods?
- **Q2: Effectiveness.** Does incorporating the mechanistic estimate $\mathbf{y}^{\mathrm{M}}$ and its associated uncertainty $\mathbf{u}^{\mathrm{M}}$ improve predictive performance?
- **Q3: Mechanistic Interpretation and Uncertainty.** How to explain the predictions from `EpiDiff`?
- **Q4: Sensitivity.** How does the performance of `EpiDiff` vary with hyper-parameter settings?

4.1. Experimental Setup

Datasets. We evaluate the effectiveness of `EpiDiff` on three representative datasets covering both COVID-19 and influenza-like illness: ***covid-US***, ***covid-JP***, and ***influenza-US***. ***covid-US*** contains state-level daily COVID-19 case counts in the United States from April 15, 2020 to April 15, 2021 (366 days), sourced from The New York Times. ***covid-JP*** comprises prefecture-level daily COVID-19 cases in Japan from April 1, 2020 to September 21, 2021, based on the NHK COVID-19 database. ***influenza-US*** provides weekly state-level influenza incidence in the United States over 158 weeks (2019–2021), estimated from CDC reports. Region-level graphs are constructed according to geographical adjacency following Wang et al. (2022). Data distributions and other details are provided in Appendix C.1.

Metrics. The metrics used to evaluate forecasting performance are: root mean squared error ($\mathcal{R}$, RMSE), which measures the square root of the average squared difference between predicted and true values, and mean absolute error ($\mathcal{P}$, MAE), which measures the average absolute difference between predicted and true values. Both $\mathcal{R}$ and $\mathcal{P}$ take values in $[0, +\infty)$, with smaller values indicating better performance. These two evaluation metrics are consistent with previous works (Liu et al., 2023; Wan et al., 2025). Detailed definitions and calculation formulas are provided in Appendix C.2.

Baselines. We compare our approach against a diverse set of classic and state-of-the-art forecasting models. Every model is run three times. Detailed descriptions are provided in Appendix C.3.

- *Mechanistic models*: Our proposed mechanistic algorithm **PatchSIR** in Section 3.1 based on **SIR** (Kermack & McKendrick, 1927).
- *Time-series models*: statistical approaches including **Autoregressive (AR)** and **Autoregressive Moving Average (ARMA)** (Contreras et al., 2003), as well as neural sequence models such as **Gated Recurrent Unit (GRU)** (Cho et al., 2014) and **Long Short-Term Memory (LSTM)** (Hochreiter & Schmidhuber, 1997).
- *Graph-based spatiotemporal models*: general-purpose GNN-based models including **STGCN** (Yu et al., 2017) and **ColaGNN** (Deng et al., 2020), a Neural ODE-based model **STGODE** (Fang et al., 2021), and a diffusion-based model **DiffSTG** (Wen et al., 2023). In addition, we include epidemic-aware models that incorporate domain-specific inductive biases: the attention-based model **STAN** (Gao

Table 2. Ablation study of different variants. Best results are **in bold** and runner-ups are underlined.

Variants	*covid-US*				*covid-JP*				*influenza-US*			
	$h = 7$		$h = 14$		$h = 7$		$h = 14$		$h = 7$		$h = 14$	
	$\mathcal{R}\downarrow$	$\mathcal{P}\downarrow$	$\mathcal{R}\downarrow$	$\mathcal{P}\downarrow$	$\mathcal{R}\downarrow$	$\mathcal{P}\downarrow$	$\mathcal{R}\downarrow$	$\mathcal{P}\downarrow$	$\mathcal{R}\downarrow$	$\mathcal{P}\downarrow$	$\mathcal{R}\downarrow$	$\mathcal{P}\downarrow$
w/o Diffusion	1791	891	2077	886	313	98	332	106	323	82	454	138
w/o Steering	**1080**±2	**348**±1	**1199**±6	**386**±11	370±22	94±5	417±4	121±1	438±19	224±7	1787±87	928±59
w/o Uncertainty	1099±5	376±1	1253±5	438±18	316±5	97±2	365±1	122±1	328±1	147±1	446±0.4	155±0.3
EpiDiff	1096±5	374±1	1232±5	422±16	**188**±7	**62**±2	**247**±2	**86**±1	**304**±1	**82**±0.4	**423**±1	**125**±0.1

et al., 2021), GNN-based model **EpiColaGNN** (Liu et al., 2023), and a Neural ODE-based model **EARTH** (Wan et al., 2025).

4.2. Performance

This section addresses **Q1**. To demonstrate the performance of our proposed EpiDiff, we conducted comprehensive experiments on various epidemic datasets. We considered multiple baselines as detailed in Tab.1. Key observations include: ***(i)*** classical time series models (AR, ARMA, LSTM, GRU) perform poorly, especially for non-stationary or long-term epidemic dynamics; ***(ii)*** spatiotemporal graph models (STGCN, ColaGNN, DiffSTG) and ODE-based methods (STGODE, EARTH) provide moderate improvements but still struggle to adapt to distribution shifts; ***(iii)*** epidemic-specific neural models (STAN, EpiColaGNN) achieve better robustness, but may lack effective uncertainty modeling; ***(iv)*** EpiDiff consistently outperforms nearly all baselines across datasets and horizons, achieving the lowest or near-lowest RMSE and MAE, and demonstrates strong robustness, especially under distribution shifts such as those seen in ***covid-JP*** and ***influenza-US***. These results highlight the superior accuracy and stability of our approach for real-world epidemic forecasting.

4.3. Effectiveness

This section addresses **Q2**. Tab.2 (Row 1) shows mechanistic point estimation (Section 3.1). Row 2 reports the diffusion backbone with standard DDIM (no steering). Row 3 uses only point estimate steering, while the final row shows our full model.

On ***covid-US***, a stationary dataset, the accuracy of EpiDiff primarily comes from the strong fitting ability of the diffusion backbone. Since parameter estimation is difficult for non-trending data, uncertainty-steered inference performs slightly worse, but remains very close to direct diffusion inference. In contrast, on ***covid-JP*** (positive non-stationarity) and ***influenza-US*** (negative non-stationarity, both see Appendix C.1), the diffusion backbone alone struggles, while uncertainty-steered inference leads to a substantial improvement, outperforming both the mechanistic model and the diffusion backbone alone.

Fig.3 further illustrates the performance of different variants

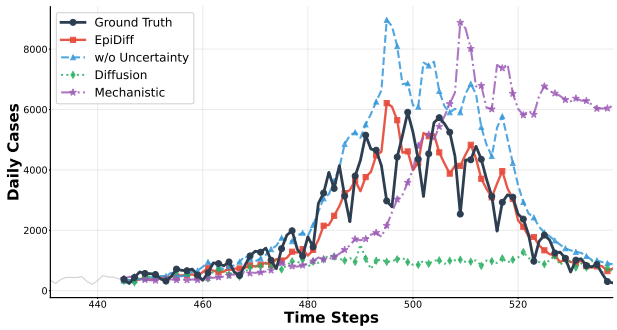


Figure 3. Full test set timeline comparison on Node 12 of the ***covid-JP*** dataset ($h = 14$).

across the full test set for Node 12 in the ***covid-JP*** dataset. We can see that EpiDiff not only captures the overall trend but also fits multiple outbreak peaks accurately, while other variants fail to track these dynamics as effectively.

4.4. Mechanistic Interpretation and Uncertainty

This section addresses Q3 on the mechanistic interpretation and uncertainty of EpiDiff. As shown in the left panel of Fig.4, for Node 30 in ***covid-JP*** at selected windows, the predictive uncertainty bands highlight where the mechanistic model is less confident. In these intervals with high uncertainty, the model's reliance shifts towards the data-driven diffusion backbone, since the mechanistic prior offers weaker guidance. The right panel of Fig.4 visualizes the uncertainty for all nodes and prediction windows as a heatmap. In general, the uncertainty grows with the prediction horizon, while some node–window regions show distinctly high uncertainty. Such visualization helps interpret when and where the model is uncertain, suggesting these high-uncertainty regions as priorities for more fine-grained data collection to further enhance epidemic forecasting. We further provide quantitative uncertainty evaluation using CRPS and empirical coverage in Appendix C.9; EpiDiff improves CRPS over the mechanistic baseline by about 22% across all datasets, although calibration remains challenging under simplified SIR dynamics.

