Coarse-to-Fine Latent Guidance: A Multi-Scale Diffusion Transformer for Traffic Flow Forecasting

Zetao Li[1] , **Silin Zhou**[1] , **Zheng Hu**[1*] , **Shimin Cai**[1*] and **Tao Zhou**[1]

[1]University of Electronic Science and Technology of China, Chengdu 610054, China
1998huzheng@gmail.com,shimincai@uestc.edu.cn

Abstract

Accurate traffic flow prediction is fundamental to Intelligent Transportation Systems (ITS). However, traffic dynamics exhibit inherent multi-scale heterogeneity, where stable global trends are often masked by stochastic local fluctuations. Existing methods struggle to reconcile these conflicting resolutions, leading to sub-optimal forecasting. To address this, we propose the **Multi-Scale Spatial-Temporal Diffusion Transformer (MS-STDT)**. Deviating from previous diffusion approaches that treat generation as a monolithic task, we reframe the forward diffusion process as an intrinsic temporal coarse-graining operation. By leveraging the data's multi-scale hierarchy as a structural anchor, we introduce a coarse-to-fine latent guidance strategy that enables the model to reconstruct stable global trends before refining fine-grained details. This ensures physical consistency and generative stability without requiring external labels. Extensive experiments across six real-world datasets confirm that MS-STDT performs the best, demonstrating significant improvements in predictive accuracy and zero-shot robustness against sensor failures. We provide code and data at https://github.com/ZetaoLiPhD/MS-STDT.

1 Introduction

Traffic flow forecasting (TFF) constitutes the cornerstone of modern Intelligent Transportation Systems (ITS), enabling critical downstream applications ranging from congestion mitigation [Lana *et al.*, 2018] and resource optimization [Liu *et al.*, 2023] to the facilitation of sustainable urban mobility [Maimaris and Papageorgiou, 2016]. While urban traffic exhibits inherent periodicity driven by daily commutes and societal routines, the underlying dynamics are deeply stochastic and non-linear [Ma *et al.*, 2020; Wang *et al.*, 2021]. Real-world flows are governed by a complex interplay of spatial-temporal dependencies, where stable global patterns are continuously modulated by volatile perturbations and external factors [Carianni and Gemma, 2025]. Consequently, achieving high-fidelity predictions remains a formidable challenge despite the apparent regularity of the data.

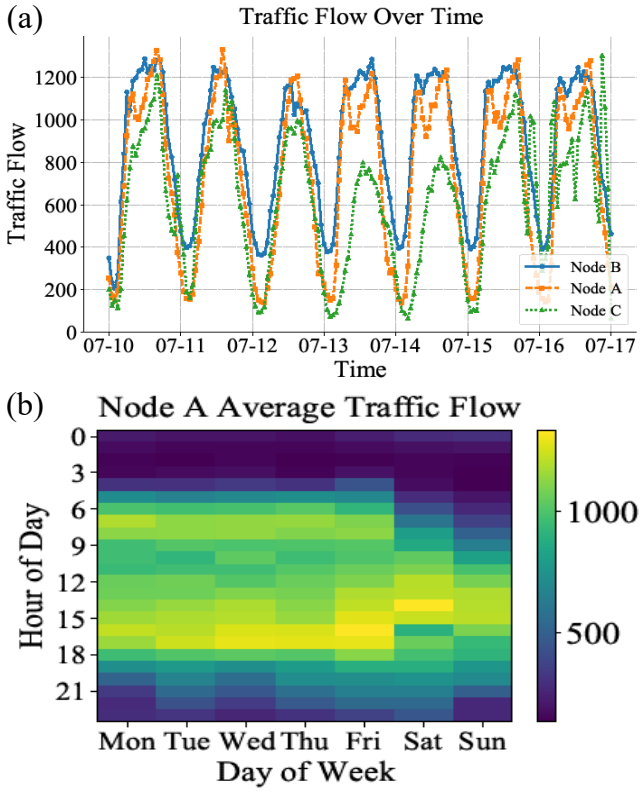

Figure 1: Multi-scale characteristics of traffic flow.

