LEARNING SPATIO-TEMPORAL FOUNDATION MODELS FOR TRAFFIC FORECASTING FROM PURE SYNTHETIC DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

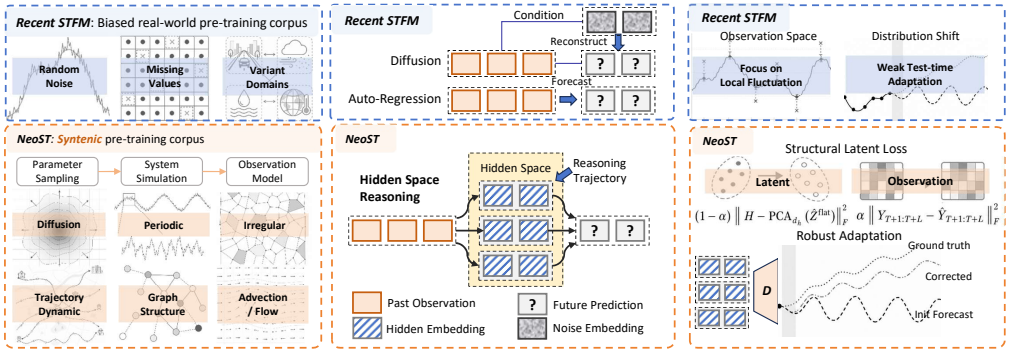
Spatio-temporal foundation models (STFMs) offer a promising path toward universal traffic forecasting by learning transferable priors across heterogeneous road networks and cities However, existing approaches suffer from distributional bias in real-world pre-training data, structural bottlenecks of autoregressive or diffusion-based paradigms, and objectives that overemphasize point-wise reconstruction in noisy observation space. We propose **NeoST**, the first STFM for traffic forecasting pre-trained solely on procedurally generated synthetic systems and driven by a latent-space reasoning architecture that turns this synthetic supervision into transferable structural priors. NeoST introduces a scalable synthetic pre-training corpus to mitigate real-world bias, iterative latent-space refinement of multiple future trajectories without sequential error accumulation, and latent-space objectives that emphasize structural dynamics and enable inference-time correction under distribution shifts. Extensive experiments across diverse real-world traffic benchmarks show that NeoST consistently outperforms existing STFMs in diverse real-world traffic systems, achieves superior long-horizon stability and inference efficiency.

1 INTRODUCTION

Traffic forecasting Jiang & Luo (2022) is a fundamental task underlying modern urban mobility systems, informing applications such as congestion warning, adaptive signal control, and route planning Zhang et al. (2024c). Conventional approaches Jin et al. (2023a); Yu et al. (2017) are typically trained as task-specific models tied to a single road network, city, or sensor deployment, learning patterns that are tightly coupled to the particular topology and traffic dynamics observed during training. The recent success of foundation models, especially Spatio-Temporal Foundation Models (STFMs) Liang et al. (2025); Yu et al. (2017), which are pretrained at scale and adapted to new scenarios through zero-shot inference or lightweight tuning, offers a promising alternative to this fragmented paradigm by distilling transferable structural priors that hold across diverse traffic systems. Building a spatio-temporal foundation model capable of generalizing across heterogeneous road networks and cities therefore represents a compelling direction toward universal traffic forecasting.

Despite their promise, the development of STFMs faces fundamental challenges across the entire modeling pipeline, beginning with the quality and diversity of available training data Fang et al. (2026). Models pre-trained on specific real-world collections often overfit to the idiosyncratic topology and temporal rhythms of those specific sensor deployments, hindering their generalizability to unseen scenarios. This imbalance induces strong distributional bias during pre-training: a model trained mainly on freeway sensor networks may fail to capture the signal-controlled dynamics of an unseen city's arterial roads, underperforming precisely when transferred to less-represented systems. Additionally, real-world spatio-temporal data are often corrupted by sensor noise, irregular sampling, and extensive missing values that follow fixed, uncontrolled patterns, such as loop detectors that drop out during the very accidents or storms that congestion warning most needs to anticipate Wang et al. (2023). Such imperfections not only degrade representation quality but also entangle intrinsic system dynamics with measurement artifacts, making it difficult for STFMs to distill clean and universal structural priors of space and time from noisy, incomplete observations.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Data (b) Model (c) Training & Inference

Figure 1: Comparison of Paradigms: NeoST vs. Recent STFM.

Beyond data limitations, current STFMs, including traffic-specific models such as OpenCity Li et al. (2024c) and UniST Yuan et al. (2024), are largely built upon either autoregressive Liang et al. (2024); Benidis et al. (2022); Lim & Zohren (2021) or diffusion-based paradigms Yang et al. (2026; 2024); Wen et al. (2023); Hu et al. (2024), each of which introduces structural bottlenecks. Autoregressive models generate future states sequentially, conditioning each prediction on previously generated outputs. While conceptually straightforward, this paradigm is prone to error accumulation: an underestimated queue length at one intersection can propagate into systematically wrong congestion forecasts for an adaptive signal controller several intersections downstream, hindering stable long-term forecasting. Diffusion-based models alleviate some issues of sequential dependency by modeling complex distributions through iterative denoising, but they typically require dozens to hundreds of sampling steps during inference. Such computational overhead is prohibitive in time-sensitive scenarios, such as a city-wide traffic management center that must refresh congestion forecasts across thousands of road segments every few minutes, where rapid and reliable predictions are essential.

Challenges further arise in training objectives and inference paradigms Fang et al. (2026). Spatio-temporal signals often exhibit sparse or weak semantics in the raw time domain, where point-wise losses Jadon et al. (2024) encourage models to focus on local fluctuations while overlooking global structures such as trends, periodicities, and latent dynamical mechanisms Cuturi & Blondel (2017); Le Guen & Thome (2019), capturing an instantaneous speed dip while missing the build-up of a rush-hour congestion wave that route-planning apps need to anticipate. In addition, most existing STFMs lack effective test-time adaptation: when a lane closure or a signal-timing change reshapes traffic patterns overnight, these models treat the observed input sequence as fixed context, unable to leverage incoming observations to recalibrate predictions, which constrains robustness and reliability in open-world deployment under evolving infrastructure conditions.

To tackle these challenges, we propose **NeoST**, the first spatio-temporal foundation model for traffic forecasting **pre-trained solely on synthetic spatio-temporal data**. Our main contributions can be summarized as the following three aspects:

- On the **data** side, we depart from conventional reliance on large but biased real-world datasets and instead construct a synthetic pre-training corpus with virtually unlimited scale and controllable diversity, spanning a broad range of dynamics and structural patterns. Rather than avoiding noise and missing values, we deliberately inject them in controllable, unbiased form, training the model to disentangle intrinsic dynamics from measurement artifacts and transfer robustly to real-world traffic data across diverse cities and sensor networks.
- On the **architecture** side, we propose a latent-space reasoning paradigm for STFMs, replacing the dominant autoregressive and diffusion-based formulations. Our model generates multiple candidate future trajectories directly in latent space and performs iterative refinement through patch-level and trajectory-level correction mechanisms. By decoupling reasoning from step-wise generation, NeoST is capable of deeper deliberation over alternative futures under a fixed computational budget.
- On the **training & inference** side, NeoST compute losses in latent space rather than directly in the raw time domain. This approach encourages the model to learn and reason about the underlying evolutionary dynamics and structural patterns of spatio-temporal systems, rather than merely fitting superficial value-level details. Moreover, the latent reasoning process naturally enables inference-time correction based on observed inputs, enhancing robustness under distribution shifts.

2 Related Works

The success of large-scale pre-training in NLP has inspired analogous efforts in time series and spatio-temporal data Jin et al. (2023b); Ni et al. (2026). Time Series Foundation Models (TSFMs) Liang et al. (2024) such as TimesFM Das et al. (2024), Chronos Ansari et al. (2024), and Moirai Liu et al. (2024) consolidate multiple forecasting tasks within a single pre-trained backbone, demonstrating the viability of foundation-scale temporal representation learning. However, these models ignore the rich spatial structure present in real-world systems. Emerging STFMs extend this paradigm to spatially resolved domains Liang et al. (2025); Goodge et al. (2025). OpenCity Li et al. (2024c) integrates Transformer-GNN hybrids pre-trained on heterogeneous traffic datasets to enable zero-shot urban forecasting, UniST Yuan et al. (2024) introduces knowledge-guided prompts for unified urban spatio-temporal prediction across few-shot and zero-shot settings, and UrbanGPT Li et al. (2024b) couples spatio-temporal encoders with LLM prompting for instruction-based spatial reasoning under limited supervision. In climate science Liang et al. (2023); Feng et al. (2024), models such as Pangu-Weather Bi et al. (2023) and ClimaX Nguyen et al. (2023) apply large Transformer architectures to global reanalysis data, achieving spatio-temporal generality across physical domains. Though promising, existing STFMs share common limitations: they rely on real-world data that is domain-biased and noisy, adopt either error-prone autoregressive decoding or computationally intensive diffusion sampling, and optimize directly in the observation space without structural abstraction. *An expanded version of the related works can be found in the Appendix A.*

3 Synthetic Spatio-Temporal Data

We construct our synthetic data engine by explicitly parameterizing a family of stochastic graph dynamical systems defined on a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $|\mathcal{V}| = N$. Let $\mathbf{A} \in \mathbb{R}^{N\times N}$ denote the adjacency matrix and $\mathbf{L}$ the corresponding normalized graph Laplacian. At each time step t, the latent dynamical state is represented as $\mathbf{H}_t \in \mathbb{R}^{N\times d_h}$, where d_h denotes the latent channel dimension. The observable signal $\mathbf{x}_t \in \mathbb{R}^N$ is obtained through a random linear projection of $\mathbf{H}_t$, while the latent state itself remains fully accessible during training. Further analysis is provided in Appendix B.

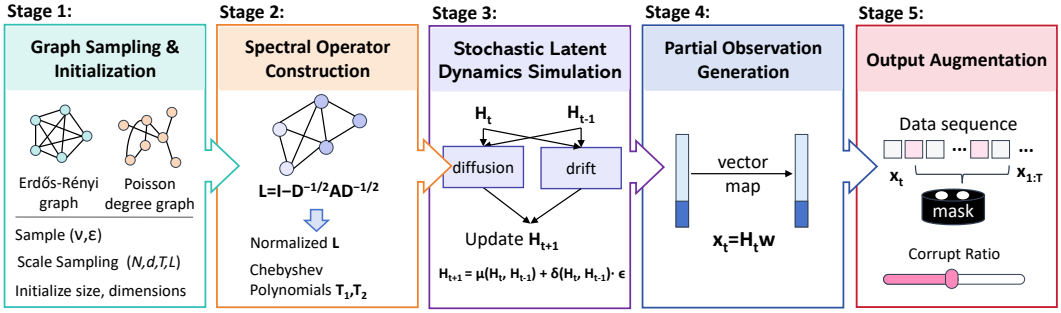


Figure 2: Synthetic data pipeline.

