WaveDiST: A Wavelet Diffusion Transformer for Spatio-Temporal Estimation on Unobserved Locations

Huiling Qin[1], **Yuanxun Li**[2], **Weijia Jia**[1,3*],
[1]Beijing Normal University
[2]King Soft
[3]Beijing Normal-Hong Kong Baptist University
qinhuiling@bnu.edu.cn, genkunabe@gmail.com, jiawj@bnu.edu.cn

Abstract

Spatio-temporal estimation plays a vital role in numerous scientific and engineering tasks, particularly for novel or unobserved locations lacking historical references. Many areas remain unobserved by sensors due to their non-core location or pending development status. The states of these areas can only be estimated through similar nodes in the geospace, rather than through historical data with temporal trends. Estimating these unobserved node states is crucial for city-wide spatio-temporal sensing and urban development, extending beyond simple point or block data imputation. In this study, we introduce a diffusion point process in high-frequency space to develop a robust spatio-temporal diffusion transformer for urban estimation where partial historical reference data is lacking. Our approach decomposes spatio-temporal data into high and low-frequency components through wavelet transform, and trains a diffusion model of spatial temporal data with a transformer that operates on high frequency signals. We incorporate low-frequency signals as diffusion conditions in the transformer architecture to capture overall spatio-temporal profiles and gradual trends. To enhance the learning of each step, we design an diffusion model featuring a spatio-temporal attention module that adaptively captures interdependencies between time and space. Extensive experiments across diverse domains including traffic, economics, and environment demonstrate that our method significantly outperforms state-of-the-art baselines.

Introduction

Spatio-temporal estimation of unknown areas plays a central role in maintaining urban stability and facilitating urban development. For example, local transport authorities rely heavily on road-based sensors (Marisca, Cini, and Alippi 2022), to obtain real-time traffic information for daily operations. However, due to the high cost of sensor deployment, authorities can only monitor major urban arteries (as illustrated in Figure 1), leaving many high-traffic ordinary road sections without monitoring and unable to respond quickly to anomalies. Therefore, estimating road conditions for these unobserved yet crucial road sections that lack historical reference data is essential for ensuring travel safety. Similarly, in urban commercial development, numerous untapped areas present valuable opportunities. For instance, planning the layout of offline retail stores like Sam's Club or developing commercial complexes such as Galeries Lafayette requires extensive field research and demand analysis (Qin et al. 2021a). Thus, estimating demand for these unestablished areas is crucial for rapidly assessing their commercial potential. However, estimating such unobserved areas without historical reference data presents numerous challenges:

- **Data Sparsity:** Urban sensors, such as traffic cameras and air quality detectors, can only monitor a small portion of urban areas. These sensors are primarily deployed in key urban areas rather than providing city-wide coverage, resulting in sparse observations (Zhuang et al. 2022) of urban states from source data.
- **Sampling Noise:** In real-world scenarios, urban sensor data often contains significant noise due to human interference (Qin et al. 2021b) (e.g., vehicle sensors affected by driver operations, as shown in Figure 1 Bottom) or sensor communication issues (e.g., signal propagation disrupted by thunderstorms). This leads to potentially unreliable data with high variance.
- **Lack of Historical Reference:** Unlike traditional missing data imputation, urban area estimation faces a unique challenge: unobserved or blank areas have no historical data available for reference. This absence of historical

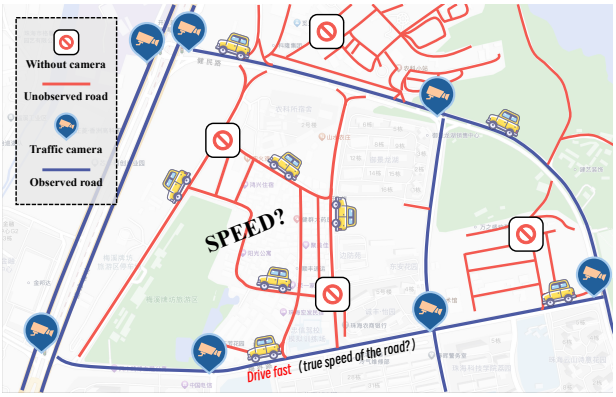


Figure 1: Limited and noisy urban traffic state monitoring.

data makes it impossible for models to leverage temporal trends for state estimation.

- **Complex Spatio-temporal Patterns:**Urban states are heavily influenced by urban rhythms, human activities, and development patterns, which create highly heterogeneous spatio-temporal patterns in urban sensor data (Li and Zhu 2021), making it particularly challenging to infer unknown area states from partially observed data.

To address the aforementioned challenges, we propose a Wavelet Transform-based Spatial-Temporal Diffusion Transformer for urban area estimation with partially observed data. Through wavelet transform, we decompose the original time series signal into multiple high-frequency signals and one low-frequency signal (Michau, Frusque, and Fink 2022; Zhang et al. 2023; Yang et al. 2023), capturing both overall trends and detailed variations. As sensor sampling noise typically resides in high-frequency signals, we introduce the denoising diffusion probabilistic model (Nichol and Dhariwal 2021) in the high-frequency latent space post wavelet decomposition. This enables robust estimation of unknown areas while simultaneously denoising observed data. The low-frequency signal serves as a condition for the probability distribution, providing both an overall spatio-temporal profile and slowly evolving trends for unknown area reconstruction. To enhance step-wise learning in the reverse diffusion process, we develop an elaborate spatial-temporal transformer. Since unknown areas can only be estimated using observable data from neighboring regions, we design a novel structure that adaptively captures interdependencies between time and space to obtain fused spatio-temporal representations. The transformer further incorporates the low-frequency signal from wavelet decomposition into its scale-shift module, improving its ability to capture general trends and ultimately achieving estimation from sparse to complete urban states. Our main contributions are summarized as follows:

- We propose a wavelet-enhanced diffusion process that decomposes spatio-temporal signals and applies diffusion modeling selectively. By separating the denoising process across frequency domains, it overcomes noise interference and trend instability in traditional full-signal diffusion approaches.
- We develop a spatio-temporal diffusion transformer that integrates multi-scale wavelet decomposition with adaptive attention mechanisms. By incorporating conditional diffusion blocks and specialized spatio-temporal attention modules, it captures complex interdependencies across frequency scales and spatial-temporal dimensions.
- Extensive experiments across diverse urban domains demonstrate that WaveDiST significantly outperforms state-of-the-art baselines. The model shows robust performance under various missing rates and excels with limited training samples.

Problem Statement

Definition 1. **Adjacency matrix.** We consider the urban sensor network as a graph $G = \{R, A\}$, where sensors are considered as the nodes V in graph G, and A is the adjacency matrix representing the connectivity between sensors.

Definition 2. **Partially observed spatial temporal data.** We denote X as the spatial temporal data extracted from sensors on the urban sensor network G. Let x_t^v be the record of a sensor $v \in V$ at time step t, and we write the records for all sensors in V over the time period $[t_0 : t_n]$ as a matrix $X_{t_0:t_n} = \left[x_{t_0:t_n}^v | v \in V\right]$.

Definition 3. **Observability Mask.** We define a mask matrix M, with $M = \left[m^v | v \in V\right]$ denoting the observability of sensors over the all time period $[t_0 : t_n]$. We set $m^v = 1$ if x^v is observed, and $m^v = 0$ if x^v is unobserved.

Our goal is to reliably estimate all the unobserved entries in spatial temporal data $X_{t_0:t_n}$ by constructing a filled matrix $\hat{X}_{t_0:t_n}$ of a given sensor network G and time interval $[t_0 : t_n]$, We denote our urban area estimation task as a mapping: $\mathcal{F}(X, M) \mapsto [\hat{X}]$.

Wavelet-Transform based-Diffusion Process

Urban sensing data inherently contains two primary noise sources: sampling bias introduced during data collection by urban sensors, and missing data noise where unobserved areas can be treated as a special form of noisy observations. To handle such limited and noisy data, we incorporate the diffusion process to enhance the model's robustness and expressiveness. During training, the diffusion process learns the mapping from noise to clean signals, enabling the model to distinguish inherent data patterns from noise. This allows the model to preserve dominant patterns while suppressing random noise during data generation.

According to signal processing theory, noise in time series typically exhibits broadband characteristics across the frequency spectrum. For a spatio-temporal sequence x_i with additive noise ϵ, J-level wavelet decomposition $\mathcal{W} : x_i \mapsto \{x_i^{h_1}, x_i^{h_2}, \ldots, x_i^{h_J}, x_i^l\}$ naturally separates noise from signal, where $x_i^{h_j}$ $(j = 1, 2, \ldots, J)$ are high-frequency components at level j, and x_i^l is the low-frequency component at the coarsest level J The effectiveness of wavelet-based denoising stems from asymmetric energy distribution between signal and noise. For white noise with uniform power spectral density, the multi-scale structure results in:

$$\mathbb{E}\left[\sum_{j=1}^{J} \|\mathcal{W}^{h_j}(\epsilon_i)\|_2^2\right] > \mathbb{E}[\|\mathcal{W}^{l}(\epsilon_i)\|_2^2] \tag{1}$$

This occurs because high-frequency subbands collectively span a broader frequency range than the single low-frequency subband, naturally accumulating more noise energy while signal trends are predominantly preserved in low-frequency components. As noted by Donoho and Johnstone (Donoho and Johnstone 1994), this asymmetry forms the foundation for wavelet-based diffusion process.

High-Frequency Space Conditional Diffusion. Based on the signal processing theory, we design a conditional diffusion process that operates exclusively in high-frequency

space while using low-frequency components as stable conditioning information. Typically, diffusion models consist of two Markov chain processes with step length K: a forward (noising) process that gradually adds noise to the original data X_0, and a reverse (denoising) process that recovers the original distribution from the noisy state.

We decompose the original spatio-temporal data into high-frequency components $x^h = \{x^{h_1}, x^{h_2}, \ldots, x^{h_J}\}$ and a low-frequency component x^l through the wavelet transform operator $\mathcal{W}$. In the forward process, noise is sequentially added only to high-frequency components, where the incremental noising is mathematically defined as a Markov chain that transitions the high-frequency data from its original state $x_0^{h_j}$ to a noised state $x_K^{h_j}$:

$$\begin{aligned} q(x_{1:K}^{h_j}|x_0^{h_j}) &= \prod_{k=1}^{K} q(x_k^{h_j}|x_{k-1}^{h_j}) \\ q(x_k^{h_j}|x_{k-1}^{h_j}) &= \mathcal{N}(x_k^{h_j}; \sqrt{1-\beta_k}x_{k-1}^{h_j}, \beta_k \mathbf{I}) \end{aligned} \tag{2}$$

where β_k represents the noise schedule. For practical gradient-based optimization, a reparameterization approach is employed such that $x_k^{h_j} = \sqrt{\bar{\alpha}_k}x_0^{h_j} + \sqrt{1-\bar{\alpha}_k}\epsilon$, and $\bar{\alpha}_k = \prod_{i=1}^{k}(1-\beta_i)$. It directly targets the noise-dominated frequency bands while preserving global trends.