Due to the complex spatial-temporal patterns of TFF, existing methods predominantly rely on deep learning [Manibardo *et al.*, 2021]. Within this domain, Transformer-based approaches primarily focus on identifying long-range dependencies, utilizing attention mechanisms to connect distant spatial and temporal events that influence current traffic conditions [Jin *et al.*, 2023; Jiang *et al.*, 2023; Cai *et al.*, 2024]. In parallel, diffusion models address the inherent uncertainty of the system. Instead of producing a single deterministic prediction, these generative frameworks learn the full probability distribution of traffic states, allowing them to better

*Corresponding authors.

handle noise and missing data [Rühling Cachay *et al.*, 2023; Wen *et al.*, 2023; Fan *et al.*, 2024; Hu *et al.*, 2024a].

Despite these methodological advances, current models still fail to capture the natural **multi-scale physical structure** of traffic data, specifically the hierarchical dependency where stable global trends govern stochastic local fluctuations. As illustrated in Fig. 1, this hierarchy manifests as a complex superposition of distinct temporal behaviors. For the **coarse scale**, the average heatmaps (b) reveal a stable "skeleton"—robust low-frequency patterns governed by weekly routines that clearly differentiate the functional rhythms of distinct regions, such as the airport dynamics of Node A versus the arterial intersection patterns of Node C. Conversely, for the **fine scale**, the time series (a) exposes the stochastic reality: jagged high-frequency fluctuations superimposed on, yet structurally constrained by, these stable trends. However, existing methods ignore this intrinsic dependency. GNNs often aggressively over-smooth high-frequency details to capture spatial consistency [Choi *et al.*, 2022; Kong *et al.*, 2024], while monolithic diffusion models treat generation as a flat, isotropic process. By attempting to denoise all frequencies simultaneously without structural prioritization, they fail to anchor the prediction to the global context, often leading to unstable or physically inconsistent results [Fan *et al.*, 2024]. This motivates our research: accurate forecasting requires a hierarchical framework that explicitly uses stable coarse-grained trends to guide the generation of fine-grained details.

To realize this, we propose the **Multi-Scale Spatial-Temporal Diffusion Transformer (MS-STDT)**. Our framework is founded on the spectral alignment between physical data and generative physics. We frame the multi-scale hierarchy as the intrinsic resolution trajectory of the diffusion process: since the forward injection of noise functions as a coarse-graining operation that naturally dissipates fine details while retaining global trends, the reverse denoising process logically mirrors this hierarchy through a structured **coarse-to-fine reconstruction**. Consequently, the generative trajectory prioritizes recovering robust low-frequency structures in early sampling steps to establish a stable physical skeleton, effectively constraining the subsequent restoration of volatile high-frequency fluctuations. Guided by this insight, we employ hierarchical spatial-temporal encodings to represent distinct levels of granularity and introduce the **Coarse-to-Fine Latent Guidance** mechanism. This module mathematically aligns the diffusion steps with the data's physical hierarchy, using intrinsic macro-trends as a structural anchor to stabilize the generation. Finally, we incorporate a **cross-scale attention mechanism** to orchestrate the interaction between resolutions, dynamically re-weighting information flow to adaptively focus on broad trends or local variances as traffic conditions evolve.