Graph Sampling & Operator Construction. To promote structural diversity, the graph topology is sampled on-the-fly from random graph families, including Erdős–Rényi and Poisson degree models. Self-loops are added to stabilize the spectrum. Given $\mathbf{A}$, we compute the normalized Laplacian $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$, where $\mathbf{D}$ is the degree matrix. To model spatial interactions in a principled manner, we leverage Chebyshev polynomials of the rescaled Laplacian. Let $\widehat{\mathbf{L}}$ denote the spectrum-normalized Laplacian, and define $\mathbf{T}_1 = \widehat{\mathbf{L}}, \mathbf{T}_2 = 2\widehat{\mathbf{L}}^2 - \mathbf{I}$. These operators provide localized yet expressive spectral filters without requiring eigen-decomposition during sequence generation. To ensure scale generalization, we dynamically sample system configurations from broad distributions, including graph size N, latent dimension d, and horizons (T, L), exposing the model to highly heterogeneous spatial scales and temporal regimes.

Stochastic Latent Dynamics. The temporal evolution follows a stochastic recurrence:

$$\mathbf{H}_{t+1} = a\mathbf{H}_t + \boldsymbol{\mu}(\mathbf{H}_t, \mathbf{H}_{t-1}, \mathbf{A}) + \boldsymbol{\sigma}(\mathbf{H}_t, \mathbf{H}_{t-1}, \mathbf{A}) \odot \boldsymbol{\epsilon}_t + \mathbf{b}, \tag{1}$$

where $a \in (0, 1)$ controls system dissipation, $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ introduces stochastic excitation, $\mathbf{b}$ is a sample-wise bias term, and $\odot$ denotes element-wise multiplication. The drift term $\boldsymbol{\mu}$ and diffusion

term $\boldsymbol{\sigma}$ are parameterized as

$$\boldsymbol{\mu}(\cdot) = \tanh\Big([\mathbf{T}_1\mathbf{H}_t \parallel \mathbf{T}_2\mathbf{H}_{t-1}]\mathbf{W}_1\Big)\mathbf{W}_2, \quad \boldsymbol{\sigma}(\cdot) = \text{sigmoid}\Big([\mathbf{T}_1\mathbf{H}_t \parallel \mathbf{T}_2\mathbf{H}_{t-1}]\mathbf{W}_3\Big), \tag{2}$$

where $\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3 \in \mathbb{R}^{d_h \times d_h}$ are learnable parameters of the generator and $[\cdot\|\cdot]$ denotes channel-wise concatenation. This formulation couples first- and second-order spectral responses with nonlinear transformations, producing rich spatio-temporal behaviors including diffusion-like smoothing, oscillatory propagation, and stochastic perturbations. Notably, the generator parameters are randomly initialized for each sample, yielding a distribution over dynamical systems rather than a single fixed simulator. Consequently, the pre-training corpus spans a broad manifold of graph-coupled stochastic processes with heterogeneous spectral characteristics and stability regimes.

Observation Generation. Observable signals are obtained through a random linear projection $\mathbf{x}_t = \mathbf{H}_t\mathbf{w}$, where $\mathbf{w} \in \mathbb{R}^d$ is sampled per instance. This projection induces partial observability, mimicking real-world sensing processes that capture low-dimensional measurements of higher-dimensional latent states. To approximate realistic data imperfections while maintaining full control of the generative mechanism, we randomly mask entries in $\mathbf{X}_{1:T}$ with a controllable ratio. Since these corruptions are externally imposed and independent of the underlying dynamics, the model can in principle disentangle intrinsic evolution patterns from observational artifacts.

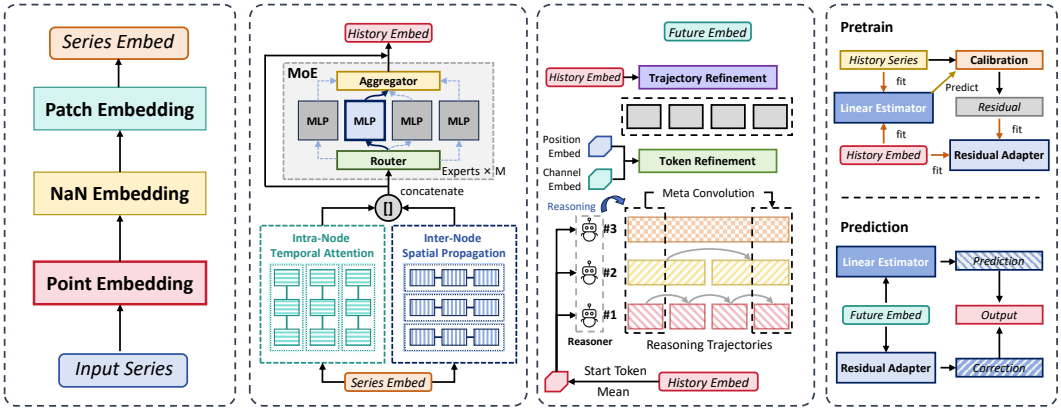
Output Augmentation. To bridge the gap between ideal synthetic signals and imperfect real-world measurements, we randomly mask a proportion of the generated observable sequence with missing values. This simulates the missing data ubiquitous in real-world sensor networks. Crucially, while missing data in real-world datasets often exhibits fixed topological biases (e.g., broken sensors at specific locations), our random masking strategy enforces an unstructured, unbiased distribution of missingness across both space and time. This prevents the model from relying on superficial, dataset-specific failure topologies and instead forces it to learn robust, position-invariant spatiotemporal dependencies, significantly enhancing its zero-shot transferability to unseen downstream systems.

4 Model Architecture

A Spatio-Temporal Foundation Model (STFM) is defined as a parametric function $f_\theta : (\mathbf{X}_{1:T}, \mathbf{A}) \mapsto \widehat{\mathbf{Y}}_{T+1:T+L}$, where θ denotes the model parameters, $\mathbf{X}_{1:T} \in \mathbb{R}^{T\times N}$ is the historical observation, $\mathbf{A} \in \mathbb{R}^{N\times N}$ encodes spatial dependencies, and $\widehat{\mathbf{Y}}_{T+1:T+L} \in \mathbb{R}^{L\times N}$ is the predicted future sequence. Formally, an STFM should support arbitrary input lengths $T \in \mathbb{N}^+$, arbitrary graph sizes $N \in \mathbb{N}^+$, and arbitrary forecasting horizons $L \in \mathbb{N}^+$. NeoST follows a latent-space spatio-temporal modeling paradigm that decouples representation learning, global reasoning, and output adaptation within a unified foundation backbone (Figure 3). Raw observations are first transformed into structured patch-level embeddings capturing local temporal continuity and node-level semantics, then processed by a latent **Spatio-Temporal Encoder** that jointly models temporal dependencies and graph-aware spatial interactions. A **Hierarchical Latent Reasoner** then generates future trajectory embeddings via multi-step global reasoning directly in latent space, constructing and refining candidate futures without strictly autoregressive decoding. Finally, an **Output Adapter** maps latent predictions back to the observation space to produce $\widehat{\mathbf{Y}}_{T+1:T+L}$.

4.1 Preprocessing

Given an input sequence $\mathbf{X}_{1:T} \in \mathbb{R}^{T\times N}$, we first apply instance normalization Kim et al. (2021) along the temporal dimension to reduce scale heterogeneity. For each node v_i, we compute the temporal mean $\mu^{(i)}$ and standard deviation $\sigma^{(i)}$, and normalize the signal via $\hat{x}_t^{(i)} = (x_t^{(i)} - \mu^{(i)})/\tilde{\sigma}^{(i)}$, where $\tilde{\sigma}^{(i)} = \max(\sigma^{(i)}, 1)$ prevents small-variance sequences from being over-amplified. Each normalized scalar observation is then lifted to a d_p-dimensional vector through a fixed random linear projection $\mathbf{z}_t^{(i)} = \hat{x}_t^{(i)}\mathbf{w} + \mathbf{b}$, where $\mathbf{w}$ and $\mathbf{b}$ are randomly initialized and kept non-trainable to stabilize optimization, and masked values are mapped to zero vectors. The resulting embeddings are segmented into temporal patches of length p with stride s, optionally allowing overlap, and each flattened patch is projected to the hidden dimension d to form patch tokens. An optional trend decomposition branch computes first-order differences between adjacent patches and injects the encoded residual signal into the patch representation, enabling sensitivity to short-range dynamics.



(a) Series Preprocess (b) Spatio-Temporal Encoder (c) Hierarchical Reasoner (d) Output Adapter

Figure 3: Overall architecture of NeoST.

4.2 Spatio-Temporal Encoder

The spatio-temporal encoder transforms the patch-level representation $\mathbf{Z}_{1:P_T} \in \mathbb{R}^{P_T \times N \times d}$, where P_T denotes the number of patches obtained from patchifying the input sequence, into a contextual latent memory. This is accomplished through alternating between two complementary mechanisms: intra-node temporal attention, which captures temporal evolution within individual nodes, and inter-node spatial propagation, which disseminates information across nodes according to learned relational affinities. The encoder stacks L_{enc} such layers, with each layer jointly implementing both mechanisms in a unified computation block. Critically, the two pathways are designed to be mutually exclusive in their information sources, ensuring that each mechanism contributes a distinct representation.

Intra-node Temporal Attention. For each node v_i, the patch sequence $\mathbf{Z}^{(i)}_{1:P_T} \in \mathbb{R}^{P_T \times d}$ is processed independently by a causal gate self-attention module Qiu et al. (2025) equipped with Rotary Position Embeddings (RoPE) Su et al. (2024), which encodes relative temporal positions without relying on additive positional encodings. This operation produces two complementary value representations per node $\mathbf{V}^{\text{res}}$ and $\mathbf{V}^{\text{dis}}$. $\mathbf{V}^{\text{res}} \in \mathbb{R}^{P_T \times N \times d}$ captures the self-contained temporal dynamics of each node and serves as a residual pathway that preserves node-specific history. $\mathbf{V}^{\text{dis}} \in \mathbb{R}^{P_T \times N \times d}$ produces temporal features that are subsequently made available for spatial distribution to neighboring nodes. The temporal pathway operates independently per node while the spatial pathway is explicitly prevented from collapsing into identity by masking self-connections, ensuring that each mechanism contributes a distinct and complementary inductive bias to the final representation.