4.5. Sensitivity

This section addresses **Q4** by investigating the sensitivity of EpiDiff to the guidance scale parameter τ in (13) (see Fig. 5). As the guidance scale approaches zero, the model behaves like the pure diffusion backbone, while larger val-

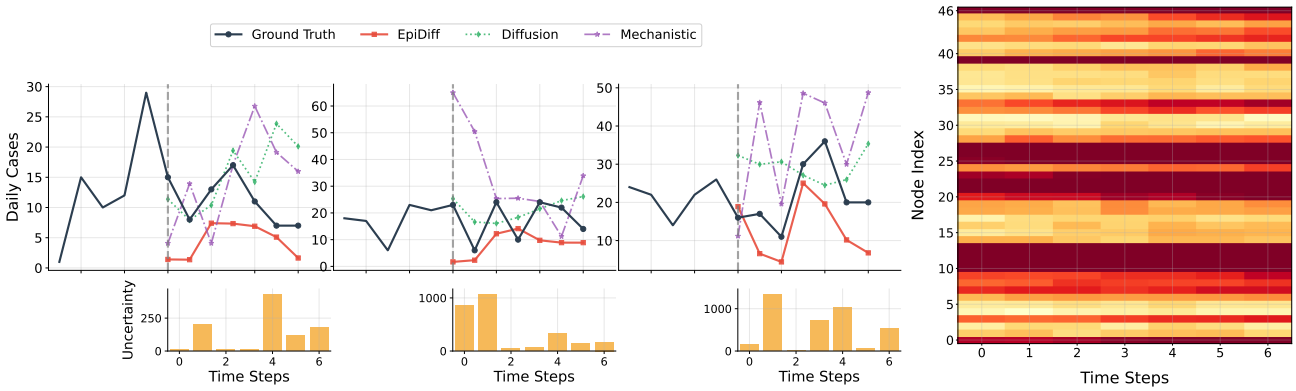
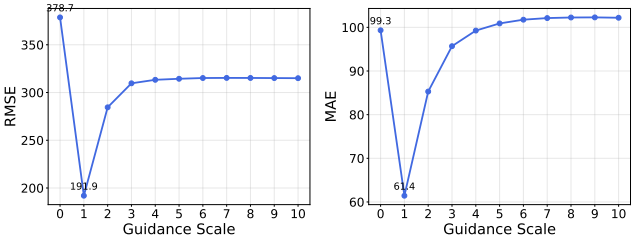
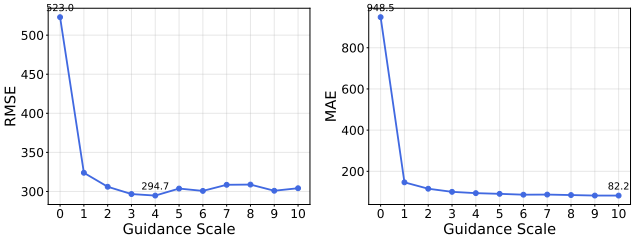


Figure 4. **Left:** Node 30 prediction and uncertainty at windows 30, 60, and 70 on ***covid-JP***. **Right:** Uncertainty heatmap for nodes and steps on ***covid-JP***. Darker squares indicate greater uncertainty.



(a) ***covid-JP***



(b) ***influenza-US***

Figure 5. RMSE and MAE with respect to guidance scale (h=7) on ***covid-JP*** and ***influenza-US***.

ues push the results toward mechanistic estimation. We observe that the best prediction accuracy is usually achieved at an intermediate guidance scale, highlighting the benefit of combining both data-driven and mechanistic signals. In practice, the optimal scale can be efficiently selected by grid search on the validation set prior to test-time inference.

5. Related Works

5.1. Mechanistic-Informed Epidemic Forecasting

Traditional epidemic modeling has progressed from purely mechanistic compartmental models (SIR, SEIR) to hybrid frameworks that combine epidemiological domain knowledge with neural networks (Wang et al., 2022; Song et al., 2020b; Tomy et al., 2022; Liu et al., 2023). Early methods mainly used compartmental models to synthesize training data for GNNs in source detection and surveillance (Shah et al., 2020; Ru et al., 2023; Meirom et al., 2021). More recent work explicitly incorporates mechanistic priors into neural architectures—e.g., CausalGNN (Wang et al., 2022) leverages SIRD states as causal features for spatiotemporal learning, and EARTH (Wan et al., 2025) employs SIR-inspired neural ODEs for continuous-time modeling. However, most methods lack explainable fusion mechanisms for mechanistic guidance. In addition, the noise in observations and time-varying parameters render hard-constraint approaches like PINNs (Karniadakis et al., 2021; Raissi et al., 2019) not suited for robust epidemic forecasting. These limitations highlight the need for flexible probabilistic frameworks that can adaptively incorporate mechanistic knowledge and uncertainty.

5.2. Inference-time Steering for Diffusion Models

In contrast to classifier-free guidance (CFG) (Ho & Salimans, 2022), which requires modifying the model architecture during training, inference-time steering methods intervene only during sampling, providing greater flexibility. A canonical example is classifier guidance (Dhariwal & Nichol, 2021), which alters the denoising trajectory via gradients from an external classifier. This gradient-based approach extends to generic differentiable rewards through methods like universal guidance (Bansal et al., 2023). Particle-based strategies such as SMC (Wu et al., 2023) broaden reward formats but often suffer from inefficiency. Recently, derivative-free approaches (Li et al., 2024) have been unified under the Feynman–Kac framework (Singhal et al., 2025), formalizing steering as sampling from tilted distributions. Building on this, we propose leveraging principled uncertainty quantification to robustly steer the diffusion process through the noisy dynamics inherent in epidemic forecasting.

6. Conclusion

In this paper, we introduce `EpiDiff`, a novel hybrid framework for epidemic forecasting that unifies mechanistic mod-

els and diffusion-based deep learning. By incorporating mechanistic priors and uncertainty quantification into the diffusion process, `EpiDiff` achieves robust and accurate predictions, even under non-stationary or out-of-distribution scenarios. Extensive experiments on multiple real-world datasets demonstrate the effectiveness and adaptability of our approach. Our work provides practical insights into bridging classical epidemic modeling with modern generative models, paving the way for future advances in epidemiological forecasting and decision-making.

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive comments. This work was supported in part by the NVIDIA Academic Grant Program and the National Science Foundation under Grant No. 2402438. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Impact Statement

This paper proposes `EpiDiff`, a diffusion-based spatiotemporal forecasting framework that integrates mechanistic epidemic models via uncertainty-aware guidance at inference time. Improved short-term forecasts of infectious disease trajectories may support public health planning (e.g., early warning and resource allocation), particularly under changing transmission dynamics.

Despite these potential benefits, forecasts can be misused or over-interpreted: inaccurate predictions may trigger delayed or unnecessary interventions and lead to suboptimal or inequitable allocation of limited resources. In addition, surveillance data often suffer from reporting delays, underreporting, and demographic/geographic biases; graph construction and proxy signals (e.g., mobility) may further encode structural disparities. Our mechanistic component may be misspecified for some pathogens or settings, and the uncertainty estimates are approximate, which could still bias guided sampling.

We therefore stress that `EpiDiff` should not be used as a stand-alone decision tool. Responsible use requires reporting uncertainty (e.g., prediction intervals), auditing errors across regions and subpopulations, stress-testing under plausible shift scenarios (policy changes, reporting disruptions), and clearly documenting data sources and modeling assumptions. Any real-world deployment should comply with applicable privacy and governance requirements.

References

Anderson, R. *Infectious diseases of humans: dynamics and control*. Oxford University Press, 1991.

Bansal, A., Chu, H.-M., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., and Goldstein, T. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 843–852, 2023.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

Brockwell, P. J. and Davis, R. A. *Introduction to time series and forecasting*. Springer, 2002.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. *Handbook of markov chain monte carlo*. CRC press, 2011.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing*, 16(5):1190–1208, 1995.

Cairns, C. and Kang, K. National hospital ambulatory medical care survey: 2019 emergency department summary tables. 2022.

Castro, R., Santos, R., Sousa, G., Pinheiro, Y., Martins, R., Pereira, M., and Silva, R. Spatial dynamics of the covid-19 pandemic in brazil. *Epidemiology & Infection*, 149: e60, 2021.

Chang, S., Pierson, E., Koh, P. W., Gerardin, J., Redbird, B., Grusky, D., and Leskovec, J. Mobility network models of covid-19 explain inequities and inform reopening. *Nature*, 589(7840):82–87, 2021.

Chimmula, V. K. R. and Zhang, L. Time series forecasting of covid-19 transmission in canada using lstm networks. *Chaos, solitons & fractals*, 135:109864, 2020.

Chinazzi, M., Davis, J. T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., Pastore y Piontti, A., Mu, K., Rossi, L., Sun, K., et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (covid-19) outbreak. *Science*, 368(6489):395–400, 2020.

Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

Chowell, G., Luo, R., Sun, K., Roosa, K., Tariq, A., and Viboud, C. Real-time forecasting of epidemic trajectories using computational dynamic ensembles. *Epidemics*, 30: 100379, 2020.

Contreras, J., Espinola, R., Nogales, F. J., and Conejo, A. J. Arima models to predict next-day electricity prices. *IEEE transactions on power systems*, 18(3):1014–1020, 2003.

Cramer, E. Y., Ray, E. L., Lopez, V. K., Bracher, J., Brennen, A., Castro Rivadeneira, A. J., Gerding, A., Gneiting, T., House, K. H., Huang, Y., et al. Evaluation of individual and ensemble probabilistic forecasts of covid-19 mortality in the united states. *Proceedings of the National Academy of Sciences*, 119(15):e2113561119, 2022.