To rigorously validate our framework, we conducted extensive experiments across six real-world datasets, benchmarking MS-STDT against 12 state-of-the-art baselines. These evaluations extend beyond standard accuracy metrics to specifically stress-test the model's zero-shot robustness against sensor failures and assess its computational efficiency. **Our contributions are summarized as follows:**

- We identify a critical **coarse-to-fine dependency** in traffic data, where stable global trends serve as structural anchors for local fluctuations. This finding challenges flat modeling paradigms, demonstrating that robust forecasting requires explicitly disentangling these conflicting multi-scale resolutions.
- We propose the **MS-STDT** featuring **Coarse-to-Fine Latent Guidance**. By leveraging intrinsic data structures to guide the diffusion process, our framework effectively reconciles global consistency with local stochasticity without requiring external labels.
- Extensive evaluations on six benchmarks confirm that MS-STDT consistently outperforms existing baselines, improving MAE by **10.1%** and CRPS by **12.8%**. Beyond accuracy, the results highlight the model's exceptional zero-shot resilience to sensor failures.

2 Preliminaries

2.1 Problem Formulation

We represent the topology of the road network as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \boldsymbol{A})$, where $\mathcal{V}$ is a set of N nodes (sensors or intersections), and $\boldsymbol{A} \in R^{N\times N}$ denotes the weighted adjacency matrix encoding spatial connectivity. The dynamic state of the system is captured by the traffic flow tensor $\mathcal{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T) \in R^{T\times N\times D}$, where $\boldsymbol{X}_t$ represents the D-dimensional traffic features at time step t.

Traffic flow prediction seeks to forecast future traffic states based on historical data. Specifically, the objective is to model the conditional distribution of future time steps $(\boldsymbol{X}_{t_0}, \cdots, \boldsymbol{X}_T)$ given a historical context window and a graph. Formally, the problem can be formulated as:

$$q\left(\boldsymbol{X}_{t_0:T} \mid \boldsymbol{X}_{1:t_0-1}, \mathcal{G}\right) = \prod_{t=t_0}^{T} q\left(\boldsymbol{X}_t \mid \boldsymbol{X}_{1:t-1}, \mathcal{G}\right), \quad (1)$$

Unlike deterministic approaches that output a single point estimate, this probabilistic formulation allows us to capture the stochastic uncertainty inherent in complex traffic dynamics.

2.2 Background: Conditional Diffusion

To model the complex distribution of traffic flows, we adopt the Denoising Diffusion Probabilistic Model (DDPM) [Ho *et al.*, 2020] as our generative backbone. While standard DDPMs generate samples from pure Gaussian noise, traffic forecasting requires conditioning on historical states. Therefore, we align our approach with the autoregressive formulation introduced in TimeGrad [Rasul *et al.*, 2021].

In this framework, the temporal dependencies are managed by a recurrent component (RNN) that summarizes the history $\boldsymbol{X}_{1:t-1}$ into a hidden state $\mathbf{h}_{t-1}$. The diffusion process is then applied to the current observation $\boldsymbol{X}_t$ conditioned on this state. The training involves a forward diffusion process that gradually corrupts $\boldsymbol{X}_t$ into Gaussian noise $\boldsymbol{x}_t^n$ over N steps, and a parameterized reverse process that learns to denoise it. Instead of predicting the clean data directly, we train a denoising network ϵ_θ to estimate the noise component added at each step. The training objective minimizes the re-weighted variational bound:

$$\mathcal{L} = E_{t,n,\epsilon,\boldsymbol{X}_t}\left[\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_n}\boldsymbol{X}_t + \sqrt{1-\bar{\alpha}_n}\epsilon, \mathbf{h}_{t-1}, n\right)\right\|^2\right], \quad (2)$$

where n is the diffusion step, $\bar{\alpha}_n$ follows a fixed variance schedule, and $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. (See Appendix A for the full derivation of the variational lower bound and transition kernels). This design allows the model to generate diverse and realistic future flow by iteratively refining random noise under the guidance of historical traffic patterns.