Inter-node Spatial Propagation. Given the temporally encoded features, the encoder performs spatial information routing through a graph attention mechanism operating across nodes at each patch position. Queries $\mathbf{Q} \in \mathbb{R}^{P_T \times N \times d}$ and keys $\mathbf{K} \in \mathbb{R}^{P_T \times N \times d}$ are computed from the current patch representations, and a node-to-node attention map is derived as $\mathbf{A}_p = \text{softmax}\left(a\mathbf{Q}_p\mathbf{K}_p^\top + \mathbf{M}\right) \in \mathbb{R}^{N \times N}$, where $a = 1/\sqrt{d/H}$ is the scalar factor, H is the number of attention heads, $\mathbf{M}$ is a masking matrix with $\mathbf{M}_{ii} = -\infty$ and $\mathbf{M}_{ij} = 0$ for $i \neq j$. The diagonal masking explicitly prohibits each node from attending to itself during spatial aggregation, ensuring that the spatial pathway carries exclusively cross-node information. The attention-weighted aggregation of $\mathbf{V}^{\text{dis}}$ yields a spatially-propagated representation $\mathbf{V}^{\text{agg}} = \mathbf{A}_p\mathbf{V}^{\text{dis}}_p$, which is then modulated by a learned sigmoid gate $\mathbf{g} = \sigma(\mathbf{W}_g\mathbf{Z})$ to adaptively control how much spatial context is incorporated at each position.

Output Combination. The outputs of both pathways are concatenated and linearly projected to produce the updated representation, followed by a position-wise feed-forward network:

$$\mathbf{Z}' = \mathbf{W}_o\left[\mathbf{V}^{\text{res}} \,\|\, \mathbf{V}^{\text{agg}} \odot \mathbf{g}\right] + \mathbf{Z}, \qquad \mathbf{Z}'' = \text{FFN}(\mathbf{Z}') + \mathbf{Z}'. \tag{3}$$

This formulation ensures that each node's output integrates both its own temporal trajectory and a gated summary of information propagated from the rest of the graph. The final encoder output $\mathbf{Z}_{1:P_T} \in \mathbb{R}^{P_T \times N \times d}$ serves as the contextual memory that conditions all subsequent latent reasoning. Furthermore, the dense residual pathways embedded within this integration step facilitate stable gradient flow, guaranteeing robust representation learning during the end-to-end pre-training phase.

4.3 HIERARCHICAL REASONER

Given the contextual patch memory $\mathbf{Z}_{1:P_T} \in \mathbb{R}^{P_T \times N \times d}$ produced by the spatio-temporal encoder, the hierarchical reasoner is responsible for generating future trajectory embeddings directly in latent space. Rather than decoding predictions step-by-step in the observation domain, the reasoner operates entirely over patch tokens, producing $\hat{\mathbf{Z}}_{1:P_L} \in \mathbb{R}^{P_L \times N \times d}$, where P_L is the number of output patch tokens corresponding to the target horizon L. The overall reasoning process follows a coarse-to-fine pipeline: An ensemble of multi-scale predictors first generates candidate trajectory hypotheses at different temporal resolutions in latent space, which are then consolidated and refined at the token and trajectory levels to align with the encoded historical context.

Multi-Scale Latent Prediction. To capture future dynamics at multiple temporal resolutions, we instantiate a set of S independent predictors $\{\mathcal{P}_s\}_{s=1}^{S}$, each operating at a distinct temporal stride Δ_s patch steps. Predictor $\mathcal{P}_s$ maps a global context summary $\bar{\mathbf{z}} \in \mathbb{R}^{N \times d}$ to a sequence of latent states at its native stride and truncated to P_L steps, which can be written as $\hat{\mathbf{Z}}^{(s)}_{1:P_L} = \mathcal{P}_s(\bar{\mathbf{z}},\ P_L) \in \mathbb{R}^{P_L \times N \times d}$. Here, $\bar{\mathbf{z}} = (1/P)\sum_{p=1}^{P} \mathbf{Z}_p$ condenses the average observed historical context into a compact initialization signal. The S resulting trajectories are concatenated along the feature dimension to form a multi-scale composite representation $\hat{\mathbf{Z}}^{\text{cat}}_{1:P_L} = \left[\hat{\mathbf{Z}}^{(1)} \,\|\, \hat{\mathbf{Z}}^{(2)} \,\|\, \cdots \,\|\, \hat{\mathbf{Z}}^{(S)}\right] \in \mathbb{R}^{P_L \times N \times Sd}$, to which a learnable scale embedding $\mathbf{e}_{\text{ch}} \in \mathbb{R}^{Sd}$ is added to distinguish each scale's contribution prior to fusion. Each predictor $\mathcal{P}_s$ is realized as a ChebConv-LSTM Seo et al. (2016) with polynomial order K, capturing K-hop spatial dependencies across the road network.

Token-Level Refinement via Conditional Temporal Convolution. The multi-scale composite is subsequently fused into a unified d-dimensional token sequence. Since the appropriate fusion strategy depends on the specific input context, we formulate token-level refinement as a context-conditioned operation, with fusion weights dynamically generated from the observed context rather than fixed as shared parameters. A lightweight meta-network ϕ_{meta} takes the global context summary $\bar{\mathbf{z}}$ and produces an instance-specific convolutional kernel $\mathbf{W}^{\text{dyn}} = \phi_{\text{meta}}(\bar{\mathbf{z}}) \in \mathbb{R}^{d \times (Sd+d) \times k}$, where k is the kernel width, applied independently per node along the temporal dimension to an augmented input $\left[\hat{\mathbf{Z}}^{\text{cat}} + \mathbf{e}_{\text{ch}} \,\|\, \mathbf{E}^{\text{pos}}\right]$, which concatenates the channel embedding $\mathbf{e}_{\text{ch}} \in \mathbb{R}^{Sd}$ with sinusoidal position embeddings $\mathbf{E}^{\text{pos}} \in \mathbb{R}^{P_L \times d}$, yielding $\hat{\mathbf{Z}}^{\text{fused}} \in \mathbb{R}^{P_L \times N \times d}$.

Trajectory-Level Refinement via Encoder-Decoder Reasoning. While token-level fusion reconciles scale inconsistencies locally, the resulting trajectory may still exhibit global structural incoherence or misalignment with the historical context. We therefore apply a two-stage refinement that jointly reasons over the full trajectory sequence: a trajectory encoder $\mathcal{R}_{\text{enc}}$ first processes $\hat{\mathbf{Z}}^{\text{fused}}$ to propagate long-range temporal dependencies across future tokens, after which a trajectory decoder $\mathcal{R}_{\text{dec}}$ grounds the refined trajectory in the observed history $\mathbf{Z}_{1:P_T}$ through cross-attention. Both are instantiated as stacks of transformer layers: $\mathcal{R}_{\text{enc}}$ employs causal gate self-attention with RoPE to model temporal ordering within the predicted trajectory, while $\mathcal{R}_{\text{dec}}$ performs cross gate attention between the refined trajectory queries and $\mathbf{Z}_{1:P_T}$ to enforce temporal consistency.

5 TRAINING & INFERENCE

Pre-training Phase. Beyond providing ground-truth future values $\mathbf{Y}_{T+1:T+L}$, the engine exposes the underlying latent state vectors $\{\mathbf{H}_t\}_{t=T+1}^{T+L}$ that directly govern the generative dynamics of each synthetic system. This privileged supervisory signal, unavailable from real-world datasets, forms the cornerstone of our pre-training objective. Concretely, we flatten $\hat{\mathbf{Z}}^{\text{out}}$ to $\hat{\mathbf{Z}}^{\text{flat}} \in \mathbb{R}^{(P_L N) \times d}$ and apply the Principal Component Analysis (PCA) Wold et al. (1987) to project it into the d_h-dimensional space of the synthetic latent matrix $\mathbf{H} \in \mathbb{R}^{(P_L N) \times d_h}$. The resulting pre-training objective, which we term **Latent Alignment Loss (LAL)**, interpolates between a latent alignment term and a standard observation-space reconstruction term via a scalar hyperparameter $\alpha \in [0, 1]$:

$$\mathcal{L}_{\text{LAL}} = (1-\alpha)\left\|\mathbf{H} - \text{PCA}_{d_h}\big(\hat{\mathbf{Z}}^{\text{flat}}\big)\right\|_F^2 + \alpha\left\|\mathbf{Y}_{T+1:T+L} - \hat{\mathbf{Y}}_{T+1:T+L}\right\|_F^2. \tag{4}$$

By placing the primary supervisory weight on the latent space, $\mathcal{L}_{\text{LAL}}$ steers NeoST toward internalizing the structural and dynamical properties of spatio-temporal systems rather than overfitting to the superficial statistical patterns of any particular observation domain.

Test-Time Adaptation. Let $\tilde{\mathbf{Z}}^{\text{enc}} \in \mathbb{R}^{T\times N\times d_s}$ denote the time-resolved encoder embeddings obtained by depatchifying the encoder output, where $d_s = d/p$ is the per-timestep embedding dimension and p is the patch length. Collecting all spatiotemporal tokens that are not masked or missing, we form the feature matrix $\mathbf{Z}_{\text{obs}} \in \mathbb{R}^{M\times d_s}$ and the corresponding observation vector $\mathbf{v}_{\text{obs}} \in \mathbb{R}^M$, where M is the number of valid observations. The adapter solves the ridge regression problem by utilizing the following equation: $\mathbf{w}^*, b^* = \arg\min_{\mathbf{w},\, b} \left\|\mathbf{Z}_{\text{obs}}\mathbf{w} + b\,\mathbf{1} - \mathbf{v}_{\text{obs}}\right\|^2 + \lambda\|\mathbf{w}\|^2$. The equation admits the closed-form solution $\mathbf{w}^* = (\mathbf{Z}_{\text{obs}}^\top \mathbf{Z}_{\text{obs}} + \lambda\mathbf{I})^{-1}\mathbf{Z}_{\text{obs}}^\top \tilde{\mathbf{v}}_{\text{obs}}$, where $\tilde{\mathbf{v}}_{\text{obs}}$ denotes the bias-corrected target. Denoting the depatchified future embeddings as $\tilde{\mathbf{Z}}^{\text{out}} \in \mathbb{R}^{L\times N\times d_s}$, the predicted future sequence is obtained by applying the fitted mapping $\hat{\mathbf{Y}}_{T+1:T+L} = \tilde{\mathbf{Z}}^{\text{out}}_{\text{flat}}\,\mathbf{w}^* + b^*$, where $\tilde{\mathbf{Z}}^{\text{out}}_{\text{flat}} \in \mathbb{R}^{(LN)\times d_s}$ is the reshaped future embedding matrix.

6 Experiments

In our experiments, we aim to address the following research questions (RQs):

- **(RQ1)**: How effectively does NeoST transfer to diverse unseen real-world systems?
- **(RQ2)**: How does each component and hyperparameter in NeoST contribute to model performance?
- **(RQ3)**: How efficient is NeoST compared to other foundation models?
- **(RQ4)**: How can synthetic spatio-temporal data enhance the generalization ability of models?

6.1 Experimental Setups

Datsets & Baselines. NeoST's pre-training utilizes the synthetic data introduced in Section 3. For testing, we selected six classic spatio-temporal datasets for evaluation: PEMS03/04/07/08 Guo et al. (2021), PEMS-BAY & METR-LA Li et al. (2018). We compare NeoST with the following models:

- **Spatio-Temporal Foundation Models (STFMs)**: FactoST Zhong et al., OpenCity Li et al. (2024c), UniST Yuan et al. (2024)
- **Time Series Foundation Models (TSFMs)**: ROSE Wang et al. (2024), TimesFM Das et al. (2024), Moirai Liu et al. (2025a);

We adopted Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) as evaluation metrics.