In the reverse process, we employ the conditional diffusion probabilistic model to reconstruct the original spatio-temporal data from the noise-corrupted high-frequency state $\tilde{x}_K^h$. This denoising process is guided by both the previous noise data and the low-frequency conditional information x^l, which incorporates domain-specific constraints and provides essential global trend information. The reverse process directly estimates the clean spatio-temporal data $\hat{X}$ rather than denoising the high-frequency components:

$$\begin{aligned} p_\theta(\hat{X}_{0:K-1}|\tilde{x}_K^h, x^l) &= \prod_{k=1}^{K} p_\theta(\hat{X}_{k-1}|\tilde{x}_k^h, x^l) \\ p_\theta(\hat{X}_{k-1}|\tilde{x}_k^h, x^l) &:= \mathcal{N}(\hat{X}_{k-1}; \mu_\theta(\tilde{x}_k^h, k, x^l), \sigma_\theta(\tilde{x}_k^h, k, x^l)^2 \mathbf{I}) \end{aligned} \tag{3}$$

where $\mu_\theta(\tilde{x}_k^h, k, x^l)$ and $\sigma_\theta(\tilde{x}_k^h, k, x^l)$ represent the learnable mean and variance of the reverse process at each step k, respectively. Both are parameterized by θ and modulated by the low-frequency conditional x^l.

The low-frequency conditioning x^l provides two critical advantages: trend preservation by constraining the reconstruction to follow global patterns, and convergence acceleration by reducing the search space for reconstruction. This ensures $\text{Var}[\hat{X}|x^l] \leq \text{Var}[\hat{X}]$, providing more stable optimization compared to full-signal diffusion.

WaveDiST Architecture

In this section, we introduce WaveDiST, a wavelet-enhanced diffusion model for spatio-temporal estimation at unobserved locations, as illustrated in Figure 3. The framework first applies wavelet transform to decompose time series into high-frequency components capturing detailed variations and one low-frequency component representing the underlying trend. We conduct the diffusion process in the high-frequency space. The noisy high-frequency components are

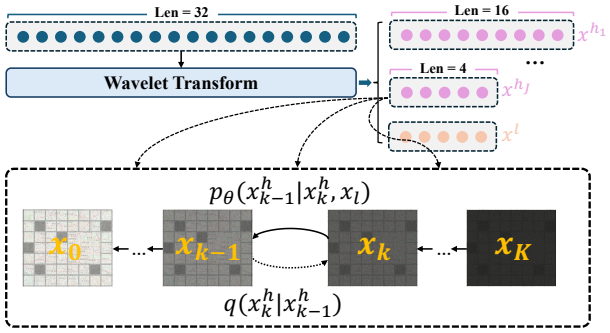


Figure 2: The overview of the wavelet transform-based diffusion process.

then fed into WaveDiST, which employs spatial-temporal fusion attention mechanisms to learn spatio-temporal features and generate denoised complete time-series data. Meanwhile, the low-frequency component serves as conditional guidance to ensure stable urban state estimation.

Wavelet-Based Feature Representation.

Spatio-Temporal Signal Encoding. The Transformer encoder receives two types of inputs: multi-scale high-frequency noisy embeddings and spatio-temporal positional embeddings. The noisy embeddings are derived from high-frequency components $x^h = \{x^{h_1}, \cdots, x^{h_n}\}$ obtained from wavelet decomposition, where each component captures signal information at different frequency ranges. For each time series, we independently perform wavelet decomposition and add standard Gaussian noise $\mathcal{N}(0, I)$ to the high-frequency components in the forward diffusion process. Since these components have inconsistent lengths, we employ dimensional transformation to scale them to the original sequence length L, then concatenate them along the channel dimension, resulting in $E_h \in \mathbb{R}^{N \times L \times D}$ where D is the number of high-frequency components. The low-frequency component x_l serves as conditional information, detailed in the next subsection.

For positional encoding, we use sinusoidal encoding for temporal sequences and learnable encoding for spatial nodes. These are combined via element-wise addition with the noisy embeddings before input to Transformer blocks. Detailed wavelet implementation and spatio-temporal positional encoding are provided in supplementary materials.

Conditional Information Encoding. The conditioning mechanism integrates two complementary signals: denoising step embeddings that capture the current noise level, and low-frequency embeddings that preserve global temporal trends. To encode the denoising steps into a continuous representation, we employ sinusoidal positional embedding followed by a two-layer MLP:

$$\begin{aligned} freq &= \exp(-\log(K_{max}) \cdot 2i/d) \quad \text{for } i \in [0, d/2-1] \\ k_{emb} &= MLP([\cos(k \cdot freq); \sin(k \cdot freq)]) \end{aligned} \tag{4}$$

where K_{max} is the maximum period and d is the embedding dimension. The $freq$ represents frequency factors at different scales, enabling the model to capture patterns at different