3 Methodology

3.1 Framework Overview

As shown in Fig. 2, MS-STDT framework operates through a progressive, coarse-to-fine strategy comprising five core components:

1. **Multi-Scale Sampling (MSS):** Instead of processing raw data $\mathcal{X}$ solely at its finest resolution, we project it into a latent space $\boldsymbol{X}_{\text{data}}$ and progressively downsample it via strided convolutions. This yields a hierarchy of representations $\mathcal{X}_{\text{ms}} = \{\boldsymbol{X}^0_{\text{data}}, \ldots, \boldsymbol{X}^M_{\text{data}}\}$, where the g-th scale $\boldsymbol{X}^g_{\text{data}}$ captures dynamics at a temporal resolution of $T/2^g$. The sampling process is recursively defined as $\boldsymbol{X}^g_{\text{data}} = \text{Conv}\left(\boldsymbol{X}^{g-1}_{\text{data}}, \text{stride} = 2\right)$ for $g \in \{1, \ldots, M\}$, allowing the model to explicitly decouple macro-trends from micro-variations.
2. **Contextual Embedding:** To ground these representations in physical space and time, we augment the data features with a composite embedding. We sum learnable Laplacian eigenvectors $\boldsymbol{X}^g_{\text{spe}}$ (encoding graph topology [Belkin and Niyogi, 2003]), periodic embeddings $\boldsymbol{X}^g_w, \boldsymbol{X}^g_d$ (capturing weekly and daily cycles [Ding *et al.*, 2023]), and temporal positional encodings $\boldsymbol{X}^g_{\text{tpe}}$ [Vaswani *et al.*, 2017]. The final input at each scale is defined as: $\boldsymbol{X}^g = \boldsymbol{X}^g_{\text{data}} + \boldsymbol{X}^g_{\text{spe}} + \boldsymbol{X}^g_w + \boldsymbol{X}^g_d + \boldsymbol{X}^g_{\text{tpe}}$.
3. **Spatial-Temporal Process Module (STPM):** Serving as the conditional encoder, this module operates independently at each scale g. It utilizes a unified masked attention mechanism to extract robust spatial-temporal features $\boldsymbol{C}^g$, ensuring the model remains resilient to sensor failures and missing data.
4. **Guided Diffusion Process Module (GDPM):** This serves as the generative engine of our framework. Unlike standard diffusion models that treat denoising as a flat, isotropic task, GDPM utilizes the multi-scale representations as structural anchors. This allows the reverse process to stabilize the generation of fine-grained details by conditioning on the coarser, more stable trends.
5. **Multi-Scale Fusion and Output (MSFO):** Finally, we synthesize the predictions across all resolutions. An adaptive attention mechanism dynamically weights the generated samples $\boldsymbol{S}^g$ from each scale, fusing them into a single, cohesive forecast: $\text{Output} = \text{Ensemble}\left(\{\text{Attn}_g(\boldsymbol{S}^g)\}_{g=0}^M\right)$.

3.2 Spatial-Temporal Process Module (STPM)

Functioning as the scale-specific encoder, the STPM operates independently at each resolution g to extract robust features for the diffusion process. To distinguish intrinsic sparsity (e.g., low volume) from extrinsic missingness (e.g., sensor failure), it employs a **unified masked modeling** strategy. By forcing the reconstruction of randomly masked tokens via spatial-temporal context, the module ensures that structural dependencies are preserved across the hierarchy—guaranteeing that both stable macro-trends and stochastic micro-variations are accurately encoded even in the presence of sensor failures.

Unified Masking Strategies. We employ a diverse set of masks, each designed to teach the model a specific dependency: *Local Robustness:* **Random Masking** (M_{rand}) drops random patches to encourage local interpolation [Feichtenhofer *et al.*, 2022]. *Sensor Reliability:* **Tube Masking** (M_{tube}) hides specific nodes across all time steps, simulating sensor outages [Yuan *et al.*, 2024]. *Global Context:* **Block Masking** (M_{block}) removes large contiguous regions, forcing the model to infer states from distant global signals [Li *et al.*, 2022a]. *Topology Awareness:* **Short/Long Masking** (M_{short}, M_{long}) selectively hides neighbors based on graph distance, refining spatial propagation [Jiang *et al.*, 2023]. *Forecasting Capability:* **Future Masking** (M_{future}) strictly hides future steps, aligning the training objective with the autoregressive nature of forecasting [Tan *et al.*, 2023].