Implementation Details. All experiments are conducted on a single NVIDIA A100 80GB GPU paired with an Intel Xeon Platinum 8358P CPU (32C @ 2.60GHz). During the synthetic data generation process, the number of nodes and the latent dimension for each spatio-temporal graph are uniformly sampled from the ranges $[64, 512]$ and $[4, 16]$, respectively, producing 1,000 distinct samples per epoch. The core model architecture is configured with a hidden dimension of 96 and 4 attention heads. The spatio-temporal encoder stacks 4 layers, while the trajectory refinement module is structured with a 2-layer encoder and a 2-layer decoder. The hyper-connections throughout the network maintain an expansion rate of 4. We pre-train the model using AdamW Loshchilov & Hutter (2017) with a batch size of 8 for 15 epochs. The other settings are the same as those in Zhong et al..

6.2 Model Comparison (RQ1)

We evaluate the models on six established spatio-temporal datasets in the zero-shot scenario, with the results summarized in Table 1. NeoST consistently surpasses both spatio-temporal and time series foundation models, achieving the lowest MAE and RMSE in ten out of the twelve evaluated settings, with particularly substantial improvements on the challenging 96-step long-horizon scenarios across PEMS-03, PEMS-04, PEMS-07, and METR-LA. This sustained long-term accuracy validates the effectiveness of our latent reasoning paradigm in mitigating the error accumulation typical of autoregressive architectures. Although FactoST leads marginally on the short horizon of PEMS-Bay and TimesFM on the long horizon of PEMS-08, NeoST remains highly competitive while dominating the overall benchmark, confirming that pre-training exclusively on diverse synthetic data successfully circumvents the distributional biases of real-world datasets.

Table 1: Quantitative comparison of zero-shot forecasting performance. Results are reported as MAE / RMSE, where lower values indicate better predictive performance. The best and second-best results are highlighted in **red** and blue, respectively. Datasets marked with an asterisk (*) were included in the pre-training corpora of specific baselines and are therefore excluded from the ranking.

Model Type		Ours	STFMs			TSFMs		
Dataset	Horizon	NeoST	FactoST	OpenCity	UniST	ROSE	TimesFM	Moirai
PEMS-03	12→12	**25.16 / 34.90**	30.12 / 46.92	30.37 / 47.49	101.87 / 129.94	29.74 / 47.27	29.71 / 49.21	28.23* / 46.36*
	96→96	**84.49 / 107.85**	113.62 / 144.80	125.18 / 159.23	102.87 / 137.90	122.93 / 158.76	108.59 / 152.74	74.85* / 110.78*
PEMS-04	12→12	**30.98 / 43.63**	38.65 / 56.59	39.34* / 58.50*	67.91 / 87.63	38.35 / 57.34	35.00 / 53.27	34.65* / 52.13*
	96→96	**105.25 / 137.35**	142.11 / 175.99	153.18* / 188.95*	115.76 / 153.58	151.39 / 189.29	127.39 / 171.13	105.95* / 141.40*
PEMS-Bay	12→12	2.67 / 4.80	**2.02 / 4.59**	3.23* / 6.91*	14.89 / 16.92	2.72 / 6.41	6.59 / 14.99	1.97* / 4.69*
	96→96	**4.86 / 8.14**	6.12 / 10.47	6.92* / 11.79*	9.29 / 12.72	7.02 / 11.90	14.16 / 24.01	6.51* / 11.70*
PEMS-07	12→12	**35.70 / 49.48**	45.47 / 66.68	45.18 / 67.15	104.93 / 133.65	44.79 / 67.33	42.10 / 64.59	35.64* / 50.25*
	96→96	**144.18 / 175.65**	156.07 / 190.59	172.20 / 211.25	144.67 / 184.62	172.07 / 212.94	151.99 / 205.18	125.91* / 169.43*
PEMS-08	12→12	**26.28 / 36.42**	32.14 / 47.27	32.45* / 48.42*	73.46 / 93.14	31.70 / 47.77	29.68 / 45.18	38.23* / 53.12*
	96→96	94.03 / **121.87**	121.43 / 151.59	128.48* / 161.75*	104.77 / 136.21	129.05 / 163.15	**92.72** / 126.10	119.10* / 151.83*
METR-LA	12→12	**4.90 / 8.21**	5.98 / 14.26	4.30* / 8.37*	24.33 / 29.31	6.05 / 14.36	6.59 / 14.99	5.55 / 13.79
	96→96	**8.31 / 13.70**	12.68 / 18.82	10.85* / 17.60*	25.88 / 30.19	12.96 / 23.61	14.16 / 24.01	12.87 / 22.19

6.3 Ablation and Hyperparameter Analysis (RQ2)

To validate NeoST's core components, we evaluate zero-shot forecasting MAE under short-term (12-step) and long-term (96-step) horizons on PEMS03 and PEMS07. This analysis ablates critical architectural modules, examines the latent alignment trade-off hyperparameter $\alpha \in [0, 1]$ (Figure 4), and isolates the efficacy of test-time adaptation (Figure 5).

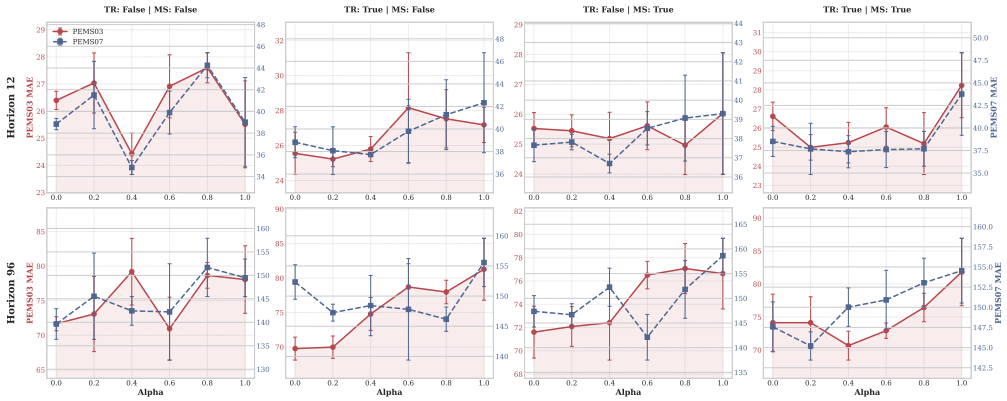


Figure 4: Ablation and hyperparameter analysis.

- **Architectural Contributions:** The full NeoST framework attains the lowest MAE and exhibits superior predictive stability across most evaluated scenarios. Specifically, disabling multi-scale prediction significantly impairs forecasting robustness over long-term horizons, whereas omitting the trajectory refinement module results in a substantial increase in cumulative structural errors.
- **Component Synergy:** The four configurations in Figure 4 show the two components are complementary: removing either alone already widens the MAE and its variance, while removing both (TR: False | MS: False) yields the largest error and widest bars almost everywhere. This indicates that trajectory refinement mainly curbs cumulative drift while multi-scale prediction stabilizes cross-horizon consistency, and that the two mechanisms are best deployed jointly.
- **Objective Trade-off (α):** Model predictive performance demonstrates high sensitivity to the balance between latent and observation-space objectives. We find that relying exclusively on latent alignment ($\alpha = 0.0$) results in suboptimal error rates, while pure observation-space optimization ($\alpha = 1.0$) restricts robust generalization under distributional shifts. The optimal balance consistently resides within $\alpha \in [0.2, 0.6]$, validating our proposed synergistic dual-objective design.
- **Test-Time Adaptation (TTA):** As shown in Figure 5, disabling the TTA module consistently degrades short-horizon (12-step) forecasting accuracy across all four datasets, with the largest drop observed on PEMS07. This confirms that leveraging inference-time observation context is especially effective for mitigating distribution shifts in short-term zero-shot transfer.

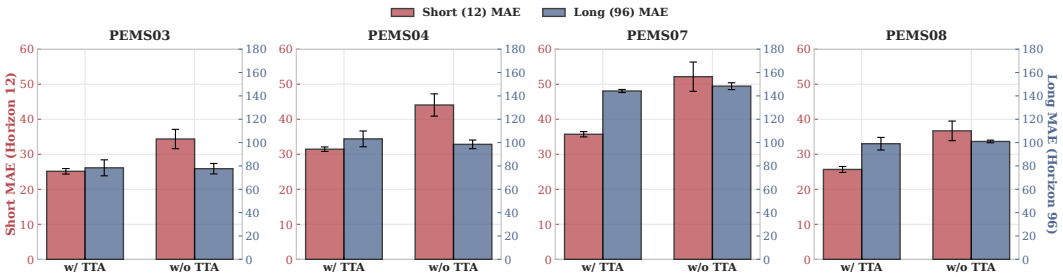


Figure 5: Impact of Test-Time Adaptation (TTA).

6.4 Inference Efficiency (RQ3)

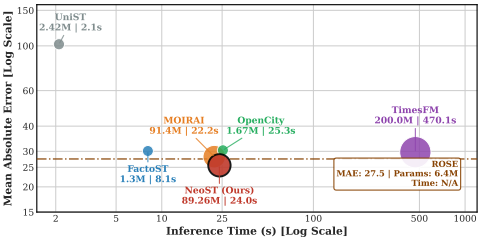


Figure 6: Efficiency comparison.

To answer RQ3, we compare NeoST's inference time, parameter scale, and MAE against representative baselines on PEMS03, as illustrated in Figure 6. NeoST achieves the lowest MAE among all evaluated models. While lightweight models like UniST (2.1s) and FactoST (8.1s) run faster, they incur substantially higher prediction errors, and massive models such as TimesFM (200.0M params, 470.1s) incur prohibitive computational costs without outperforming our approach. Operating with 89.26M parameters and a 24.0s inference time, NeoST's footprint is comparable to MOIRAI (22.2s) and OpenCity (25.3s) but delivers substantially better accuracy, establishing a favorable trade-off between inference efficiency and spatio-temporal modeling performance for real-world deployment. This gain follows directly from replacing step-by-step autoregressive or diffusion-based decoding with a single latent-space refinement pass.

6.5 Manifold Coverage of Synthetic Data (RQ4)

To address RQ4, we visualize the representational manifold of our synthetic corpus against six real-world benchmarks using t-SNE. After instance-normalizing fixed-length (12-step) temporal sequences from all datasets to isolate intrinsic dynamics, we projected them into a 2D space. As illustrated in Figure 7, the synthetic data forms an expansive, continuous manifold that envelopes the localized clusters of all real-world test sets. This empirical evidence validates our core hypothesis: pre-training exclusively on this highly diverse synthetic manifold equips the foundation model with universal structural priors, mitigating real-world data biases and enabling robust zero-shot.