Delussu, F., Tizzoni, M., and Gauvin, L. The limits of human mobility traces to predict the spread of covid-19: A transfer entropy approach. *PNAS nexus*, 2(10):pgad302, 2023.

Deng, S., Wang, S., Rangwala, H., Wang, L., and 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*, pp. 245–254, 2020.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Fang, Z., Long, Q., Song, G., and Xie, K. Spatial-temporal graph ode networks for traffic flow forecasting. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 364–373, 2021.

Flaxman, S., Mishra, S., Gandy, A., Unwin, H. J. T., Mellan, T. A., Coupland, H., Whittaker, C., Zhu, H., Berah, T., Eaton, J. W., et al. Estimating the effects of non-pharmaceutical interventions on covid-19 in europe. *Nature*, 584(7820):257–261, 2020.

Gao, J., Sharma, R., Qian, C., Glass, L. M., Spaeder, J., Romberg, J., Sun, J., and Xiao, C. Stan: spatio-temporal attention network for pandemic prediction using real-world evidence. *Journal of the American Medical Informatics Association*, 28(4):733–743, 2021.

Haug, N., Geyrhofer, L., Londei, A., Dervic, E., Desvars-Larrive, A., Loreto, V., Pinior, B., Thurner, S., and Klimek, P. Ranking the effectiveness of worldwide covid-19 government interventions. *Nature human behaviour*, 4(12):1303–1312, 2020.

He, Y., Chen, C., Wang, S., Min, G., and Li, J. Demystify epidemic containment in directed networks: Theory and algorithms. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pp. 905–913, 2025.

Hethcote, H. W. The mathematics of infectious diseases. *SIAM review*, 42(4):599–653, 2000.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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

Holmdahl, I. and Buckee, C. Wrong but useful—what covid-19 epidemiologic models can and cannot tell us. *New England Journal of Medicine*, 383(4):303–305, 2020.

Ioannidis, J. P., Cripps, S., and Tanner, M. A. Forecasting for covid-19 has failed. *International journal of forecasting*, 38(2):423–438, 2022.

Jewell, N. P., Lewnard, J. A., and Jewell, B. L. Predictive mathematical models of the covid-19 pandemic: underlying principles and value of projections. *Jama*, 323(19): 1893–1894, 2020.

Julier, S. J. and Uhlmann, J. K. New extension of the kalman filter to nonlinear systems. In *Signal processing, sensor fusion, and target recognition VI*, volume 3068, pp. 182–193. Spie, 1997.

Kalman, R. E. A new approach to linear filtering and prediction problems. 1960.

Kapoor, A., Ben, X., Liu, L., Perozzi, B., Barnes, M., Blais, M., and O'Banion, S. Examining covid-19 forecasting using spatio-temporal graph neural networks. *arXiv preprint arXiv:2007.03113*, 2020.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., and Yang, L. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

Keeling, M. J. and Rohani, P. *Modeling infectious diseases in humans and animals*. Princeton university press, 2008.

Kermack, W. O. and McKendrick, A. G. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

Li, X., Zhao, Y., Wang, C., Scalia, G., Eraslan, G., Nair, S., Biancalani, T., Ji, S., Regev, A., Levine, S., et al. Derivative-free guidance in continuous and discrete diffusion models with soft value-based decoding. *arXiv preprint arXiv:2408.08252*, 2024.

Liu, M., Liu, Y., and Liu, J. Epidemiology-aware deep learning for infectious disease dynamics prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 4084–4088, 2023.

Liu, Z., Wan, G., Prakash, B. A., Lau, M. S., and Jin, W. A review of graph neural networks in epidemic modeling. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 6577–6587, 2024.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328, 2003.

Meirom, E., Maron, H., Mannor, S., and Chechik, G. Controlling graph dynamics with reinforcement learning and graph neural networks. In *International Conference on Machine Learning*, pp. 7565–7577. PMLR, 2021.

Min, G., He, Y., and Chen, C. Scaling epidemic inference on contact networks: Theory and algorithms. *Advances in Neural Information Processing Systems*, 38:13700–13735, 2025.

Neal, R. M. et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

Pastor-Satorras, R., Castellano, C., Van Mieghem, P., and Vespignani, A. Epidemic processes in complex networks. *Reviews of modern physics*, 87(3):925–979, 2015.

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

Reed, C., Angulo, F. J., Swerdlow, D. L., Lipsitch, M., Meltzer, M. I., Jernigan, D., and Finelli, L. Estimates of the prevalence of pandemic (h1n1) 2009, united states, april–july 2009. *Emerging infectious diseases*, 15(12): 2004, 2009.

Reich, N. G., Brooks, L. C., Fox, S. J., Kandula, S., McGowan, C. J., Moore, E., Osthus, D., Ray, E. L., Tushar, A., Yamana, T. K., et al. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the united states. *Proceedings of the National Academy of Sciences*, 116(8):3146–3154, 2019.

Ru, X., Moore, J. M., Zhang, X.-Y., Zeng, Y., and Yan, G. Inferring patient zero on temporal networks via graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9632–9640, 2023.

Shah, C., Dehmamy, N., Perra, N., Chinazzi, M., Barabási, A.-L., Vespignani, A., and Yu, R. Finding patient zero: Learning contagion source with graph neural networks. *arXiv preprint arXiv:2006.11913*, 2020.

Singhal, R., Horvitz, Z., Teehan, R., Ren, M., Yu, Z., McKeown, K., and Ranganath, R. A general framework for inference-time scaling and steering of diffusion models. *arXiv preprint arXiv:2501.06848*, 2025.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Song, S., Zong, Z., Li, Y., Liu, X., and Yu, Y. Reinforced epidemic control: Saving both lives and economy. *arXiv preprint arXiv:2008.01257*, 2020b.

Stocks, T., Britton, T., and Höhle, M. Model selection and parameter estimation for dynamic epidemic models via iterated filtering: application to rotavirus in germany. *Biostatistics*, 21(3):400–416, 2020.

Takeuchi, J. S., Yamamoto, K., Kurokawa, M., Fukano, K., Kamikawa, A., Hatano, E., Takayanagi-Nishisako, S., Motohashi, A., Takamatsu, Y., Mitsuya, H., et al. Large-scale screening of sars-cov-2 variants in tokyo, japan: A 3-year and 9-month longitudinal survey. *Global Health & Medicine*, 7(2):151–160, 2025.

Tomy, A., Razzanelli, M., Di Lauro, F., Rus, D., and Della Santina, C. Estimating the state of epidemics spreading with graph neural networks. *Nonlinear Dynamics*, 109(1):249–263, 2022.

Wan, E. A. and Van Der Merwe, R. The unscented kalman filter for nonlinear estimation. In *Proceedings of the IEEE 2000 adaptive systems for signal processing, communications, and control symposium (Cat. No. 00EX373)*, pp. 153–158. Ieee, 2000.

Wan, G., Liu, Z., Shan, X., Lau, M. S., Prakash, B. A., and Jin, W. Earth: Epidemiology-aware neural ode with continuous disease transmission graph. In *Forty-second International Conference on Machine Learning*, 2025.

Wang, L., Adiga, A., Chen, J., Sadilek, A., Venkatramanan, S., and Marathe, M. Causalgnn: Causal-based graph neural networks for spatio-temporal epidemic forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 12191–12199, 2022.

Wang, S., Cao, J., and Philip, S. Y. Deep learning for spatio-temporal data mining: A survey. *IEEE transactions on knowledge and data engineering*, 34(8):3681–3700, 2020.

Wang, S., Pan, X., Ruan, S., Han, H., Wang, Z., Yuan, H., Zhu, J., and Li, Q. Diffcrime: A multimodal conditional diffusion model for crime risk map inference. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3212–3221, 2024.

Wen, H., Lin, Y., Xia, Y., Wan, H., Wen, Q., Zimmermann, R., and Liang, Y. Diffstg: Probabilistic spatio-temporal graph forecasting with denoising diffusion models. In *Proceedings of the 31st ACM international conference on advances in geographic information systems*, pp. 1–12, 2023.

Wu, L., Trippe, B., Naesseth, C., Blei, D., and Cunningham, J. P. Practical and asymptotically exact conditional sampling in diffusion models. *Advances in Neural Information Processing Systems*, 36:31372–31403, 2023.

Yu, B., Yin, H., and Zhu, Z. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*, 2017.

Zheng, Y., Zhong, L., Wang, S., Yang, Y., Gu, W., Zhang, J., and Wang, J. Diffuflow: Robust fine-grained urban flow inference with denoising diffusion model. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 3505–3513, 2023.