Masked Attention Mechanism. To process these masked inputs, we utilize a factorized attention mechanism. For a given scale embedding $\boldsymbol{X}^g$, we compute attention separately along the spatial and temporal axes. Let $\boldsymbol{X}^g_{t::}$ and $\boldsymbol{X}^g_{:n:}$ denote temporal and spatial slices, respectively.

Spatial and Temporal Mask Attention. We adopt masked attention mechanisms to encode structured priors in both spatial and temporal dimensions. For each attention head g, we first construct the query, key, and value matrices as:

$$\begin{cases} \boldsymbol{Q}^{(S,g)}_t = \boldsymbol{X}^g_{t::}\boldsymbol{W}^{(S,g)}_Q, \boldsymbol{Q}^{(T,g)}_n = \boldsymbol{X}^g_{:n:}\boldsymbol{W}^{(T,g)}_Q, \\ \boldsymbol{K}^{(S,g)}_t = \boldsymbol{X}^g_{t::}\boldsymbol{W}^{(S,g)}_K, \boldsymbol{K}^{(T,g)}_n = \boldsymbol{X}^g_{:n:}\boldsymbol{W}^{(T,g)}_K, \\ \boldsymbol{V}^{(S,g)}_t = \boldsymbol{X}^g_{t::}\boldsymbol{W}^{(S,g)}_V, \boldsymbol{V}^{(T,g)}_n = \boldsymbol{X}^g_{:n:}\boldsymbol{W}^{(T,g)}_V. \end{cases} \tag{3}$$

Here, the subscripts t and n denote temporal step and spatial node, respectively. Then masked attention scores and outputs are computed as:

$$\begin{cases} \boldsymbol{A}^{(S,g)}_t = \dfrac{\boldsymbol{Q}^{(S,g)}_t(\boldsymbol{K}^{(S,g)}_t)^\top}{\sqrt{d'}}, \boldsymbol{A}^{(T,g)}_n = \dfrac{\boldsymbol{Q}^{(T,g)}_n(\boldsymbol{K}^{(T,g)}_n)^\top}{\sqrt{d'}}, \\ \boldsymbol{O}^{(S,g)}_t = \text{Softmax}(\boldsymbol{A}^{(S,g)}_t \odot \boldsymbol{M}_S)\boldsymbol{V}^{(S,g)}_t, \\ \boldsymbol{O}^{(T,g)}_n = \text{Softmax}(\boldsymbol{A}^{(T,g)}_n \odot \boldsymbol{M}_T)\boldsymbol{V}^{(T,g)}_n, \end{cases} \tag{4}$$

where $\boldsymbol{M}_S \in \{M_{\text{rand}}, M_{\text{tube}}, M_{\text{block}}, M_{\text{short}}, M_{\text{long}}\}$, $\boldsymbol{M}_T \in \{M_{\text{rand}}, M_{\text{tube}}, M_{\text{block}}, M_{\text{future}}\}$.

Multi-Head Fusion. To fully integrate diverse spatial and temporal dependencies while maintaining computational efficiency, we apply multi-head attention across various masking strategies. Each head focuses on a specific aspect of the data, and their outputs are concatenated and linearly projected to