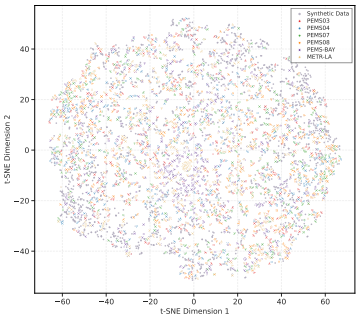


Figure 7: t-SNE visualization.

7 Conclusion & Future Works

In this work, we introduced NeoST, a spatio-temporal foundation model that overcomes the limitations of biased real-world datasets and computationally expensive generation paradigms for traffic forecasting. By pre-training exclusively on a procedurally generated synthetic corpus, NeoST mitigates distributional bias and measurement artifacts, while its latent-space reasoning paradigm replaces autoregressive and diffusion-based generation with iterative refinement, alleviating error accumulation and enabling efficient inference. Extensive experiments across diverse traffic benchmarks demonstrate that NeoST achieves superior generalization, long-horizon stability, and inference efficiency in real-world traffic systems. For future work, we plan to extend NeoST beyond traffic to other spatio-temporal domains with heterogeneous topologies, such as environmental sensing and epidemic modeling, and to enrich the synthetic generation engine with more complex physical constraints and non-stationary dynamics, further narrowing the remaining gap between synthetic pre-training and real-world deployment across diverse practical application scenarios.

8 AI USE STATEMENT

In accordance with the ICLR 2027 policy on the use of AI tools, we disclose that a large language model (LLM) was used during the preparation of this manuscript solely to assist with language polishing, i.e., improving the clarity, grammar, and fluency of certain sentences. The LLM was not used to generate research ideas, design the methodology, conduct experiments, derive theoretical results, or produce any of the technical content presented in this paper. All scientific contributions, including the problem formulation, the proposed method, the experimental design, and the analysis of results, are entirely the original work of the authors, who have carefully reviewed and take full responsibility for all content in this paper, including any text that was AI-polished.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. arXiv preprint arXiv:2403.07815, 2024.

Jost Arndt, Utku Isil, Michael Detzel, Wojciech Samek, and Jackie Ma. Synthetic datasets for machine learning on spatio-temporal graphs using pdes. arXiv preprint arXiv:2502.04140, 2025.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 15619–15629, 2023.

Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. arXiv preprint arXiv:2511.08544, 2025.

Konstantinos Benidis, Syama Sundar Rangapuram, Valentin Flunkert, Yuyang Wang, Danielle Maddix, Caner Turkmen, Jan Gasthaus, Michael Bohlke-Schneider, David Salinas, Lorenzo Stella, et al. Deep learning for time series forecasting: Tutorial and literature survey. ACM Computing Surveys, 55(6):1–36, 2022.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. Nature, 619(7970):533–538, 2023.

Marco Cuturi and Mathieu Blondel. Soft-dtw: a differentiable loss function for time-series. In International conference on machine learning, pp. 894–903. PMLR, 2017.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. ICML, 2024. Pretrained TimesFM on 100B time points from Google Trends, Wikipedia, synthetic and real-world data.

Thomas DP Edwards, James Alvey, Justin Alsing, Nam H Nguyen, and Benjamin D Wandelt. Scaling-laws for large time-series models. arXiv preprint arXiv:2405.13867, 2024.

Yuchen Fang, Hao Miao, Yuxuan Liang, Liwei Deng, Yue Cui, Ximu Zeng, Yuyang Xia, Yan Zhao, Torben Bach Pedersen, Christian S Jensen, et al. Unraveling spatio-temporal foundation models via the pipeline lens: A comprehensive review. IEEE Transactions on Knowledge and Data Engineering, 2026.

Yutong Feng, Qiongyan Wang, Yutong Xia, Junlin Huang, Siru Zhong, and Yuxuan Liang. Spatio-temporal field neural networks for air quality inference. In Kate Larson (ed.), Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24, pp. 7260–7268. International Joint Conferences on Artificial Intelligence Organization, 8 2024. doi: 10.24963/ijcai.2024/803. URL https://doi.org/10.24963/ijcai.2024/803. AI for Good.

Nan Gao, Hao Xue, Wei Shao, Sichen Zhao, Kyle Kai Qin, Arian Prabowo, Mohammad Saiedur Rahaman, and Flora D Salim. Generative adversarial networks for spatio-temporal data: A survey. ACM Transactions on Intelligent Systems and Technology (TIST), 13(2):1–25, 2022.

Adam Goodge, Wee Siong Ng, Bryan Hooi, and See Kiong Ng. Spatio-temporal foundation models: Vision, challenges, and opportunities. arXiv preprint arXiv:2501.09045, 2025.

Shengnan Guo, Youfang Lin, Huaiyu Wan, Xiucheng Li, and Gao Cong. Learning dynamics and heterogeneity of spatial-temporal graph data for traffic forecasting. IEEE Transactions on Knowledge and Data Engineering, 34(11):5415–5428, 2021.

Junfeng Hu, Xu Liu, Zhencheng Fan, Yuxuan Liang, and Roger Zimmermann. Towards unifying diffusion models for probabilistic spatio-temporal graph learning. In Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems, pp. 135–146, 2024.

Aryan Jadon, Avinash Patil, and Shruti Jadon. A comprehensive survey of regression-based loss functions for time series forecasting. In International Conference on Data Management, Analytics & Innovation, pp. 117–147. Springer, 2024.

Weiwei Jiang and Jiayun Luo. Graph neural network for traffic forecasting: A survey. Expert systems with applications, 207:117921, 2022.

Guangyin Jin, Yuxuan Liang, Yuchen Fang, Zezhi Shao, Jincai Huang, Junbo Zhang, and Yu Zheng. Spatio-temporal graph neural networks for predictive learning in urban computing: A survey. IEEE transactions on knowledge and data engineering, 36(10):5388–5408, 2023a.

Ming Jin, Qingsong Wen, Yuxuan Liang, Chaoli Zhang, Siqiao Xue, Xue Wang, James Zhang, Yi Wang, Haifeng Chen, Xiaoli Li, et al. Large models for time series and spatio-temporal data: A survey and outlook. arXiv preprint arXiv:2310.10196, 2023b.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In International conference on learning representations, 2021.

Vincent Le Guen and Nicolas Thome. Shape and time distortion loss for training deep time series forecasting models. Advances in neural information processing systems, 32, 2019.

Lihuan Li, Hao Xue, Yang Song, and Flora Salim. T-jepa: A joint-embedding predictive architecture for trajectory similarity computation. In Proceedings of the 32nd ACM international conference on advances in geographic information systems, pp. 569–572, 2024a.

Lihuan Li, Hao Xue, Shuang Ao, Yang Song, and Flora Salim. Hit-jepa: A hierarchical self-supervised trajectory embedding framework for similarity computation. arXiv preprint arXiv:2507.00028, 2025.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In International Conference on Learning Representations (ICLR '18), 2018.

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

Zhonghang Li, Long Xia, Lei Shi, Yong Xu, Dawei Yin, and Chao Huang. Opencity: Open spatio-temporal foundation models for traffic prediction. arXiv preprint arXiv:2408.10269, 2024c.

Yuxuan Liang, Yutong Xia, Songyu Ke, Yiwei Wang, Qingsong Wen, Junbo Zhang, Yu Zheng, and Roger Zimmermann. Airformer: Predicting nationwide air quality in china with transformers. In Proceedings of the AAAI conference on artificial intelligence, volume 37, pp. 14329–14337, 2023.

Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. Foundation models for time series analysis: A tutorial and survey. In Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining, pp. 6555–6565, 2024.

Yuxuan Liang, Haomin Wen, Yutong Xia, Ming Jin, Bin Yang, Flora Salim, Qingsong Wen, Shirui Pan, and Gao Cong. Foundation models for spatio-temporal data science: A tutorial and survey. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2, pp. 6063–6073, 2025.

Bryan Lim and Stefan Zohren. Time-series forecasting with deep learning: a survey. Philosophical transactions of the royal society a: mathematical, physical and engineering sciences, 379(2194), 2021.

Chenghao Liu, Taha Aksu, Juncheng Liu, Xu Liu, Hanshu Yan, Quang Pham, Silvio Savarese, Doyen Sahoo, Caiming Xiong, and Junnan Li. Moirai 2.0: When less is more for time series forecasting. arXiv preprint arXiv:2511.11698, 2025a.

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

Xu Liu, Juncheng Liu, Gerald Woo, Taha Aksu, Yuxuan Liang, Roger Zimmermann, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Moirai-moe: Empowering time series foundation models with sparse mixture of experts. arXiv preprint arXiv:2410.10469, 2024.

Xu Liu, Taha Aksu, Juncheng Liu, Qingsong Wen, Yuxuan Liang, Caiming Xiong, Silvio Savarese, Doyen Sahoo, Junnan Li, and Chenghao Liu. Empowering time series analysis with synthetic data: A survey and outlook in the era of foundation models. arXiv preprint arXiv:2503.11411, 2025b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. arXiv preprint arXiv:2301.10343, 2023.

Juntong Ni, Shiyu Wang, Ming Jin, Qi He, and Wei Jin. Streasoner: Empowering llms for spatio-temporal reasoning in time series via spatial-aware reinforcement learning. arXiv preprint arXiv:2601.03248, 2026.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, et al. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. arXiv preprint arXiv:2505.06708, 2025.

Youngjoo Seo, Michaël Defferrard, Pierre Vandergheynst, and Xavier Bresson. Structured sequence modeling with graph convolutional recurrent networks, 2016. URL https://arxiv.org/abs/1612.07659.

Jingzhe Shi, Qinwei Ma, Huan Ma, and Lei Li. Scaling law for time series forecasting. Advances in Neural Information Processing Systems, 37:83314–83344, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. Neurocomputing, 568:127063, 2024.

Xu Wang, Hongbo Zhang, Pengkun Wang, Yudong Zhang, Binwu Wang, Zhengyang Zhou, and Yang Wang. An observed value consistent diffusion model for imputing missing values in multivariate time series. In Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining, pp. 2409–2418, 2023.

Yihang Wang, Yuying Qiu, Peng Chen, Kai Zhao, Yang Shu, Zhongwen Rao, Lujia Pan, Bin Yang, and Chenjuan Guo. Rose: Register-assisted general time series forecasting with decomposed frequency learning. 2024.

Haomin Wen, Youfang Lin, Yutong Xia, Huaiyu Wan, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. Diffstg: Probabilistic spatio-temporal graph forecasting with denoising diffusion models. In Proceedings of the 31st ACM international conference on advances in geographic information systems, pp. 1–12, 2023.

Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. Chemometrics and intelligent laboratory systems, 2(1-3):37–52, 1987.