A. Theoretical Proof

A.1. Proof of Theorem 3.2

Proof. Let $\theta \sim \mathcal{N}(\hat{\theta}, \Sigma)$ and denote the centered variable $\delta = \theta - \hat{\theta}$. For any twice continuously differentiable function $g : \mathbb{R}^d \to \mathbb{R}$, we expand g around $\hat{\theta}$ using a second-order Taylor expansion,

$$g(\hat{\theta} + \delta) = g(\hat{\theta}) + \nabla g(\hat{\theta})^\top \delta + \tfrac{1}{2}\, \delta^\top \nabla^2 g(\hat{\theta})\, \delta + R_3(\delta), \tag{14}$$

where the remainder $R_3(\delta)$ is bounded in norm by a constant times $\|\delta\|^3$. Taking expectation under the Gaussian distribution gives

$$\mathbb{E}[g(\theta)] = g(\hat{\theta}) + \tfrac{1}{2}\, \mathrm{tr}\big(\nabla^2 g(\hat{\theta})\, \Sigma\big) + \mathbb{E}[R_3(\delta)]. \tag{15}$$

The nonlinear transform constructs $2d{+}1$ support points $\{\theta^{(m)}\}$ with associated weights $\{w^{(m)}\}$ such that the discrete distribution exactly matches the first two moments of the Gaussian: the weights sum to one, the weighted deviations from $\hat{\theta}$ vanish, and the weighted outer products reproduce Σ. Propagating these support points through g gives the approximation

$$\widehat{\mathbb{E}}[g(\theta)] = \sum_m w^{(m)} g(\theta^{(m)}). \tag{16}$$

Substituting the Taylor expansion into this expression and using the moment-matching properties shows that the constant and quadratic terms are reproduced exactly, while the linear term cancels. The only discrepancy comes from the remainder terms $R_3(\theta^{(m)} - \hat{\theta})$, which are of order $\|\theta - \hat{\theta}\|^3$. Thus, the estimated mean matches the true mean up to third-order error.

A similar argument applies to the variance. Since g truncated at second order is a quadratic form in δ, and the support points are constructed to match the Gaussian mean and covariance, the variance of the quadratic approximation is captured exactly by the support point rule. The difference between the true variance and the approximate variance again arises only through the cubic remainder, and is therefore of order $\mathcal{O}(\|\theta - \hat{\theta}\|^3)$.

In summary, the nonlinear transform with $2d{+}1$ support points reproduces the first two moments of $g(\theta)$ exactly up to second order, with errors only in higher-order terms. This establishes the second-order accuracy of the propagation scheme. □

B. Details of Algorithms

B.1. Nonlinear Transform

The nonlinear transform, also known as the *unscented transform* (Julier & Uhlmann, 1997; Wan & Van Der Merwe, 2000), constructs a deterministic set of $2d{+}1$ weighted *support points* that approximates the Gaussian distribution $\mathcal{N}(\hat{\theta}, \Sigma)$ up to second order. By propagating these support points through a nonlinear mapping $g(\cdot)$, the first two moments of $g(\theta)$ can be recovered with second-order accuracy, while avoiding the need for explicit numerical integration.

While the nonlinear transform is traditionally employed for nonlinear filtering and state estimation, we adopt it in a different setting. In our framework, the support points are not used to approximate hidden state dynamics, but instead to propagate uncertainty in the mechanistic parameters of epidemic models, e.g., infection and recovery rates. This allows us to quantify how estimation uncertainty in model parameters translates into predictive uncertainty over the epidemic trajectories, and to incorporate this uncertainty directly into the steering of diffusion inference. This application of the nonlinear transform for *uncertainty-aware mechanistic guidance* is, to our knowledge, novel.

Remark. In the SIR dynamics considered in this paper, the parameter dimension is $d = 2$, so only $2d{+}1 = 5$ support points are required per node.

B.2. PatchSIR

We summarize our method of mechanistic SIR estimation as a robust algorithm **PatchSIR**.

To ensure reliable estimates and uncertainties even under numerical instabilities, we design the following fallback algorithm.

B.3. Steered Diffusion Inference

We present the detailed inference algorithm corresponding to Section 3.1. Our procedure follows the standard deterministic DDIM sampling scheme (Song et al., 2020a), while incorporating an uncertainty-weighted mechanistic prior at each

Algorithm 1 Nonlinear Transform for Support Point Construction

Input: Mean $\hat{\theta} \in \mathbb{R}^d$, covariance $\Sigma \in \mathbb{R}^{d\times d}$, scalars $\alpha > 0$ and ν.
Output: Support points $\{\theta^{(m)}\}_{m=0}^{2d}$ and weights $\{w^{(m)}\}_{m=0}^{2d}$.

1. Set $\lambda \leftarrow \alpha^2(d+\nu) - d$.
2. Compute the Cholesky factor L of Σ such that $\Sigma = LL^\top$.
3. Set $S \leftarrow \sqrt{d+\lambda}\, L$.
4. Set $\theta^{(0)} \leftarrow \hat{\theta}$ and $w^{(0)} \leftarrow \lambda/(d+\lambda)$.
5. For $i = 1, \ldots, d$:
   (a) Let e_i be the i-th standard basis vector in $\mathbb{R}^d$.
   (b) Set $\theta^{(i)} \leftarrow \hat{\theta} + Se_i$ and $\theta^{(i+d)} \leftarrow \hat{\theta} - Se_i$.
   (c) Set $w^{(i)} \leftarrow 1/[2(d+\lambda)]$ and $w^{(i+d)} \leftarrow 1/[2(d+\lambda)]$.
6. **return** $\{\theta^{(m)}, w^{(m)}\}_{m=0}^{2d}$.

Algorithm 2 PatchSIR

Input: History y_{hist}, horizon T_p, SIR states sir_{hist}, priors, dispersion α_{NB}, constraints $\mathcal{B}$.
Output: Forecast $\hat{\mathbf{y}}$ and uncertainty $\mathbf{u}$.

1. Initialize normalized SIR states x_0^{hist} and x_0^{fore}; set rolling history $\tilde{y} \leftarrow y_{\text{hist}}$.
2. For $t = 1, \ldots, T_p$:
   (a) Estimate MAP parameters $\hat{\eta}$ and covariance Σ_η using Laplace approximation.
   (b) Generate support points $\{(\eta^{(m)}, w^{(m)})\}$ by the nonlinear transform.
   (c) Calibrate scale s from the last observed incidence and the model incidence.
   (d) Simulate one-step incidence $z^{(m)}$ for each support point.
   (e) If samples are invalid or extreme, set
   $$(\hat{y}_t, u_t) \leftarrow \textsc{FallbackFromHistMean}(\tilde{y}, 1, \alpha_{\text{NB}}).$$
   (f) Otherwise, aggregate with weighted geometric mean to obtain $\hat{y}_t$, and compute variance u_t.
   (g) Commit $(\hat{y}_t, u_t)$ and update $\tilde{y} \leftarrow \tilde{y} \cup \{\hat{y}_t\}$.
3. **return** $(\hat{\mathbf{y}}, \mathbf{u})$.

denoising step. Concretely, the clean trajectory $\hat{\mathbf{Y}}_0^{t+1:t+h}$ predicted from the noisy state $\mathbf{Y}_s^{t+1:t+h}$ is adjusted toward the mechanistic estimate $\mathbf{Y}^{\text{M}}$ with strength modulated by the covariance $\mathbf{\Sigma}_{\text{M}}$ and guidance scale τ. The full algorithm is given below.

C. Details of Experiments

Infrastructure and Implementation. All experiments were conducted on a server equipped with NVIDIA A100 GPUs (80GB memory) and an AMD EPYC 7473X CPU with 48 cores and 503GB RAM. Our implementation is based on PyTorch 2.4 with CUDA 12.1.

Hyperparameter Settings. For all experiments, we set the historical window length to $T_h = 14$ days. The model was trained with a batch size of 8 for 300 epochs using the Adam optimizer with learning rate 0.002 and weight decay 1×10^{-5}. Early stopping was applied with a patience of 10 epochs. The diffusion process followed a quadratic noise schedule with $\beta_{\text{end}} = 0.1$ and 200 sampling steps under the DDPM strategy.

Our backbone is the STGTransformer with 4 layers, hidden dimension $d_h = 32$, 4 attention heads, dropout rate 0.2, and channel multipliers $[1, 2]$. Spatial embeddings of size 32 were learned for each nodes. The input is first projected with a 1×1 convolution, followed by adaptive layer normalization, multi-head attention, and feed-forward networks in each block. The total number of learnable parameters is $75{,}549$. At inference time, we adopt deterministic DDIM sampling with 40 steps, which provides efficient generation while maintaining comparable accuracy to the longer ancestral DDPM trajectories.

Algorithm 3 FallbackFromHistMean

Input: Rolling history $\tilde{y}$, horizon $h = 1$, dispersion α_{NB}.
Output: Fallback forecast $\hat{y}^{\mathrm{fb}}$ and uncertainty u^{fb}.