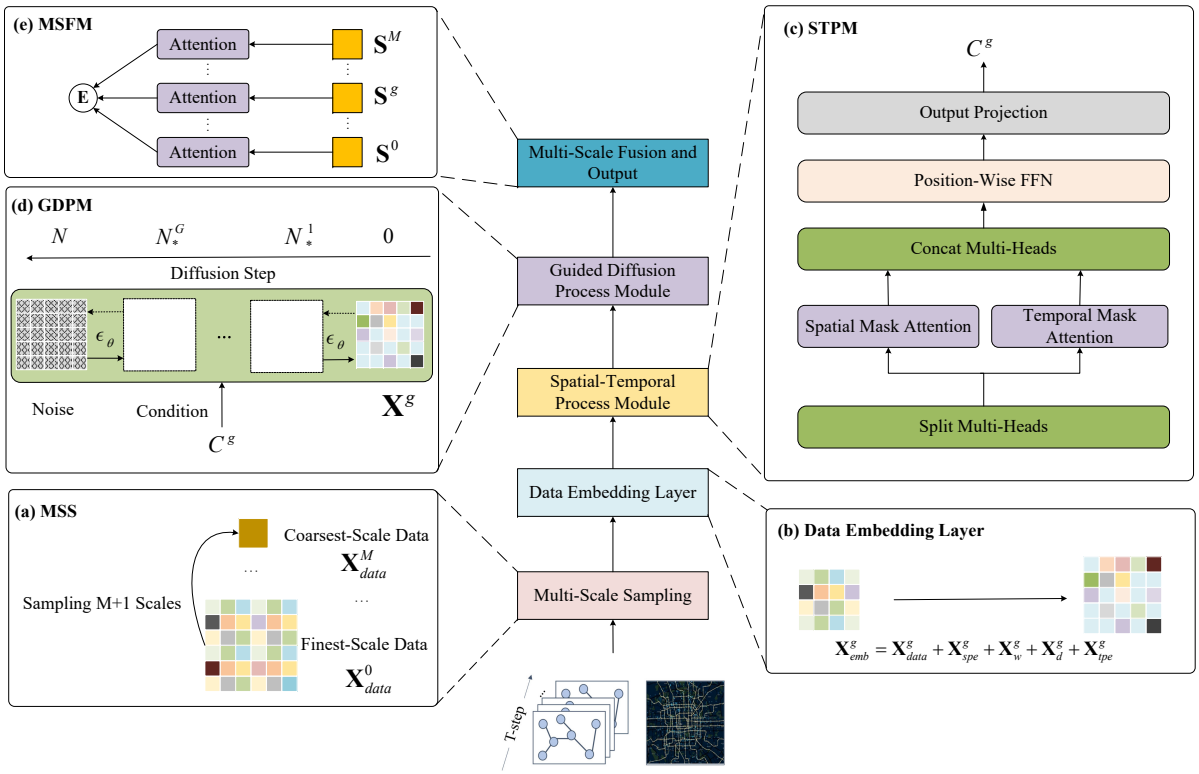

Figure 2: Overall architecture of MS-STDT.

form the final representation.

$$\begin{aligned} \boldsymbol{O}^g = \mathrm{Concat}(&\boldsymbol{O}^{(S,g,\mathrm{rand})}_{1:h_{(S,\mathrm{rand})}}, \boldsymbol{O}^{(S,g,\mathrm{tube})}_{1:h_{(S,\mathrm{tube})}}, \boldsymbol{O}^{(S,g,\mathrm{block})}_{1:h_{(S,\mathrm{block})}}, \\ &\boldsymbol{O}^{(g,\mathrm{short})}_{1:h_{(S,\mathrm{short})}}, \boldsymbol{O}^{(g,\mathrm{long})}_{1:h_{(S,\mathrm{long})}}, \boldsymbol{O}^{(T,g,\mathrm{rand})}_{1:h_{(T,\mathrm{rand})}}, \\ &\boldsymbol{O}^{(T,g,\mathrm{tube})}_{1:h_{(T,\mathrm{tube})}}, \boldsymbol{O}^{(T,g,\mathrm{block})}_{1:h_{(T,\mathrm{block})}}, \boldsymbol{O}^{(g,\mathrm{future})}_{1:h_{(T,\mathrm{future})}})\boldsymbol{W}^l, \end{aligned} \tag{5}$$

where $\mathrm{Concat}(\cdot)$ denotes concatenation, and $\boldsymbol{W}^l$ is a learnable projection matrix. Following the standard Transformer architecture, the fused output is passed through a position-wise feed-forward network, with residual connections and layer normalization applied to stabilize training and enhance representational capacity. The resulting feature representation is denoted as $\mathcal{X}_o \in R^{T\times N\times d}$. Subsequently, a 1×1 convolutional projection is employed to map $\mathcal{X}_o$ into a conditional embedding $\boldsymbol{C}^g \in R^{T\times N\times d_c}$, which serves as the scale-specific spatial-temporal representation.