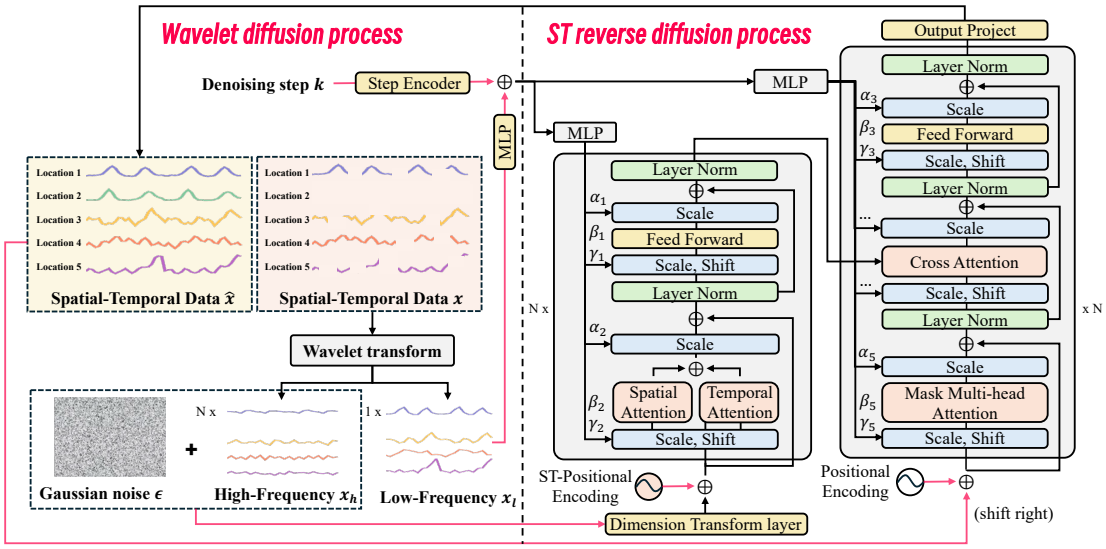


Figure 3: Overall Framework of WaveDiST.

temporal scales. The obtained frequency embedding k_{freq} is then transformed through an MLP to get the final timestep representation k_{emb}. For the low-frequency component x_l, we use an MLP layer to transform its dimensions to match those of the denoising step embedding. The two embeddings are then added element-wise to obtain the final condition embedding, which provides both temporal scale information from the denoising steps and global trend information from the low-frequency components.

Spatio-Temporal Attention Module.

To learn the dynamics of different spatio-temporal nodes and obtain effective spatio-temporal state representations, we design a spatio-temporal attention module in the Diffusion Transformer. Following the initial dimension scaling, we process $E_h \in \mathbb{R}^{L\times N\times D}$ through two specialized self-attention modules that capture temporal and spatial dimensions independently, where L is sequence length, N is number of nodes, and D is embedding dimension.

Temporal Attention. Considering the presence of sampling noise and missing temporal samples in spatio-temporal data, modeling based on the temporal dependencies of observable areas enables the model to learn how urban states evolve over time, which helps correct inaccurate information introduced during sensor sampling. Using the scale-shift output h_{ss}, we utilize the self-attention mechanism to model pairwise temporal interactions between time steps:

$$H_t = A_t(h_{ss}W_{vt}); A_t = \text{softmax}(\frac{h_{ss}W_{qt}(h_{ss}W_{kt})^T}{\sqrt{d_k}}) \quad (5)$$

the computation is performed independently for each node across the temporal dimension.

Spatial Attention. Meanwhile, for unobserved spatial nodes, estimation cannot rely on historical temporal information. Therefore, it is reasonable to learn node transfer patterns by incorporating spatial relationships between nodes. We integrate graph attention mechanism into Transformer, where the adjacency matrix $\mathbf{A}$ defines the direction of information flow between nodes (Spatial graph construction is detailed in supplementary materials.), and weights are computed based on pairwise correlations of node embeddings:

$$H_s = A_s(h_{ss}W_{vs}); A_s = \text{softmax}(\frac{h_{ss}W_{qs}(h_{ss}W_{ks})^T}{\sqrt{d_k}} \odot \mathbf{A}) \quad (6)$$

where $\odot$ denotes element-wise multiplication. The adjacency matrix $\mathbf{A}$ not only defines the connectivity between nodes but also helps mask invalid attention weights by setting them to negative infinity, ensuring that information only flows along meaningful spatial connections.

In the encoder of diffusion transformer, we fuse the outputs from temporal self-attention and graph attention mechanisms through learnable projections:

$$H_{st} = W_h(\text{ReLU}(W_tH_t) + \text{ReLU}(W_sH_s)) \quad (7)$$

This fusion ensures that the encoder's output embeddings capture both temporal and spatial characteristics. For the decoder performing sequence-to-sequence estimation, after encoding the time series through masked multi-head attention, it utilizes the spatio-temporal embeddings from the encoder as keys and values to guide data generation.

Conditional Diffusion Transformer Block.

To enhance the efficiency of reverse diffusion in spatio-temporal scenarios, we adopt the conditional diffusion transformer as the model's foundation, as shown in the gray block in Figure 3. In addition to noised high-frequency inputs, diffusion models commonly incorporate various conditional information such as noise timesteps t and periodic signals. Following the diffusion transformer architecture, we adopt the adaLN-Zero initialization strategy (Peebles and Xie 2023), which has been proven effective in ResNet-based

models by initializing each residual block as the identity function. This is achieved by zero-initializing the final layer in each block prior to any residual connections. Considering that spatio-temporal data collected from urban sensors is influenced by urban development and daily rhythms, often exhibiting gradual urban evolution patterns, we extend the adaLN-Zero conditioning mechanism. Besides the standard time embedding t, we additionally incorporate the low-frequency signal x_l obtained from wavelet transform as a condition. These two conditions are used to generate the adaptive Layer Normalization (adaLN) parameters γ, β and the scaling parameter α through:

$$\begin{aligned} [\gamma, \beta, \alpha] &= \text{MLP}([t; x_l]) \\ h_b &= h_{ss} + \alpha \cdot \text{Block}(\gamma \cdot \text{LayerNorm}(h) + \beta) \end{aligned} \tag{8}$$

where $[t; x_l]$ denotes the concatenation of time embedding and low-frequency signal, MLP projects the concatenated conditions to a shared embedding space. Here, h represents the intermediate features within the transformer block. The Block operation can be a self-attention layer, a spatio-temporal attention layer, or a feed-forward network, depending on its position in the architecture. The scaling parameters α and β are applied immediately before any residual connections within the DiT block, as illustrated in Figure 3, ensuring the stability of overall trends in the denoised data.