1. Set $\bar{y} \leftarrow \max(\mathrm{mean}(\tilde{y}), 0)$.
2. Set $\hat{y}^{\mathrm{fb}} \leftarrow \bar{y}$.
3. Set $u^{\mathrm{fb}} \leftarrow \bar{y} + \bar{y}^2/\alpha_{\mathrm{NB}}$.
4. **return** $(\hat{y}^{\mathrm{fb}}, u^{\mathrm{fb}})$.

Algorithm 4 Inference Stage of `EpiDiff`

Input: Trained noise predictor $\hat{\varepsilon}_\psi(\cdot, s)$; schedule $\{\bar{\alpha}_s\}_{s=1}^{S}$; mechanistic estimate $\mathbf{Y}^{\mathrm{M}}$; diagonal covariance $\mathbf{\Sigma}_{\mathrm{M}}$; guidance scale $\tau > 0$.
Output: Forecast $\mathbf{Y}_0^{t+1:t+h}$.

1. Initialize $\mathbf{Y}_S^{t+1:t+h} \sim \mathcal{N}(0, I)$.
2. For $s = S, \ldots, 1$:
   (a) Predict noise:
   $$\hat{\varepsilon}_\psi \leftarrow \hat{\varepsilon}_\psi(\mathbf{Y}_s^{t+1:t+h}, s).$$
   (b) Reconstruct the clean trajectory:
   $$\hat{\mathbf{Y}}_0^{t+1:t+h} \leftarrow \frac{1}{\sqrt{\bar{\alpha}_s}} \left(\mathbf{Y}_s^{t+1:t+h} - \sqrt{1-\bar{\alpha}_s}\, \hat{\varepsilon}_\psi \right).$$
   (c) Apply mechanistic steering:
   $$\hat{\mathbf{Y}}_0^{t+1:t+h} \leftarrow \hat{\mathbf{Y}}_0^{t+1:t+h} - \tau\, \mathbf{\Sigma}_{\mathrm{M}}^{-1} \left(\hat{\mathbf{Y}}_0^{t+1:t+h} - \mathbf{Y}^{\mathrm{M}} \right).$$
   (d) Re-encode:
   $$\hat{\varepsilon} \leftarrow \frac{\mathbf{Y}_s^{t+1:t+h} - \sqrt{\bar{\alpha}_s}\, \hat{\mathbf{Y}}_0^{t+1:t+h}}{\sqrt{1-\bar{\alpha}_s}}.$$
   (e) Perform the DDIM update:
   $$\mathbf{Y}_{s-1}^{t+1:t+h} \leftarrow \sqrt{\bar{\alpha}_{s-1}}\, \hat{\mathbf{Y}}_0^{t+1:t+h} + \sqrt{1-\bar{\alpha}_{s-1}}\, \hat{\varepsilon}.$$
3. **return** $\mathbf{Y}_0^{t+1:t+h}$.

C.1. Details of Dataset

In our experiments, the three datasets represent distinct scenarios commonly encountered in epidemic forecasting. Visualizations of these datasets are shown in Fig. 6. The ***covid-US*** dataset exhibits a relatively stable distribution, with the test set remaining within the range of the training data. In contrast, ***covid-JP*** features a significant outbreak in the test period that exceeds anything observed during training, presenting a case of positive distribution shift (non-stationarity). Meanwhile, in ***influenza-US***, many influenza cases were reclassified as COVID-19 during the pandemic, resulting in lower estimated values in the test set and thus a negative distribution shift. Through extensive experiments and comprehensive comparisons with state-of-the-art baselines, we demonstrate the effectiveness of the proposed `EpiDiff` framework.

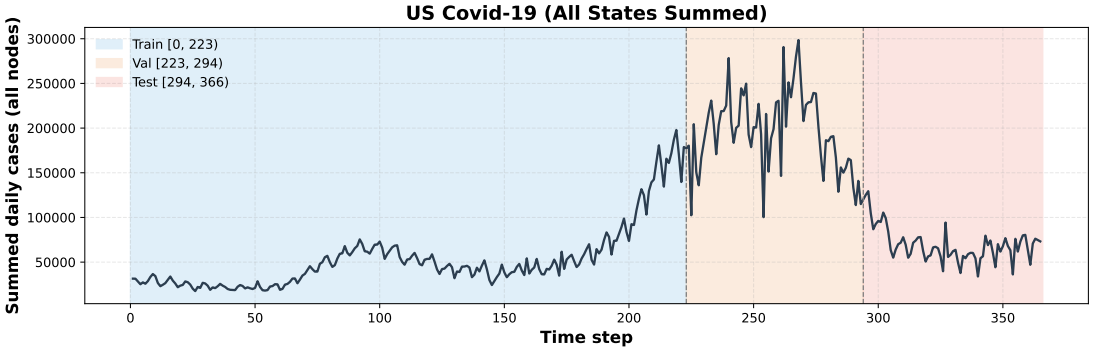
C.1.1. *covid-US*

Fig. 6a shows the temporal evolution of total daily cases (aggregated over all nodes) in the ***covid-US*** dataset from April 15, 2020 to April 15, 2021. Shaded regions indicate the data splits used in our experiments: **train** `[0, 262)`, **validation** `[262, 313)`, and **test** `[313, 366)`. The data is from daily case report from The New York Times [1].

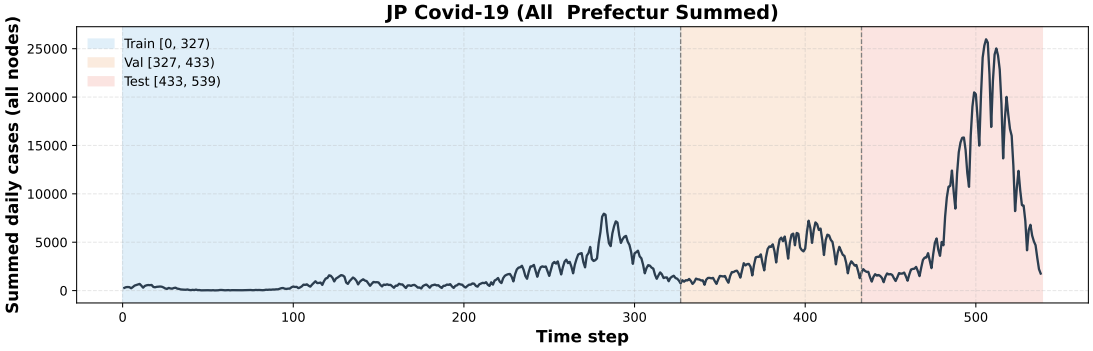
C.1.2. *covid-JP*

Trend of daily confirmed COVID-19 cases in ***covid-JP*** is shown in Fig. 6b. Daily cases are aggregated over all prefectures from April 1, 2020 to September 21, 2021. The series is partitioned into **train** `[0, 383)`, **validation** `[383, 460)`, and

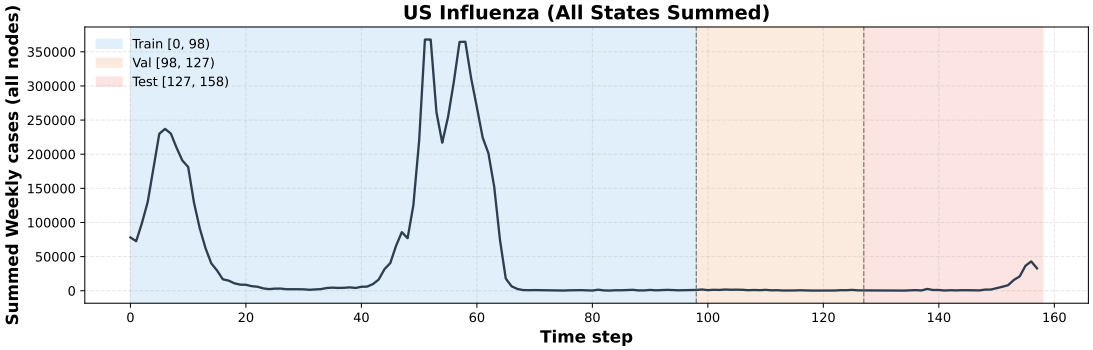
[1] https://github.com/nytimes/covid-19-data



(a) ***covid-US*** case trends.



(b) ***covid-JP*** case trends.



(c) ***influenza-US*** case trends.

Figure 6. Visualization of case trends in the three datasets: ***covid-JP***, ***covid-US***, and ***influenza-US***.

test `[460, 539)`. The data is from the NHK COVID-19 database [2].

C.1.3. *influenza-US*

Weekly estimated cases are aggregated over all states from Week 1, 2019 to the last week of 2021. The series is partitioned into **train** `[0, 98)`, **validation** `[98, 127)`, and **test** `[127, 158)`.