Lixia Wu, Haomin Wen, Haoyuan Hu, Xiaowei Mao, Yutong Xia, Ergang Shan, Jianbin Zheng, Junhong Lou, Yuxuan Liang, Liuqing Yang, et al. Lade: The first comprehensive last-mile express dataset from industry. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 5991–6002, 2024.

Yutong Xia, Chang Xu, Yuxuan Liang, Qingsong Wen, Roger Zimmermann, and Jiang Bian. Causal time series generation via diffusion models. arXiv preprint arXiv:2509.20846, 2025.

Yiyuan Yang, Ming Jin, Haomin Wen, Chaoli Zhang, Yuxuan Liang, Lintao Ma, Yi Wang, Chenghao Liu, Bin Yang, Zenglin Xu, Jiang Bian, Shirui Pan, and Qingsong Wen. A survey on diffusion models for time series and spatio-temporal data. 2024.

Yiyuan Yang, Ming Jin, Haomin Wen, Chaoli Zhang, Yuxuan Liang, Lintao Ma, Yi Wang, Chenghao Liu, Bin Yang, Zenglin Xu, Shirui Pan, and Qingsong Wen. A survey on diffusion models for time series and spatio-temporal data. ACM Comput. Surv., 58(8), February 2026. ISSN 0360-0300. doi: 10.1145/3783986. URL https://doi.org/10.1145/3783986.

Qingren Yao, Chao-Han Huck Yang, Renhe Jiang, Yuxuan Liang, Ming Jin, and Shirui Pan. Towards neural scaling laws for time series foundation models. arXiv preprint arXiv:2410.12360, 2024.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. arXiv preprint arXiv:1709.04875, 2017.

Yuan Yuan, Jingtao Ding, Jie Feng, Depeng Jin, and Yong Li. Unist: A prompt-empowered universal model for urban spatio-temporal prediction. Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 4095–4106, 2024.

Huaiwu Zhang, Yutong Xia, Siru Zhong, Kun Wang, Zekun Tong, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. Predicting parking availability in singapore with cross-domain data: A new dataset and a data-driven approach. arXiv preprint arXiv:2405.18910, 2024a.

Kexin Zhang, Qingsong Wen, Chaoli Zhang, Rongyao Cai, Ming Jin, Yong Liu, James Y Zhang, Yuxuan Liang, Guansong Pang, Dongjin Song, et al. Self-supervised learning for time series analysis: Taxonomy, progress, and prospects. IEEE transactions on pattern analysis and machine intelligence, 46(10):6775–6794, 2024b.

Weijia Zhang, Jindong Han, Zhao Xu, Hang Ni, Hao Liu, and Hui Xiong. Urban foundation models: A survey. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 6633–6643, 2024c.

Siru Zhong, Junjie Qiu, Yangyu Wu, Xingchen Zou, Zhongwen Rao, Bin Yang, Chenjuan Guo, Hao Xu, and Yuxuan Liang. Learning to factorize spatio-temporal foundation models. In The Thirty-ninth Annual Conference on Neural Information Processing Systems.

Yuanshao Zhu, Yongchao Ye, Shiyao Zhang, Xiangyu Zhao, and James Yu. Difftraj: Generating gps trajectory with diffusion probabilistic model. Advances in Neural Information Processing Systems, 36:65168–65188, 2023.

Yuanshao Zhu, James Jianqiao Yu, Xiangyu Zhao, Qidong Liu, Yongchao Ye, Wei Chen, Zijian Zhang, Xuetao Wei, and Yuxuan Liang. Controltraj: Controllable trajectory generation with topology-constrained diffusion model. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 4676–4687, 2024.

A Related Works

Synthetic Spatio-Temporal Data. The use of synthetic data for time series and spatio-temporal modeling has gained increasing attention, particularly in the era of foundation models. A comprehensive survey is provided by Liu et al. (2025b), which highlights procedural and simulation-based generation as scalable alternatives to real-world data collection. In contrast, most existing spatio-temporal models are trained on a limited set of real-world benchmarks, predominantly from transportation and urban domains, such as PEMS Guo et al. (2021), LargeST Liu et al. (2023), SINPA Zhang et al. (2024a), and LaDe Wu et al. (2024). While influential, these datasets share similar sensing infrastructures and domain biases, restricting distributional diversity. Extending synthesis to the spatio-temporal domain introduces additional complexity, as models must jointly capture spatial dependencies and temporal dynamics. GAN-based approaches Gao et al. (2022) generate realistic mobility traces or sensor readings but are typically domain-specific. PDE-based simulators Arndt

et al. (2025) construct synthetic spatio-temporal graphs governed by physical dynamics, offering improved controllability. Diffusion models for trajectory Zhu et al. (2023; 2024) and time series Xia et al. (2025) generation have also been explored. Despite these efforts, synthetic data is rarely used as the *primary* source for STFM pre-training. Most existing STFMs still rely heavily on aggregated real-world datasets, inevitably inheriting their inherent domain biases and persistent data quality limitations.

Training Paradigms and Objectives for Foundation Models. Traditional point-wise regression losses in the observation space often struggle to capture global structural similarities or temporal dynamics, prompting the use of specialized alignment metrics Jadon et al. (2024); Cuturi & Blondel (2017); Le Guen & Thome (2019). In parallel, Self-Supervised Learning (SSL) Zhang et al. (2024b) has transformed representation learning. A particularly promising evolution is the Joint-Embedding Predictive Architecture (JEPA) Assran et al. (2023); Balestriero & LeCun (2025). By forcing models to predict missing representations entirely within an abstract latent space rather than reconstructing raw values, JEPA inherently discards high-frequency noise and focuses on semantic features. This paradigm has recently achieved notable success in spatio-temporal tasks, such as trajectory representation learning via T-JEPA Li et al. (2024a) and HiT-JEPA Li et al. (2025). These developments conceptually echo our proposed Latent Alignment Loss (LAL) in NeoST, reinforcing the necessity of latent-space reasoning for complex dynamics. Furthermore, the shift toward foundation-level modeling is fundamentally intertwined with neural scaling laws. Recent empirical studies demonstrate that TSFMs adhere to predictable scaling behaviors, where performance consistently improves alongside increases in model parameters and dataset size Yao et al. (2024); Shi et al. (2024); Edwards et al. (2024). By coupling a highly scalable synthetic data engine with efficient latent-space reasoning, NeoST is inherently positioned to exploit these scaling principles, circumventing the finite scale and inherent biases of real-world spatio-temporal datasets.

B Theoretical Properties of Synthetic Data

This appendix characterizes the dynamical properties induced by the proposed synthetic data generator. Rather than instantiating a single fixed dynamical law, the generator samples graph topologies, generator parameters, biases, observation maps, and masking patterns, thereby defining a distribution over stochastic graph-coupled dynamical systems. Consequently, the synthetic corpus can contain a broad range of spatio-temporal regimes, including spatial dependence, temporal autocorrelation, delayed feedback, diffusion-like smoothing, wave-like propagation, quasi-periodic responses, nonlinear saturation, heteroscedastic stochasticity, and partial observability.

Let Θ denote the sampled system configuration for one synthetic instance. Unless otherwise stated, all statements below are conditioned on a fixed Θ.

Distribution over dynamical systems. Since each synthetic instance samples its own graph, generator weights, bias, and observation projection, the overall data distribution can be written as

$$p(\mathbf{X}) = \int p_{\Theta}(\mathbf{X})\, d\nu(\Theta), \tag{5}$$

where ν is the sampling distribution over system configurations. This mixture interpretation is central to the design of the generator: different sampled systems may exhibit different spatial scales, temporal responses, stochastic intensities, nonlinear regimes, and observation patterns. Thus, the corpus covers a heterogeneous family of structured spatio-temporal processes rather than trajectories from a single simulator.

Proposition B.1 (Conditional second-order Markov structure)**.** *For a fixed sampled configuration Θ, the latent process satisfies*

$$p(\mathbf{H}_{t+1} \mid \mathbf{H}_t, \mathbf{H}_{t-1}, \ldots, \mathbf{H}_0, \Theta) = p(\mathbf{H}_{t+1} \mid \mathbf{H}_t, \mathbf{H}_{t-1}, \Theta). \tag{6}$$

Equivalently,

$$\widetilde{\mathbf{H}}_t = \begin{bmatrix} \mathbf{H}_t \\ \mathbf{H}_{t-1} \end{bmatrix} \tag{7}$$

forms a first-order time-homogeneous Markov chain conditioned on Θ.

Proof. Conditioned on Θ, the one-step update depends on the past only through $(\mathbf{H}_t, \mathbf{H}_{t-1})$ and the fresh innovation $\boldsymbol{\epsilon}_t$, which is independent of previous states. The augmented-state representation then gives a first-order Markov chain. □

The conditioning on Θ is essential. Since generator parameters are sampled per instance, the marginal process obtained after integrating out Θ is a mixture of Markov processes and need not itself be second-order Markov.

Proposition B.2 (Conditionally Gaussian heteroscedastic transition)**.** *For fixed Θ, the transition distribution satisfies*

$$\mathrm{vec}(\mathbf{H}_{t+1}) \mid \mathbf{H}_t, \mathbf{H}_{t-1}, \Theta \sim \mathcal{N}(\mathbf{m}_t, \boldsymbol{\Sigma}_t), \tag{8}$$

where

$$\mathbf{m}_t = \mathrm{vec}(a\mathbf{H}_t + \boldsymbol{\mu}(\mathbf{H}_t, \mathbf{H}_{t-1}, \mathbf{A}) + \mathbf{b}), \tag{9}$$

and

$$\boldsymbol{\Sigma}_t = \mathrm{diag}\left(\mathrm{vec}\left(\boldsymbol{\sigma}(\mathbf{H}_t, \mathbf{H}_{t-1}, \mathbf{A})^2\right)\right). \tag{10}$$

Thus, the latent dynamics are conditionally Gaussian with state-dependent covariance.

Proof. Given $(\mathbf{H}_t, \mathbf{H}_{t-1})$ and Θ, the only random component in the update is

$$\boldsymbol{\sigma}(\mathbf{H}_t, \mathbf{H}_{t-1}, \mathbf{A}) \odot \boldsymbol{\epsilon}_t.$$

Since $\boldsymbol{\epsilon}_t$ has independent standard Gaussian entries, element-wise scaling yields a diagonal Gaussian covariance. □

Although each one-step transition is conditionally Gaussian, the process is generally not marginally Gaussian, because both its conditional mean and conditional variance depend nonlinearly on graph-filtered latent states. This state-dependent variance introduces heteroscedastic stochasticity, a common feature of real spatio-temporal sensing systems.