Estimation Process.

During training, the model first employs wavelet transform to decompose spatio-temporal data into multiple high-frequency components and one low-frequency component. Then, WaveDiST uses a scaling layer to transform the high-frequency components of inconsistent lengths to match the original sequence length, thereby obtaining multiple high-frequency latent embeddings as the spatio-temporal input embedding for the diffusion model. Subsequently, WaveDiST estimates the denoised data $\hat{x}$ in the reverse process based on the low-frequency component x_l and diffusion timestep t. Considering that spatio-temporal data estimation is essentially a reconstruction task rather than a generation task, we directly estimate the values of original data instead of estimating Gaussian noise. Meanwhile, to create supervised samples, we randomly drop a proportion of observed sensor nodes during model training. This peration ensures the generalizability of the imputation model. We use a masking indicator M_{drop} to denote these locations where the masked values are marked as ones and others as zeros. Note that the supervision loss is only calculated on these manually droped and observed nodes, and models are forbidden to have access to the masked missing points used for evaluation. Therefore, the reconstruction loss of our model is:

$$\mathcal{L} = M \odot \|X - \hat{X}\|_2^2 + M_{drop} \odot \|X - \hat{X}\|_2^2 \tag{9}$$

Experiments

In this section, We evaluate WaveDiST on datasets from different urban scenarios to answer the following questions:

- **RQ1:** Can the unobserved location data estimated by WaveDiST provide superior accuracy under various scenarios compared to several state-of-the-art baselines?
- **RQ2:** How does WaveDiST perform in terms of robustness across different missing rates?
- **RQ3:** What are the impacts of diffusion process, wavelet transform, and spatio-temporal attention mechanism on the performance of WaveDiST?
- **RQ4:** How does WaveDiST perform in terms of spatio-temporal continuity and denoising of observed data?

Experimental setups

Baselines. Our baselines cover a broad collection of relevant methods, which can be categorized into 3 types: **Temporal-based:** Transformer(Vaswani 2017), TimeMixer(Wang et al. 2024b), TimesNet(Wu et al. 2023), CSDI(Tashiro et al. 2021), and SAITS(Du, Cote, and Liu 2023). **Spatial-Temporal based:** ImputeFormer(Nie et al. 2024), FreTS(Yi et al. 2024), and STEMGNN(Cao et al. 2020). **Variants of WaveDiST:** *w/o diffusion*, which examines the contribution of the diffusion process; *w/o encoder*, which adopts the decoder-only architecture instead of the encoder-decoder architecture to evaluate the impact of spatio-temporal encoding; *w/o wavelet*, which directly applies diffusion to the original data without decomposition to verify the enhancement provided by wavelet transform; *w/o condition*, which drops the condition from the reverse diffusion process.

Setups. We evaluate on 5 benchmark datasets from various domains: Jinan Taxi, Jinhua Sales, Beijing Air, METR-LA, PEMS08, which have been widely adopted for evaluating spatio-temporal imputation performance. Detailed introduction of implementation and datasets can be found in supplementary materials.

Experiment Results

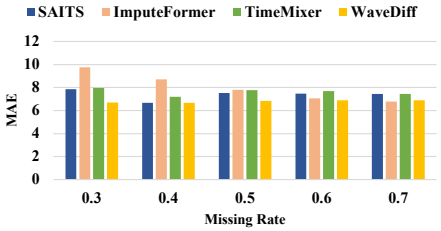
Overall Performance (RQ1). Table 1 summarizes the empirical performance of various baseline methods across five urban spatio-temporal datasets. The results demonstrate that WaveDiST consistently achieves superior performance across all scenarios. Time series-based baselines struggle to capture spatial transfer patterns without historical node data, while spatio-temporal baselines face challenges with noise and missing values. WaveDiST addresses these issues by incorporating diffusion processes and wavelet transforms to better capture spatio-temporal variation patterns.

WaveDiST shows significant improvements across different datasets: 9.8% MAE improvement over STEMGNN on Jinan Taxi, 20.9% improvement over STEMGNN on Jinhua Sales, 2.9% improvement over FreTS on Beijing Air, 9.7% improvement over CSDI on METR-LA, and 8.3% improvement over CSDI on PEMS08. Notably, WaveDiST achieves the best MAE results across all datasets, indicating superior prediction stability. This consistent performance demonstrates WaveDiST's effectiveness in learning spatio-temporal patterns across diverse urban sensing applications.

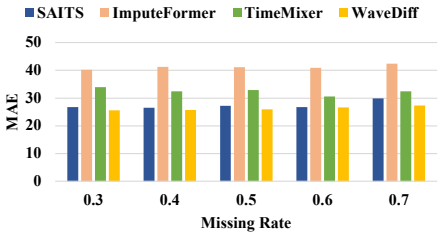
Robustness under Different Missing Rate (RQ2) To further evaluate the robustness of WaveDiST, we select three baselines (SAITS, ImputeFormer, and TimeMixer) that

Table 1: Overall accuracy. We employ "db1" as the wavelet basis and set the max decomposition level as 3 for WaveDiST.