Estimation Assumptions. We derive state-level influenza infections from ILINet reports[3] under the following assumptions: (1) We use the *percent unweighted ILI* as the weekly surveillance indicator, since the weighted version is not consistently available at the state level. (2) State populations are fixed to 2020 U.S. Census estimates and treated as constant over

[2] https://github.com/swsoyee/2019-ncov-japan

[3] https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html

2019–2021. (3) Each individual seeks outpatient care on average $\mu \approx 3$ times per year ($\mu \approx 0.06$ visits per week) (Cairns & Kang, 2022). (4) Among symptomatic individuals, the probability of seeking medical care is $p_{\text{seek}} \approx 0.45$ (Reed et al., 2009).

Estimation Formula. Let $\text{ILI}_{s,t}$ denote the unweighted ILI proportion in state s during week t, and let pop_s denote the population of state s. The estimated number of infections is then:

$$\widehat{\text{Infections}}_{s,t} = \frac{\text{ILI}_{s,t} \times \mu \times \text{pop}_s}{p_{\text{seek}}}. \tag{17}$$

Post-processing. We round $\widehat{\text{Infections}}_{s,t}$ to the nearest integer, interpolate occasional missing values linearly, and forward-fill at the boundaries.

C.1.4. Population Data

U.S. state population data is sourced from the "Annual Estimates of the Resident Population: April 1, 2020 to July 1, 2021 (CO-EST2021-alldata)" published by the U.S. Census Bureau, Population Division (March 24, 2022) [4]. Japanese prefecture population data is obtained from the Wikipedia entry "List of Japanese prefectures by population." [5]

C.2. Details of Metrics

We adopt two commonly used metrics to evaluate forecasting performance: root mean squared error ($\mathcal{R}$, RMSE) and mean absolute error ($\mathcal{P}$, MAE). RMSE measures the square root of the average squared deviation between predictions and the ground truth. By squaring the errors, RMSE penalizes large deviations more heavily than small ones. This property makes RMSE particularly sensitive to outliers or sudden spikes in epidemic case counts. In epidemic forecasting, this is useful for highlighting whether a model can capture abrupt outbreak surges. However, the squaring operation may exaggerate the impact of a few extreme errors on the overall evaluation. MAE computes the average absolute deviation between predicted and true values. Unlike RMSE, MAE treats all errors linearly and equally, without disproportionately penalizing large errors. Thus, MAE is more robust to outliers and provides a more balanced assessment of typical prediction accuracy. In epidemic settings, MAE directly reflects the average number of cases by which the model is wrong, making it easier to interpret in practical terms (e.g., "on average, predictions deviate by x cases per day").

Let $\{y_t\}_{t=1}^{T}$ denote the ground-truth sequence of epidemic values and $\{\hat{y}_t\}_{t=1}^{T}$ denote the predicted sequence. The metrics are defined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}, \tag{18}$$

$$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T}|y_t - \hat{y}_t|. \tag{19}$$

Both RMSE and MAE take values in $[0, +\infty)$, with smaller values indicating better predictive performance.

C.3. Details of Baselines

AR (Contreras et al., 2003): The Autoregressive (AR) model is a fundamental time series baseline that predicts future values based on a linear combination of past observations. Its simplicity and effectiveness make it a standard benchmark for evaluating more complex forecasting methods.

ARMA (Contreras et al., 2003): The Autoregressive Moving Average (ARMA) model combines autoregression with a moving average component, capturing both temporal dependence and short-term noise in time series data. It serves as a widely used baseline for forecasting due to its balance between simplicity and expressive power.

LSTM (Hochreiter & Schmidhuber, 1997): The Long Short-Term Memory (LSTM) network is a recurrent neural network architecture designed to capture long-range temporal dependencies by mitigating the vanishing gradient problem.

[4] https://www2.census.gov/programs-surveys/popest/datasets/2020-2021/counties/totals/co-est2021-alldata.csv
[5] https://en.wikipedia.org/wiki/List_of_Japanese_prefectures_by_population

GRU (Cho et al., 2014): The Gated Recurrent Unit (GRU) is a simplified variant of the LSTM that combines the forget and input gates into a single update gate, leading to fewer parameters and faster training while still effectively addressing the vanishing gradient problem.

STGCN (Yu et al., 2017): Spatio-Temporal Graph Convolutional Networks (STGCN) is a deep learning framework for traffic forecasting that integrates graph convolution with gated temporal convolution in a fully convolutional architecture. By formulating traffic flow as a graph-structured time series, STGCN effectively captures comprehensive spatial-temporal dependencies while maintaining fast training, fewer parameters, and easier convergence.

ColaGNN (Deng et al., 2020): ColaGNN is a cross-location attention based graph neural network designed for long-term influenza-like illness forecasting. It dynamically constructs adjacency matrices using attention scores to capture directed spatial effects, and applies multi-scale dilated convolutions to model both short- and long-term temporal patterns.

STGODE (Fang et al., 2021): STGODE is a spatial-temporal forecasting framework that leverages tensor-based ordinary differential equations to jointly capture spatial and temporal dynamics. By enabling deeper architectures, it effectively models long-range dependencies while alleviating the over-smoothing problem of shallow GNNs. STGODE incorporate both semantic adjacency information and dilated temporal convolutions.

DiffSTG (Wen et al., 2023): DiffSTG is a probabilistic spatio-temporal graph forecasting framework that generalizes denoising diffusion probabilistic models to STG data. It introduces UGnet, a denoising network that leverages a U-Net architecture for multi-scale temporal dependencies and a GNN for spatial correlations.

STAN (Gao et al., 2021): STAN is a spatio-temporal attention network that integrates real-world claims data, demographic similarity, and geographical proximity for pandemic forecasting. It employs a graph attention mechanism to capture complex spatiotemporal disease dynamics and incorporates a dynamics-based loss to improve long-term predictions.

EpiColaGNN (Liu et al., 2023): EpiColaGNN is an epidemiology-aware deep learning framework that incorporates the next-generation matrix into both model architecture and objective function. This design enables simultaneous modeling of within-location disease dynamics and cross-location transmission induced by human mobility.

EARTH (Wan et al., 2025): EARTH is an epidemiology-aware neural ODE framework designed for epidemic forecasting. It integrates mechanistic compartmental models with deep learning by modeling continuous local transmission dynamics and capturing global infection trends. A cross-attention fusion mechanism further combines global and local signals, enabling more accurate and robust epidemic predictions compared with SOTA methods.

C.4. Comparison of inference time

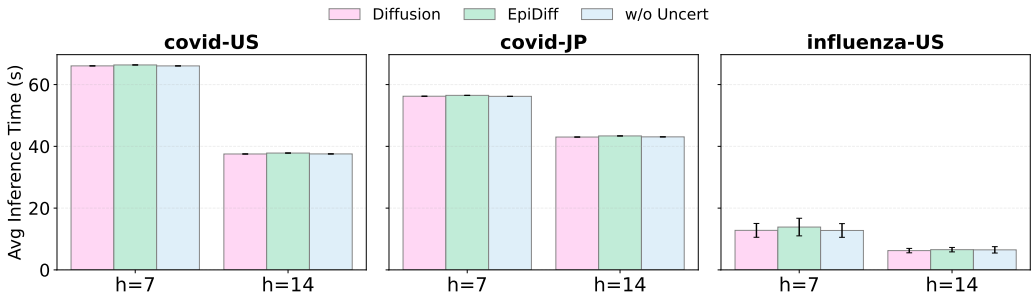


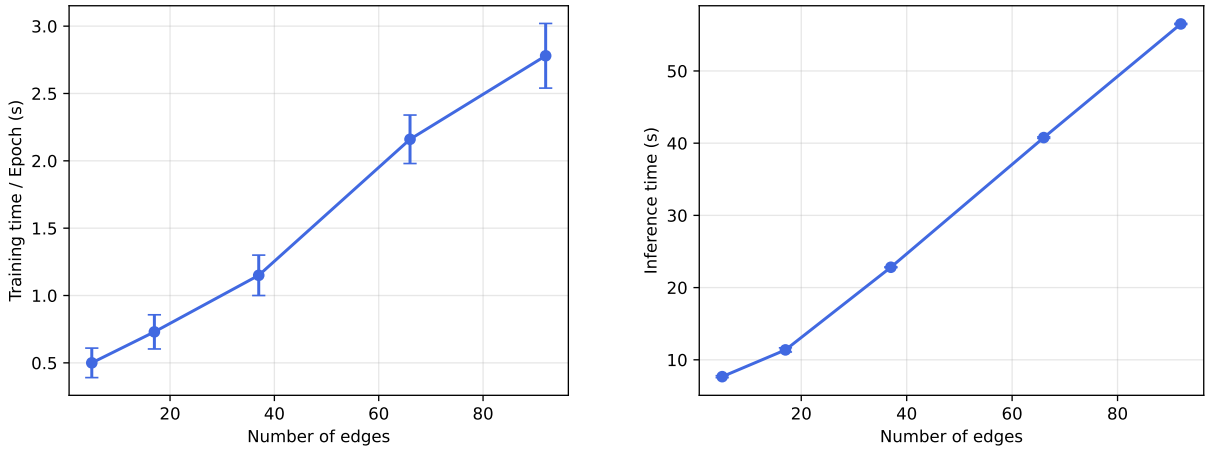
Figure 7. Average inference time (in seconds) for Diffusion baseline, `EpiDiff`, and `w/o` Uncertainty across all datasets and horizons ($h = 7, 14$).