Theorem B.3 (Spectrally bounded graph filtering)**.** *Assume the sampled graph is undirected with nonnegative weights and nonzero degrees. Then*

$$\mathrm{spec}(\widehat{\mathbf{L}}) \subset [-1, 1], \tag{11}$$

and for any Chebyshev polynomial T_k,

$$\|T_k(\widehat{\mathbf{L}})\|_2 \leq 1. \tag{12}$$

In particular,

$$\|\mathbf{T}_1\|_2 \leq 1, \qquad \|\mathbf{T}_2\|_2 \leq 1. \tag{13}$$

Proof. For an undirected graph with nonnegative weights, the normalized Laplacian has spectrum in $[0, 2]$. Hence $\widehat{\mathbf{L}} = \mathbf{L} - \mathbf{I}$ has spectrum in $[-1, 1]$. Since $|T_k(x)| \leq 1$ for $x \in [-1, 1]$, symmetry of $\widehat{\mathbf{L}}$ gives

$$\|T_k(\widehat{\mathbf{L}})\|_2 = \max_{\lambda \in \mathrm{spec}(\widehat{\mathbf{L}})} |T_k(\lambda)| \leq 1.$$

□

Theorem B.3 concerns the stability of the graph filtering operators. At the same time, $\mathbf{T}_1\mathbf{H}_t$ and $\mathbf{T}_2\mathbf{H}_{t-1}$ explicitly couple neighboring or spectrally related nodes. Therefore, spatial dependence, graph smoothness, and localized propagation are structural properties induced by the generator.

Theorem B.4 (Dissipative moment control)**.** *Fix a sampled configuration Θ. If the initial latent states have finite second moments, then there exists $C_\Theta < \infty$ such that*

$$\sup_{t \geq 0} \mathbb{E}_\Theta \|\mathbf{H}_t\|_F^2 \leq C_\Theta. \tag{14}$$

Proof. Since $\tanh(\cdot)$ is bounded and the sampled weights are finite-dimensional, there exists $M_{\mu,\Theta} < \infty$ such that

$$\|\boldsymbol{\mu}(\mathbf{H}_t, \mathbf{H}_{t-1}, \mathbf{A})\|_F \le M_{\mu,\Theta}$$

for all states. Moreover, because $\mathrm{sigmoid}(z) \in (0,1)$,

$$\mathbb{E}_\Theta \left[\|\boldsymbol{\sigma}(\mathbf{H}_t, \mathbf{H}_{t-1}, \mathbf{A}) \odot \boldsymbol{\epsilon}_t\|_F^2 \mid \mathcal{F}_t \right] \le N d_h. \tag{15}$$

Choose $\eta > 0$ such that $\alpha = (1+\eta)a^2 < 1$. Then

$$\mathbb{E}_\Theta \left[\|\mathbf{H}_{t+1}\|_F^2 \mid \mathcal{F}_t \right] \le \alpha \|\mathbf{H}_t\|_F^2 + C_{0,\Theta}, \tag{16}$$

for some $C_{0,\Theta} < \infty$. Iterating this inequality gives

$$\mathbb{E}_\Theta \|\mathbf{H}_t\|_F^2 \le \alpha^t \mathbb{E}_\Theta \|\mathbf{H}_0\|_F^2 + \frac{1-\alpha^t}{1-\alpha} C_{0,\Theta},$$

which proves the claim. □

This result establishes moment control of the generated trajectories. It should not be interpreted as a universal geometric-ergodicity or total-variation mixing-time guarantee, which would require additional irreducibility and minorization assumptions. For synthetic data generation, the main implication is that long trajectories can be generated while keeping latent magnitudes controlled.

Temporal memory and oscillatory responses. The dependence on both $\mathbf{H}_t$ and $\mathbf{H}_{t-1}$ introduces delayed feedback. This second-order structure enables temporal autocorrelation, phase-lagged responses, and oscillatory patterns. Around a reference state $\bar{\mathbf{z}}$, the conditional mean dynamics admit the local linearization

$$\delta \mathbf{z}_{t+1} \approx \mathbf{J}(\bar{\mathbf{z}}) \delta \mathbf{z}_t + \mathbf{G}(\bar{\mathbf{z}}) \boldsymbol{\epsilon}_t, \qquad \mathbf{z}_t = \begin{bmatrix} \mathrm{vec}(\mathbf{H}_t) \\ \mathrm{vec}(\mathbf{H}_{t-1}) \end{bmatrix}, \tag{17}$$

where

$$\mathbf{J}(\bar{\mathbf{z}}) = \begin{bmatrix} a\mathbf{I} + D_1 \boldsymbol{\mu}(\bar{\mathbf{z}}) & D_2 \boldsymbol{\mu}(\bar{\mathbf{z}}) \\ \mathbf{I} & \mathbf{0} \end{bmatrix}, \tag{18}$$

and $\mathbf{G}(\bar{\mathbf{z}})$ denotes the diffusion amplitude evaluated at the same reference state. Here $D_1\boldsymbol{\mu}$ and $D_2\boldsymbol{\mu}$ are the Jacobian blocks of the vectorized drift with respect to $\mathrm{vec}(\mathbf{H}_t)$ and $\mathrm{vec}(\mathbf{H}_{t-1})$.

Proposition B.5 (Local mechanism for quasi-periodic responses)**.** *Suppose* $\mathbf{J}(\bar{\mathbf{z}})$ *has a stable complex conjugate eigenvalue pair*

$$\lambda_{1,2} = r e^{\pm i\theta}, \qquad 0 < r < 1, \qquad \theta \in (0, \pi).$$

If the corresponding mode is excited by the stochastic innovation and observed by the output projection, then the observed sequence has an enhanced resonant response near

$$f_{\mathrm{osc}} = \frac{\theta}{2\pi}. \tag{19}$$

Proof sketch. For the frozen-diffusion linear system, the spectral response of a linear readout is governed by

$$(\mathbf{I} - e^{-i\omega} \mathbf{J}(\bar{\mathbf{z}}))^{-1}.$$

A complex eigenvalue pair $re^{\pm i\theta}$ enhances the response near angular frequencies $\pm\theta$. If this mode is both excited and observed, the output spectrum contains a resonant component near $\theta/(2\pi)$. □

This proposition identifies a local mechanism for quasi-periodic behavior rather than exact periodicity. With Gaussian excitation, damped oscillatory modes can be repeatedly stimulated and appear as oscillatory autocorrelation, spectral concentration, or phase-lagged propagation.

Graph-spectral diversity. The graph filters couple temporal feedback with graph-spectral structure. If $(\mathbf{q}_i, s_i)$ is an eigenpair of $\widehat{\mathbf{L}}$, then

$$\mathbf{T}_1\mathbf{q}_i = s_i\mathbf{q}_i, \qquad \mathbf{T}_2\mathbf{q}_i = (2s_i^2 - 1)\mathbf{q}_i. \tag{20}$$

Thus, different graph frequencies receive different first- and second-order feedback. Since $2s_i^2 - 1$ changes sign at $|s_i| = 1/\sqrt{2}$, some spectral components may behave like smoothing or diffusion modes, while others may behave like delayed-feedback or propagation modes. Although the nonlinear transformations and channel mixing prevent a global decomposition into independent graph modes, this view explains why diffusion-like, wave-like, and oscillatory patterns can arise in the synthetic data.

Nonlinear and irregular regimes. The bounded nonlinearity $\tanh(\cdot)$ allows the generator to interpolate between near-linear and saturated regimes. Small effective weights lead to mean-reverting stochastic autoregressive behavior, whereas larger effective weights can push the drift into saturation, producing threshold-like transitions, bursts, plateaus, and intermittent dynamics.

The deterministic skeleton can also be used to diagnose chaotic-like finite-time sensitivity. Removing the Gaussian excitation yields

$$\mathbf{z}_{t+1} = \mathbf{f}_\Theta(\mathbf{z}_t).$$

Let

$$\mathbf{\Phi}_T(\mathbf{z}_0) = D\mathbf{f}_\Theta(\mathbf{z}_{T-1}) \cdots D\mathbf{f}_\Theta(\mathbf{z}_0).$$

The finite-time Lyapunov diagnostic is

$$\lambda_T(\mathbf{z}_0) = \frac{1}{T} \log \left\| \mathbf{\Phi}_T(\mathbf{z}_0) \right\|_2 . \tag{21}$$

A positive λ_T indicates local exponential separation of nearby initial conditions over the finite window. Such behavior is not universal across sampled systems, but it can occur in high-gain nonlinear regimes. Since the drift is bounded and $a \in (0, 1)$, local expansion can coexist with global moment control, yielding bounded but irregular chaotic-like trajectories.

Observation, masking, and corpus-level coverage. The observed sequence is obtained through a random projection of the latent state, so latent modes may be amplified, attenuated, or hidden by the observation map. Random masking further introduces external observational corruption without changing the latent transition law, mimicking realistic settings where structured dynamics are only partially observed.

The randomization over graphs, weights, noise amplitudes, and observation maps provides corpus-level regime diversity. If $\mathcal{R}$ is a set of configurations inducing a particular spatio-temporal regime and $\nu(\mathcal{R}) = p_\mathcal{R} > 0$, then a corpus of M independently sampled systems contains at least one system from $\mathcal{R}$ with probability

$$1 - (1 - p_\mathcal{R})^M. \tag{22}$$

Thus, any regime with nonzero probability under the sampling distribution appears in sufficiently large synthetic corpora with high probability.

Summary. In summary, the generator produces a controlled yet diverse family of graph-coupled stochastic processes: (i) conditioned on each sampled system, the latent dynamics are second-order Markov with conditionally Gaussian and heteroscedastic transitions; (ii) Chebyshev graph filtering induces spectrally controlled spatial dependence and localized propagation; (iii) the dissipative coefficient and bounded nonlinearities provide moment-controlled long trajectories; (iv) the second-order recurrence enables temporal memory, delayed feedback, and quasi-periodic responses; (v) graph-spectral filtering supports diffusion-like, wave-like, and oscillatory patterns across different spectral components; (vi) nonlinear saturation and high-gain regimes can yield intermittent or chaotic-like behavior; and (vii) random projection and masking induce partial and imperfect observation. These properties align the synthetic corpus with the structural characteristics of real-world spatio-temporal forecasting tasks while keeping the generation process controlled and scalable.

C Supplementary Experimental Details

C.1 Evaluation Datasets

To evaluate downstream spatio-temporal (ST) forecasting, we curate a suite of real-world benchmarks as summarized in Table 2. These datasets are characterized by significant heterogeneity in spatial granularity (e.g., urban regions vs. individual sensors), temporal resolutions (ranging from 5-minute to hourly intervals), and diverse prediction targets. Such multifaceted diversity provides a rigorous testbed to validate the cross-task generalization and adaptability of FactoST across both short- and long-term forecasting horizons.

Table 2: List of evaluation spatio-temporal datasets.