Dataset	Metric	Transformer	SAITS	ImputeFormer	TimeMixer	TimesNet	FreTS	STEMGNN	GPVAE	CSDI	WaveDiST
Jinan Taxi	MAE	8.20	7.52	7.81	7.78	8.69	8.43	6.96	7.38	8.97	**6.84**
	RMSE	12.64	11.72	11.71	10.12	12.15	12.69	11.00	10.19	13.13	**9.34**
Jinhua Sales	MAE	534	427	463	587	892	1,530	422	626	454	**334**
	RMSE	12,670	12,661	12,680	12,686	13,517	17,747	**6,087**	12,128	10,943	6,159
Beijing Air	MAE	39.97	27.27	41.09	32.88	27.93	26.75	27.35	27.69	35.50	**25.97**
	RMSE	60.10	41.66	64.45	47.15	42.46	41.16	40.84	42.34	58.51	**38.24**
METR-LA	MAE	33.02	34.63	8.27	11.90	36.30	35.45	31.69	32.59	7.84	**7.08**
	RMSE	50.00	49.12	14.31	15.38	47.13	47.29	42.84	44.49	12.58	**10.20**
PEMS08	MAE	90.60	90.30	94.30	50.44	55.19	54.27	94.30	55.10	49.47	**46.31**
	RMSE	117.60	118.20	117.60	71.40	94.88	84.94	125.50	77.10	69.93	**65.40**



(a) Jinan Taxi



(b) Beijing Air

Figure 4: The performance under different missing rate for two scenarios

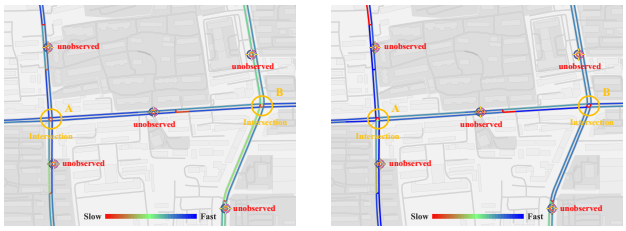
demonstrate balanced performance in Table 1 for comparison. By randomly masking all data from certain spatial locations, we gradually increase the missing rate of spatial nodes from 30% to 70%. As shown in Figure 4, while WaveDiST's estimation accuracy decreases with higher missing rates, the degradation is limited to only 10%. Moreover, even with a 70% missing rate, WaveDiST achieves comparable or better estimation performance than baselines at 30% missing rate. In contrast, the baselines, affected by different data distributions, show unstable performance across various missing rates, sometimes exhibiting lower estimation accuracy at lower missing rates. This is because when entire spatial nodes are missing, the ability to effectively transfer information from neighboring nodes becomes crucial. WaveDiST enhances its robustness across different missing rates by combining spatio-temporal information through attention with the denoising capability of the diffusion model.

Table 2: Evaluation results of WaveDiST's variations

Variants	Jinan Taxi		Sales		Beijing Air	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
w/o diffusion	7.34	9.88	349	6,267	29.79	42.65
w/o encoder	7.07	9.52	352	6,266	39.81	53.89
w/o wavelet	7.18	9.56	356	6,156	35.80	48.89
w/o condition	7.65	10.23	431	5,699	38.37	52.12
wavelet db1	6.83	9.34	334	6,159	25.97	38.23
wavelet db2	6.67	9.22	383	5,696	28.61	40.82
wavelet db4	6.97	9.32	382	5,687	28.33	42.11

Ablation Study (RQ3) Further dissection of WaveDiST performance through ablation studies emphasizes the importance of each component, as shown in Table 1. Without the diffusion process, the significant performance degradation of WaveDiST highlights the crucial role of diffusion in improving estimation accuracy for noisy data. Meanwhile, without wavelet transform or without low-frequency condition guidance, WaveDiST struggles to capture explicit overall trends and detailed variation patterns in spatio-temporal data, making it more challenging for the model to learn complex spatio-temporal patterns, resulting in reduced estimation accuracy across multiple scenarios. We also compare the commonly used decoder-only architecture with our adopted encoder-decoder architecture; the decoder-only architecture exhibits some degradation in estimation accuracy for unknown spatial nodes due to its lack of spatio-temporal representation learning. Finally, to test the influence of the wavelet basis, we test three different bases to observe the accuracy. Results show that WaveDiST is not sensitive to the choice of wavelet basis.

Spatial-Temporal Analysis (RQ4) Figure 5 demonstrates a real-world example, showing speed estimation results and actual observations in a small area of Jinan at 2 PM. Comparing the estimations and true values of unobservable roads in (a) and (b), we find that WaveDiST can effectively estimate traffic conditions for both intersections and road segments within the traffic network. Drivers typically reduce their speed when passing through intersections to ensure safety, which is observable in the Figure 5. At intersection A and B, where vehicles frequently start and stop, the traffic



(a) Estimated Traffic Data (b) Original Traffic Data

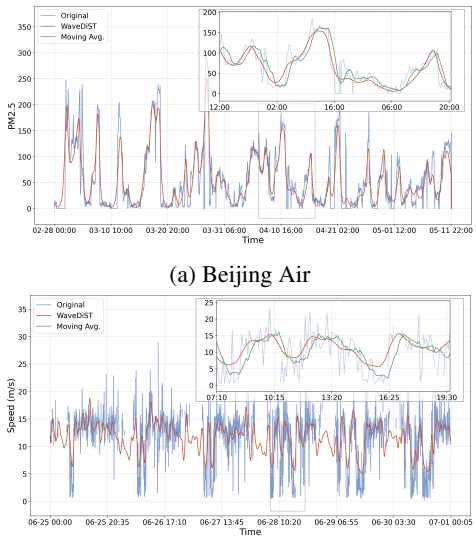
Figure 5: Spatial comparison of traffic speeds in Jinan

speed observations show distinct patterns that deviate from normal road conditions. WaveDiST demonstrates strong performance in estimating such traffic states caused by driving behaviors. Additionally, adjacent roads connected by intersections exhibit continuity in their traffic states. In this regard, WaveDiST achieves high-accuracy estimation while maintaining spatial continuity by integrating spatial relationships among different roads.