We quantitatively compare the inference time (seconds per sample) of Diffusion, `EpiDiff`, and the w/o Uncertainty variant across three datasets—`covid-US`, `covid-JP`, and `influenza-US`—and two prediction horizons ($h = 7$ and $h = 14$) in Fig.7. All models were evaluated using a 40-step DDIM sampler.

For `covid-US` at $h = 7$, the mean inference times are 66.05 (Diffusion), 66.36 (`EpiDiff`), and 66.04 (w/o Uncert), with standard deviations on the order of 10^{-2}. At $h = 14$, the corresponding means are 37.53, 37.83, and 37.55. On `covid-JP`, the mean times are 56.23, 56.51, and 56.19 for $h = 7$; 43.01, 43.37, and 43.07 for $h = 14$. For `influenza-US`, the means are 12.80, 13.87, and 12.76 at $h = 7$; 6.24, 6.55, and 6.49 at $h = 14$. In all cases, the standard deviations remain low.

These results show that integrating the mechanistic steering and uncertainty estimation in `EpiDiff` leads to nearly identical inference times compared to the original diffusion baseline and the variant without uncertainty, regardless of the dataset or prediction window. The steering module is plug-and-play and does not noticeably affect inference efficiency when using a 40-step DDIM sampler.

C.5. Scalability Analysis



(a) Per-epoch training time

(b) Inference time

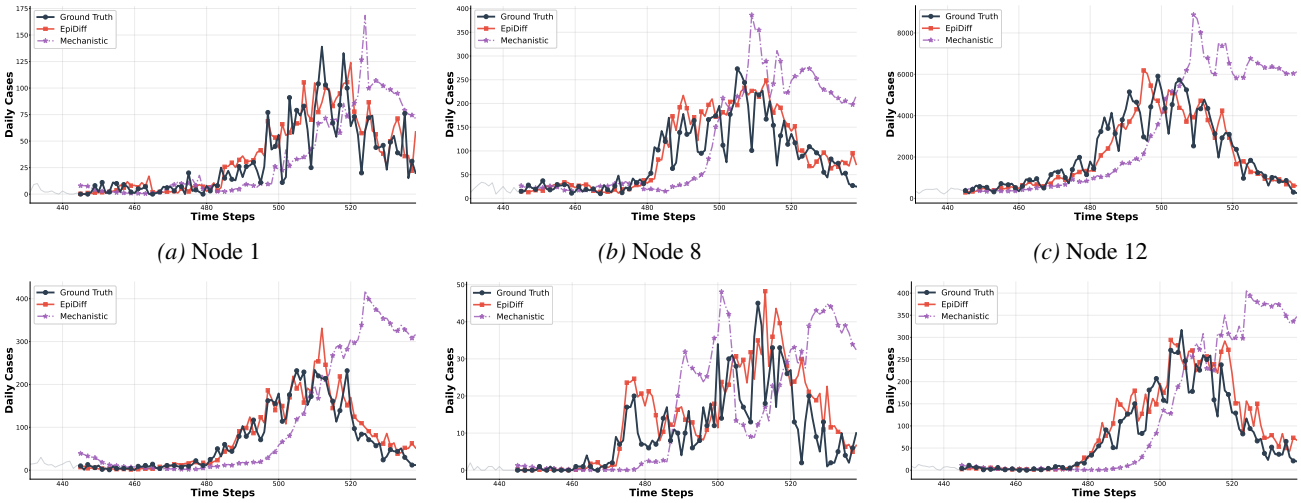
Figure 8. On the `covid-JP` dataset, per-epoch training time (left) and overall inference time (right) as we vary the number of nodes by randomly subsampling regions (prediction horizon $h = 7$).

Number of geographies N. The total runtime of `EpiDiff` consists of three parts: the time for mechanistic estimation and uncertainty quantification t_{mech}, the time for training the diffusion backbone t_{train}, and the final inference time $t_{\text{inference}}$. Since mechanistic estimation and uncertainty quantification are performed independently for each node and history window, we have $t_{\text{mech}} \propto N$. Consistent with this analysis, Figure 8a and Figure 8b show that both per-epoch training time and overall inference time increase roughly linearly as we subsample more regions. Overall, the runtime scales approximately linearly in the number of geographies.

Number of Mechanistic Parameters n. The additional cost introduced by the mechanistic component depends on the number of parameters in the chosen compartmental model. As discussed in Appendix C.4 (Figure 7), mechanistic steering itself adds almost no overhead to the diffusion backbone, so the dominant factor associated with increasing n is the computation of mechanistic estimation and uncertainty. For a mechanistic model with n parameters, the per-node, per-window cost of our Laplace-based estimator scales with the parameter dimension: approximately $\mathcal{O}(n^2)$ evaluations of the mechanistic objective are required to form the finite-difference Hessian, followed by $\mathcal{O}(n^3)$ time to invert it, on top of the forward simulation cost of the mechanistic model. For common compartmental families such as SEIR or SIRD, n remains small (typically a few parameters), so the resulting overhead is modest and parallelizable across regions. Overall, the runtime scaling with n is mild, and the diffusion backbone continues to dominate computation on larger graphs.

C.6. Mechanistic Analysis

Model performance is not overly sensitive to the exact accuracy of the mechanistic estimates, because their influence is explicitly modulated by the uncertainty term. Figure 9 shows six representative nodes on the `covid-JP` dataset (lookback window $\kappa = 14$, prediction horizon $h = 14$). in some regions (e.g., Node 1, Node 31) the SIR trajectories are relatively accurate, while in others (e.g., Node 12, Node 42) they deviate substantially and fail to capture local trends. In these regimes, EpiDiff still tracks the ground truth closely: when the mechanistic fit is poor, the posterior uncertainty becomes large, which automatically weakens the guidance and allows the diffusion backbone to correct the mismatch. Thus the mechanistic prior serves as a coarse but informative anchor rather than a brittle constraint, and EpiDiff remains robust to moderate inaccuracies in mechanistic estimation.



(a) Node 1 *(b)* Node 8 *(c)* Node 12

(d) Node 24 *(e)* Node 31 *(f)* Node 42

Figure 9. Six representative regions on the `covid-JP` dataset with lookback window $\kappa = 14$ and prediction horizon $h = 14$. For each node, we show daily cases over time for the ground truth, EpiDiff, and the SIR-based mechanistic model.

C.7. Practical Validation

Figure 10a shows the inferred nationwide effective reproduction factor $R_{\text{eff}}(t)$ for `covid-JP`, which remains within a moderate range without pathological spikes: it rises moderately above 1 during growth phases and falls to around or below 1 in declining phases, consistent with typical empirical estimates for COVID-19. Starting from approximately calendar week 25, $R_{\text{eff}}(t)$ exhibits a sharp increase above 1, indicating a period of rapid transmission; this surge coincides with the rapid rise in the proportion of the Delta variant in Japan shown in Figure 10b. Subsequently, $R_{\text{eff}}(t)$ gradually decreases back towards (and below) 1, which is consistent with the phase when Delta becomes dominant and more targeted vaccination efforts and public health interventions are implemented, suggesting that our inferred R_{eff} captures both variant-driven acceleration and policy-driven mitigation at the national level.

C.8. Sensitivity to forecast horizons and lookback windows

Figure 11 summarizes how the prediction error of `EpiDiff` evolves with the forecast horizon on `covid-JP`. The RMSE increases smoothly as h grows, which is expected as longer-term forecasts accumulate more uncertainty, but there is no sharp degradation. This suggests that the proposed uncertainty-aware guidance helps `EpiDiff` preserve reliable performance even for medium-range (up to 14 days ahead) predictions.

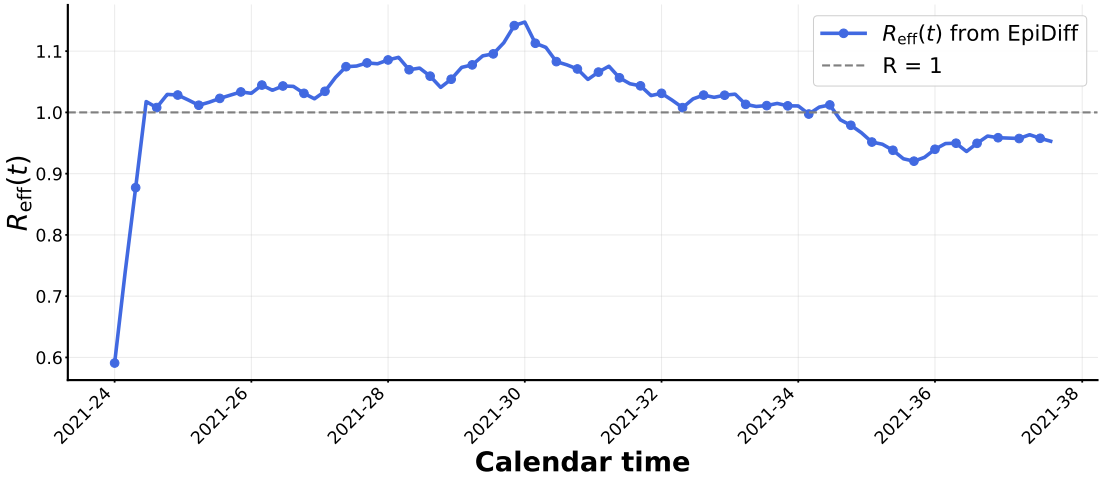
Figure 12 shows how the forecasting error of `EpiDiff` changes with different lookback windows T_h on `covid-JP`. Across $T_h \in \{8, 14, 22, 30, 36\}$, the MAE stays within a narrow range in the high 80s to low 90s, without any sharp degradation. This suggests that the proposed uncertainty-aware guidance yields a backbone that is robust to the specific choice of lookback window, rather than relying on a finely tuned history length.