Dataset	Category	# Nodes	Sample rate	Time span (Y/M/D)	# Time Points
PEMS03	Traffic flow	358	5 minutes	2018/09/01 – 2018/11/30	26208
PEMS04	Traffic flow	307	5 minutes	2018/01/01 – 2018/02/28	16992
PEMS07	Traffic flow	883	5 minutes	2017/05/01 – 2017/08/31	28224
PEMS08	Traffic flow	170	5 minutes	2016/07/01 – 2016/08/31	17856
PEMS-BAY	Traffic speed	325	5 minutes	2017/01/01 – 2017/06/30	52116
METR-LA	Traffic speed	207	5 minutes	2012/03/01 – 2012/06/27	34272

C.2 Evaluation Metrics

To evaluate the prediction performance, we employ two widely used regression metrics: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Let $\mathbf{Y} = \{Y_1, \ldots, Y_N\}$ denote the ground truth of the spatio-temporal data and $\hat{\mathbf{Y}} = \{\hat{Y}_1, \ldots, \hat{Y}_N\}$ represent the values predicted by the model, where N is the total number of test samples. These metrics are formally defined as follows:

$$\text{RMSE}(\mathbf{Y}, \hat{\mathbf{Y}}) = \sqrt{\frac{1}{N}\sum_{i}^{N}\left(Y_i - \hat{Y}_i\right)^2}, \quad \text{MAE}(\mathbf{Y}, \hat{\mathbf{Y}}) = \frac{1}{N}\sum_{i}^{N}\left|Y_i - \hat{Y}_i\right|, \tag{23}$$

C.3 Baselines and Implementation

To ensure fair comparisons, we evaluate all baseline models within a unified framework using standardized metrics (MAE and RMSE). Model hyperparameters are either retained at their default settings, as reported in the original literature, or optimized via grid search on the validation set. The specific implementation strategies for each baseline category are outlined below.

C.3.1 Time Series Foundation Models (TSFMs)

- **TimesFM** Das et al. (2024): Developed by Google Research, this is a large-scale, pretrained, decoder-only time series foundation model capable of high-accuracy univariate forecasting across various domains and frequencies. We utilize the official implementation (`https://github.com/google-research/timesfm`). The pretrained weights are sourced from `https://huggingface.co/google/timesfm-1.0-200m`.
- **Moirai** Liu et al. (2025a): A large-scale, pretrained, encoder-only time series foundation model introduced by Salesforce AI Research. We adapt the official implementation from `https://github.com/SalesforceAIResearch/uni2ts`, utilizing the checkpoint available at `https://huggingface.co/Salesforce/moirai-1.0-R-base`.

C.3.2 Spatio-Temporal Foundation Models (STFMs)

- **UniST** Yuan et al. (2024): A universal STFM driven by prompt learning, pretrained across multiple urban scenarios to achieve robust generalization. As the official codebase (`https://github.com/tsinghua-fib-lab/UniST`) inherently supports only fixed 6-step prediction horizons,

we retrain the model on 13 datasets from its original release to fit our 12- and 96-step forecasting scenarios.

- **OpenCity** Li et al. (2024c): A versatile STFM supporting both zero-shot and few-shot forecasting across a wide range of city-level applications. We integrate this model into our pipeline using the `Opencity-plus.pth` checkpoint.

C.4 Leave-One-Dataset-Out Cross-Domain Adaptation Analysis

A natural concern for any test-time adaptation (TTA) mechanism is the extent to which its effectiveness depends on having access to in-domain calibration signal at inference time. In our main experiments (Section 6), the ridge-regression adapter described in Section 5 is always fit using observations drawn from the target dataset itself. To probe how much of NeoST's zero-shot performance is attributable to this in-domain calibration versus the synthetically pre-trained backbone alone, we conduct a **Leave-One-Dataset-Out (LODO)** stress test. For each target dataset, we fit the TTA mapping w^*, b^* using pooled calibration observations from the remaining five real-world datasets, entirely withholding any in-domain signal from the target dataset, and then evaluate the resulting adapter on the held-out dataset. This setting simulates a strictly harder deployment scenario in which no calibration data from the target domain is available even at inference time, isolating the contribution of the synthetically pre-trained representation itself. Results are reported in Table 3.

Table 3: Transposed comparison between Leave-One-Dataset-Out (LODO) adaptation and full zero-shot forecasting (synthetic pre-training + in-domain TTA). For each dataset/horizon, the better result between LODO and Zero-Shot is shown in **bold**. The Improvement row reports the relative change (%) of Zero-Shot over LODO, computed as $(\text{LODO} - \text{ZeroShot})/\text{LODO} \times 100\%$; positive values (Zero-Shot better) are shown in **red**, negative values (LODO better) are shown in blue.

Setting	Metric	PEMS03		PEMS04		PEMS07		PEMS08		PEMS-Bay		METR-LA	
		12→12	96→96	12→12	96→96	12→12	96→96	12→12	96→96	12→12	96→96	12→12	96→96
LODO	MAE	26.26	137.25	33.02	172.07	38.75	**129.77**	28.30	221.97	**2.70**	15.08	5.39	8.82
	RMSE	38.35	176.64	46.64	214.65	53.29	**162.38**	38.77	274.87	**4.71**	19.02	8.50	14.48
Zero-Shot	MAE	**25.16**	**84.49**	**30.98**	**105.25**	**35.70**	144.18	**26.28**	**94.03**	**2.67**	**4.86**	**4.90**	**8.31**
	RMSE	**34.90**	**107.85**	**43.63**	**137.35**	**49.48**	175.65	**36.42**	**121.87**	4.80	**8.14**	**8.21**	**13.70**
Imp (%)	MAE	**4.19**	**38.44**	**6.18**	**38.83**	**7.87**	-11.11	**7.14**	**57.65**	**1.11**	**67.77**	**9.09**	**5.78**
	RMSE	**9.00**	**38.94**	**6.45**	**36.02**	**7.15**	-8.17	**6.06**	**55.66**	-1.91	**57.20**	**3.41**	**5.39**

The LODO results reveal a clear horizon-dependent pattern in how much NeoST relies on in-domain calibration. At the 12-step horizon, performance under LODO remains close to the vanilla zero-shot result, with average MAE degradations of only 1 to 3 points across datasets and near parity on PEMS-Bay (2.70 vs. 2.67), indicating that the encoder and hierarchical reasoner, trained purely on synthetic dynamics, already capture transferable spatiotemporal structure and that the adaptation module mainly sharpens an already competent forecast rather than compensating for a domain-specific representation. At the 96-step horizon the picture is more nuanced. On PEMS03, PEMS04, PEMS08, and PEMS-Bay, withholding in-domain calibration leads to a substantially larger MAE increase (for example 84.49 to 137.25 on PEMS03 and 94.03 to 221.97 on PEMS08), suggesting that at longer horizons the model depends more heavily on TTA to recalibrate for dataset-specific scale, sensor placement, and noise characteristics that cannot be inferred from the synthetic prior alone, whereas PEMS07 and METR-LA are notably more robust, with PEMS07 even improving slightly over its in-domain counterpart (144.18 to 129.77), likely because their traffic dynamics are already well covered by the pooled cross-domain calibration statistics from the remaining datasets. This asymmetry suggests that the benefit of in-domain TTA scales with how distinctive a target dataset's long-range statistics are relative to the rest of the benchmark suite rather than being a uniform requirement. Taken together, these results indicate that NeoST's synthetically pre-trained backbone generalizes well even without target-domain signal, especially at short horizons, while the lightweight ridge-regression TTA module provides an increasingly important dataset-specific correction as the horizon lengthens, positioning TTA as a complementary mechanism for closing the residual sim-to-real gap discussed in Appendix D.1 rather than a crutch for a biased backbone.

D Further Discussion

D.1 Limitations & Future Improvements

While NeoST demonstrates strong zero-shot generalization and inference efficiency across diverse real-world benchmarks, we acknowledge several methodological and architectural boundaries in the current framework that provide important avenues for future research.

Sim-to-Real Gap in Topology and Noise Distributions. Our synthetic data generation engine relies on Erdős-Rényi and Poisson degree models to sample spatial graph topologies. Furthermore, the temporal evolution utilizes a standard Gaussian noise formulation, $\epsilon_t \sim \mathcal{N}(0, I)$, for stochastic excitation. However, real-world complex networks (e.g., urban traffic infrastructure, epidemiological contact networks) frequently exhibit scale-free or small-world topological properties. Similarly, exogenous shocks in physical systems often manifest as heavy-tailed distributions rather than ideal Gaussian perturbations. These idealized assumptions may introduce a "sim-to-real" distribution gap, potentially constraining the model's robustness against extreme anomalies or black-swan events. Future iterations of the generative engine should incorporate scale-free network generation algorithms (e.g., the Barabási-Albert model) and non-Gaussian jump processes (e.g., Lévy flights) to more accurately reflect real-world macroscopic statistics.

Computational Scalability of Test-Time Adaptation. To mitigate distribution shifts during inference, NeoST employs a Ridge Regression-based Test-Time Adaptation (TTA) mechanism. This requires computing the closed-form solution $w^* = (Z_{obs}^\top Z_{obs} + \lambda I)^{-1} Z_{obs}^\top \tilde{v}_{obs}$. While highly effective for the moderately sized spatial graphs evaluated in our benchmarks, the matrix inversion step incurs a cubic computational complexity. For ultra-large-scale spatial networks (e.g., nationwide power grids or comprehensive city-level traffic systems with tens of thousands of nodes), this online adaptation step may introduce latency bottlenecks, potentially offsetting the high inference efficiency inherent to our latent-space reasoning paradigm. Future work could explore iterative solvers, gradient-based meta-learning, or low-rank approximations to accelerate this process at scale.

Lack of Uncertainty Quantification. The current NeoST architecture is designed exclusively for deterministic point forecasting, optimizing for reconstruction fidelity using the Frobenius norm in the objective function. In many critical spatio-temporal applications, such as extreme weather forecasting, epidemiological modeling, or autonomous fleet routing, point predictions are insufficient; decision-makers require reliable uncertainty quantification and confidence intervals to assess risk. The absence of native probabilistic forecasting capabilities limits the model's immediate utility in risk-sensitive scenarios. Future work will extend the Output Adapter to predict the parameters of continuous probability distributions or integrate quantile regression mechanisms, thereby enabling robust and calibrated uncertainty estimation for downstream tasks.

D.2 Broader Impacts

The development of NeoST, a Spatio-Temporal Foundation Model pre-trained exclusively on synthetic data, offers significant societal benefits by mitigating the demographic, geographic, and socioeconomic biases typically encoded in real-world sensor deployments. Furthermore, its efficient architecture promotes environmentally sustainable deployment compared to massive foundation models. However, deploying such models in safety-critical domains such as urban traffic management and disaster response, carries inherent risks. The "sim-to-real" distribution gap may cause the model to behave unpredictably during extreme, heavy-tailed anomalies or black-swan events, potentially endangering public safety or infrastructure security. Additionally, while our pre-training data involves no human subjects, the model's capacity to analyze sequential mobility data could be misused for mass surveillance or discriminatory practices if fine-tuned on sensitive, unanonymized tracking data. To ensure safe and equitable use, we urge practitioners to integrate robust uncertainty quantification prior to deployment, strictly adhere to human rights and data privacy regulations, and validate the model rigorously against domain-specific, real-world conditions.