Figure 6a compares WaveDiST estimations with actual PM2.5 values at an unobservable station over two months. The data shows significant fluctuations including sudden peaks from smog events, which typically challenge the estimation. By combining diffusion with spatiotemporal modeling, WaveDiST effectively distinguishes genuine patterns from noise, providing accurate estimations during both gradual changes and abrupt external events. Figure 6b shows WaveDiST estimations versus actual traffic speeds for an unobservable road segment over one week. The original Jinan Taxi data, exhibits severe fluctuations due to random driver behaviors—which represent observation noise. WaveDiST provides smoother, more continuous estimations while capturing true temporal patterns (similar to the green smoothed line), effectively eliminating unreliable noise that could impact traffic control and optimization decisions.

Related Work

Spatial-Temporal Estimation. Spatiotemporal data imputation faces challenges in capturing complex dependencies and handling extensive missing data, with real-world datasets like Uber Movement reaching 85% missing rates (Nie et al. 2024). Early statistical approaches explored tensor properties (Yu, Rao, and Dhillon 2016) but struggled with complex dependencies due to limited model capacity. Deep learning approaches demonstrated superior performance through architectural innovations. Initial temporal modeling used bidirectional RNNs in BRITS (Cao et al. 2018), enhanced by SAITS (Du, Cote, and Liu 2023) with masked self-attention mechanisms. Transformer-based architectures evolved with ImputeFormer (Nie et al. 2024) combining low-rank inductive bias with transformer expressivity. For spatial modeling, MDGCN (Liang, Zhao, and Sun 2022) pioneered graph memory networks for historical dependencies, while Wang et al. (Zhang et al. 2024) introduced learnable evolving graph structures. Recent approaches integrate spatial-temporal modeling through innovative attention mechanisms, including CrossFormer (Chen



(a) Beijing Air

(b) Jinan Taxi

Figure 6: Temporal comparison of WaveDiST estimations and actual observations for (a) PM2.5 concentrations and (b) traffic speed patterns at unobserved locations.

et al. 2023) for multi-scale temporal modeling, Autoformer (Xu et al. 2023) with decomposition architecture, and ScaleFormer (Shabani et al. 2023) incorporating multi-scale temporal features for improved accuracy and efficiency.

Diffusion Models for Time Series and Spatio-Temporal Data. Recent diffusion models have shown strong capabilities for spatio-temporal tasks. For temporal modeling, TimeGrad (Rasul et al. 2021) introduced autoregressive denoising for probabilistic forecasting, TimeDiff (Shen and Kwok 2023) proposed non-autoregressive conditional diffusion, and CSDI (Tashiro et al. 2021) developed conditional score-based models for irregular data. For spatio-temporal graphs, DiffSTG (Zhang, Wang, and Chen 2023) introduced graph-based frameworks capturing spatial-temporal dependencies, while STPP (Yuan, Ding et al. 2023), DiffUFlow (Zheng et al. 2023), and DiffTraj (Zhu et al. 2023) addressed point processes, urban flow, and trajectory generation respectively. Domain applications include MEDiC (Sharma, Dhall, and Subramanian 2023) for medical time series, D3R (Wang et al. 2024a) for anomaly detection, and PriSTI (Liu et al. 2023) for spatiotemporal imputation. Despite these advances, opportunities remain for more efficient architectures and better domain-specific constraint incorporation.

Conclusion

In this paper, we propose WaveDiST, a wavelet transform-based spatio-temporal diffusion transformer that addresses the critical challenge of estimating urban states in unobserved locations without historical reference data. By decomposing spatio-temporal signals into frequency domains and applying conditional diffusion processes in high-frequency space, our approach effectively handles sampling noise and data sparsity challenges while capturing complex

spatio-temporal interdependencies through specialized attention mechanisms. Extensive experiments across diverse urban datasets demonstrate WaveDiST's superior performance and robustness under various missing rates.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62406029, 62272050 and the grant of Beijing Normal- Hong Kong Baptist University sponsored by Guangdong Provincial Department of Education; in part by Zhuhai Science-Tech Innovation Bureau under Grant No. 2320004002772 and the Interdisciplinary Intelligence Super Computer Center of Beijing Normal University (Zhuhai).

References

Cao, D.; Wang, Y.; Duan, J.; Zhang, C.; Zhu, X.; Huang, C.; Tong, Y.; Xu, B.; Bai, J.; Tong, J.; et al. 2020. Spectral temporal graph neural network for multivariate time-series forecasting. *Advances in neural information processing systems*, 33: 17766–17778.

Cao, W.; Wang, D.; Li, J.; Zhou, H.; Li, L.; and Li, Y. 2018. BRITS: Bidirectional recurrent imputation for time series. In *Advances in Neural Information Processing Systems*, 6775–6785.

Chen, X.; Jiang, Y.; Chen, F.; and Li, T. 2023. Crossformer: Transformer Utilizing Cross-Dimension Dependency for Multivariate Time Series Forecasting. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 198–208.

Donoho, D. L.; and Johnstone, I. M. 1994. Ideal spatial adaptation by wavelet shrinkage. *biometrika*, 81(3): 425–455.