C.9. Quantitative Evaluation of Mechanistic Uncertainty

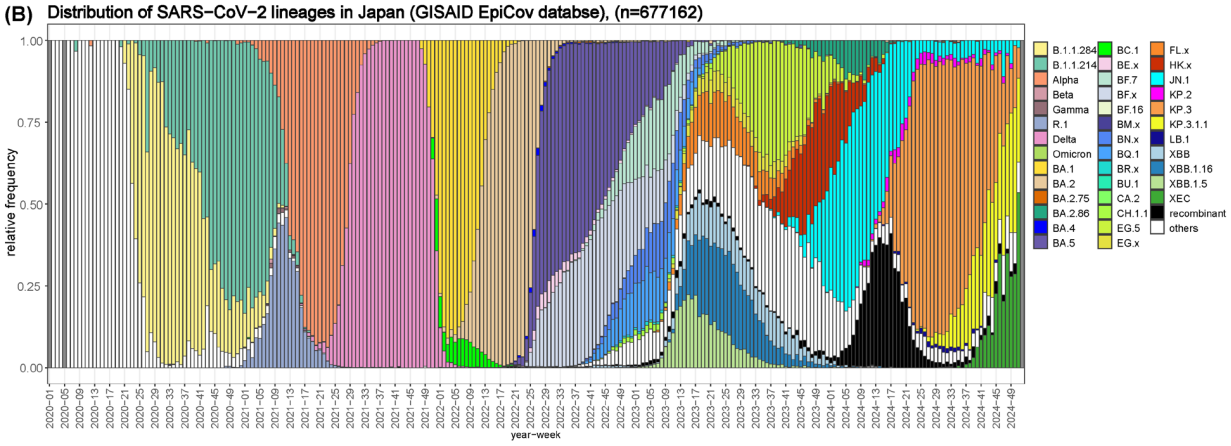
Following the probabilistic forecast evaluation protocol of Cramer et al. (2022), we report CRPS (Table 3) and empirical coverage (Table 4) to assess the quality of the mechanistic predictive distribution used for guidance.

Table 3. CRPS comparison between PatchSIR and `EpiDiff` at horizon $h = 7$. Lower is better.

Method	covid-US	covid-JP	influenza-US
PatchSIR	253.91	59.58	64.85
`EpiDiff`	196.00	46.59	49.41
Improvement	22.8%	21.8%	23.8%



(a) EpiDiff-based $R_{\text{eff}}(t)$ (7-day moving average).



(b) Distribution of SARS-CoV-2 lineages in Japan.

Figure 10. Comparison between the effective reproduction factor estimated by EpiDiff on `covid-JP` (up) and SARS-CoV-2 variant dynamics in Japan (down, reported by Takeuchi et al. (2025)).

Table 4. Empirical coverage of EpiDiff predictive intervals at horizon $h = 7$.

Nominal level	covid-US	covid-JP	influenza-US
70%	58.1%	50.3%	80.8%
90%	76.7%	69.5%	97.8%

C.10. Effect of Graph Structure

To examine whether the improvement of `EpiDiff` mainly comes from spatial message passing or from the proposed uncertainty-aware steering mechanism, we evaluate a graph-ablated variant on covid-JP by setting $A = I$. This removes all inter-region spatial edges and reduces the model to independent node-wise temporal forecasting with mechanistic steering.

As shown in the table, `EpiDiff` remains effective even without spatial edges: with $A = I$, it substantially outperforms the unsteered diffusion backbone under the same graph-ablated setting. This indicates that the core benefit of `EpiDiff` comes from uncertainty-aware mechanistic steering rather than solely from graph-based spatial propagation. At the same time, the full graph version achieves the best performance, suggesting that spatial dependencies provide additional gains when reliable adjacency information is available.

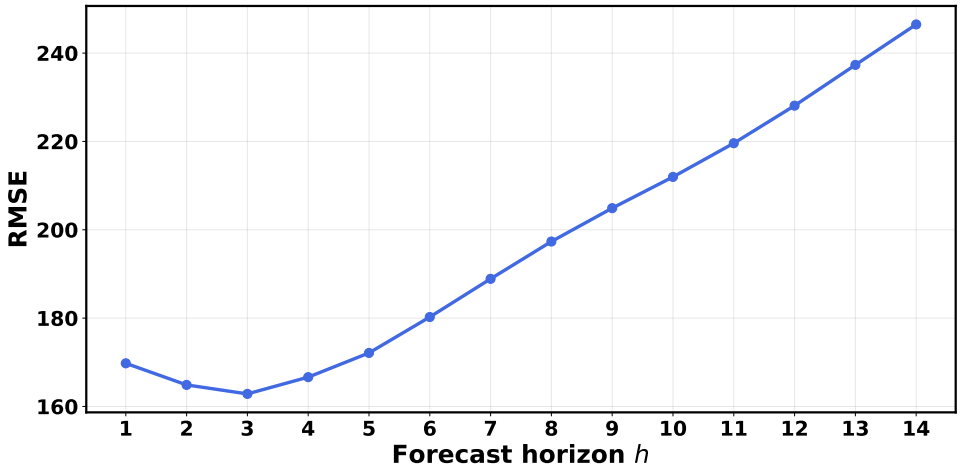


Figure 11. RMSE of `EpiDiff` evolves with the forecast horizon on `covid-JP`.

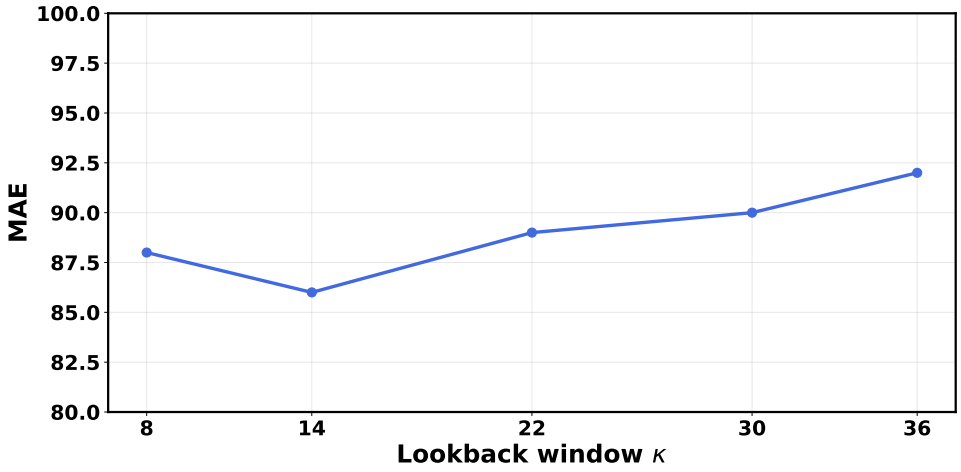


Figure 12. MAE of `EpiDiff` changes with different lookback windows T_h on `covid-JP`.

Table 5. Effect of graph structure on covid-JP. $A = I$ removes spatial connections.

Setting	$h = 7$ R / P	$h = 14$ R / P
EpiDiff (full graph)	188 / 62	247 / 86
EpiDiff ($A = I$)	197 / 66	258 / 92
w/o Steering ($A = I$)	383 / 97	468 / 140

D. Limitations

Our approach has several limitations. First, the mechanistic prior is based on simplified SIR-type dynamics with fixed parameters, which may not fully capture real-world heterogeneity or non-stationary interventions. Second, the overall performance is sensitive to the accuracy of mechanistic parameter estimation. As demonstrated in our ablation study, when the estimation is highly noisy or difficult in certain time windows, the resulting uncertainty can propagate into the guidance process and even degrade the predictive performance of the diffusion backbone. Finally, the scalability of our framework to very large graphs remains to be systematically validated. We view these limitations as opportunities for future work, including extensions to richer compartmental models, adaptive and more robust priors, and evaluation at larger scales.