Du, W.; Cote, D.; and Liu, Y. 2023. SAITS: Self-Attention-based Imputation for Time Series. *Expert Systems with Applications*, 119619.

Li, M.; and Zhu, Z. 2021. Spatial-temporal fusion graph neural networks for traffic flow forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 4189–4196.

Liang, Y.; Zhao, Z.; and Sun, L. 2022. Memory-augmented dynamic graph convolution networks for traffic data imputation with diverse missing patterns. *Transportation Research Part C: Emerging Technologies*, 143: 103826.

Liu, M.; et al. 2023. PriSTI: A Conditional Diffusion Framework for Spatiotemporal Imputation. In *ICDE*.

Marisca, I.; Cini, A.; and Alippi, C. 2022. Learning to reconstruct missing data from spatiotemporal graphs with sparse observations. *Advances in Neural Information Processing Systems*, 35: 32069–32082.

Michau, G.; Frusque, G.; and Fink, O. 2022. Fully learnable deep wavelet transform for unsupervised monitoring of high-frequency time series. *Proceedings of the National Academy of Sciences*, 119(8): e2106598119.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, 8162–8171. PMLR.

Nie, T.; Qin, G.; Ma, W.; Mei, Y.; and Sun, J. 2024. ImputeFormer: Low Rankness-Induced Transformers for Generalizable Spatiotemporal Imputation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2260–2271.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4195–4205.

Qin, H.; Ke, S.; Yang, X.; Xu, H.; Zhan, X.; and Zheng, Y. 2021a. Robust spatio-temporal purchase prediction via deep meta learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 4312–4319.

Qin, H.; Zhan, X.; Li, Y.; Yang, X.; and Zheng, Y. 2021b. Network-wide traffic states imputation using self-interested coalitional learning. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 1370–1378.

Rasul, K.; Seward, C.; Schuster, I.; and Vollgraf, R. 2021. Autoregressive Denoising Diffusion Models for Multivariate Probabilistic Time Series Forecasting. In *ICML*.

Shabani, M. A.; Abdi, A. H.; Meng, L.; and Sylvain, T. 2023. Scaleformer: Iterative Multi-scale Refining Transformers for Time Series Forecasting. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Sharma, G.; Dhall, A.; and Subramanian, R. 2023. MEDiC: Mitigating EEG Data Scarcity via Class-conditioned Diffusion Model. In *NeurIPS*.

Shen, L.; and Kwok, J. 2023. Non-autoregressive Conditional Diffusion Models for Time Series Prediction. In *ICML*.

Tashiro, Y.; Song, J.; Song, Y.; and Ermon, S. 2021. CSDI: Conditional Score-based Diffusion Models for Probabilistic Time Series Imputation. In *NeurIPS*.

Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Wang, C.; Zhuang, Z.; Qi, Q.; Wang, J.; Wang, X.; Sun, H.; and Liao, J. 2024a. Drift doesn't matter: dynamic decomposition with diffusion reconstruction for unstable multivariate time series anomaly detection. *Advances in Neural Information Processing Systems*, 36.

Wang, S.; Wu, H.; Shi, X.; Hu, T.; Luo, H.; Ma, L.; Zhang, J. Y.; and Zhou, J. 2024b. TimeMixer: Decomposable Multiscale Mixing for Time Series Forecasting. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; and Long, M. 2023. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Xu, H.; Wu, J.; Wang, J.; and Long, M. 2023. Autoformer: Decomposition Transformers with Auto-Correlation

for Long-Term Series Forecasting. In *Advances in Neural Information Processing Systems*, 1234–1244.

Yang, F.; Li, X.; Wang, M.; Zang, H.; Pang, W.; and Wang, M. 2023. WaveForM: Graph enhanced wavelet learning for long sequence forecasting of multivariate time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 10754–10761.

Yi, K.; Zhang, Q.; Fan, W.; Wang, S.; Wang, P.; He, H.; An, N.; Lian, D.; Cao, L.; and Niu, Z. 2024. Frequency-domain MLPs are more effective learners in time series forecasting. *Advances in Neural Information Processing Systems*, 36.

Yu, H.-F.; Rao, N.; and Dhillon, I. S. 2016. Temporal regularized matrix factorization for high-dimensional time series prediction. In *Advances in Neural Information Processing Systems*, 847–855.

Yuan, Y.; Ding, J.; et al. 2023. Spatio-temporal Diffusion Point Processes. In *KDD*.

Zhang, H.; Wang, X.; and Chen, P. 2023. DiffSTG: Probabilistic Spatio-Temporal Graph Forecasting with Denoising Diffusion Models. In *KDD*.

Zhang, W.; Zhang, L.; Han, J.; Liu, H.; Fu, Y.; Zhou, J.; Mei, Y.; and Xiong, H. 2024. Irregular Traffic Time Series Forecasting Based on Asynchronous Spatio-Temporal Graph Convolutional Networks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4302–4313.

Zhang, Z.; Guo, D.; Zhou, S.; Zhang, J.; and Lin, Y. 2023. Flight trajectory prediction enabled by time-frequency wavelet transform. *Nature Communications*, 14(1): 5258.

Zheng, Y.; et al. 2023. DiffUFlow: Robust Fine-grained Urban Flow Inference with Denoising Diffusion Model. In *CIKM*.

Zhu, Y.; et al. 2023. DiffTraj: Generating GPS Trajectory with Diffusion Probabilistic Model. In *NeurIPS*.

Zhuang, D.; Wang, S.; Koutsopoulos, H.; and Zhao, J. 2022. Uncertainty quantification of sparse travel demand prediction with spatial-temporal graph neural networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4639–4647.