3.3 Multi-Scale Guided Diffusion

Standard diffusion models typically neglect the intrinsic multi-scale nature of traffic data, treating the generative process as a monolithic mapping from noise to data. However, traffic flow exhibits a distinct hierarchical structure where stable global trends at coarse scales provide the foundation for local fluctuations at fine scales. To align the generative mechanism with this physical reality, we propose the **Guided Diffusion Process Module (GDPM)**.

The core design principle of our approach is that the forward diffusion process, characterized by the progressive injection of Gaussian noise, is functionally analogous to a multi-scale coarse-graining operation. As the diffusion steps advance and noise intensity increases (from step 0 to N), fine-grained high-frequency details dissipate first, leaving behind only the robust low-frequency structural information. We leverage this inherent property to formalize a direct mapping between the intermediate latent variables of the diffusion chain and the coarser resolutions of our data hierarchy. By explicitly anchoring specific diffusion time-steps to their corresponding multi-scale representations, we guide the model to reconstruct the stable global skeleton before refining the local details.

Coarse-Scale Guidance via Anchor Points. Let $q(\mathbf{x}^0)$ denote the distribution of the finest-scale data ($g=0$). A standard forward process $q(\mathbf{x}^0_{1:N}|\mathbf{x}^0_0)$ corrupts this data into noise over N steps. To inject structural guidance, we identify specific intermediate steps $N^g_* \in [1, N-1]$ as *anchor points* for each coarser scale g.

We enforce that the latent state $\mathbf{x}_{N^g_*}$ at this anchor step must be consistent with the observed coarse-scale data $\mathbf{x}^g$. Formally, we optimize the model to maximize the likelihood of the coarse observation at this specific noise level. The marginal distribution of the latent variable $\mathbf{x}_{N^g_*}$ is defined as:

$$p_\theta\left(\mathbf{x}_{N^g_*}\right) = \int p(\mathbf{x}_N) \prod_{n=N^g_*+1}^{N} p_\theta\left(\mathbf{x}_{n-1} \mid \mathbf{x}_n\right) \mathrm{d}\mathbf{x}_{(N^g_*+1):N}, \tag{6}$$

where $\mathbf{x}_N \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and each reverse transition is modeled as a Gaussian: $p_\theta\left(\mathbf{x}_{n-1} \mid \mathbf{x}_n\right) = \mathcal{N}\left(\mathbf{x}_{n-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_n, n), \boldsymbol{\sigma}_\theta(\mathbf{x}_n, n)\right)$. To enable tractable optimization, a standard approach is to maximize the variational lower bound of the log-likelihood in Eq. 6. This can be

achieved by introducing a sequence of latent variables of length $N - N_*^g$, which connects the observed coarse-scale data $\mathbf{x}^g$ to the generative process. Specifically, we define a forward diffusion process over $\mathbf{x}^g$ consisting of $N-N_*^g$ steps, producing a sequence of noisy variables $\mathbf{x}^g_{N_*^g+1}, \ldots, \mathbf{x}^g_N$. Under this construction, the log-likelihood of the observed coarse sample can be expressed as:

$$\log p_\theta\left(\mathbf{x}^g\right) = \log \int p_\theta\left(\mathbf{x}^g_{N_*^g}, \ldots, \mathbf{x}^g_N\right) \mathrm{d}\mathbf{x}^g_{(N_*^g+1):N}. \quad (7)$$

Following the DDPM framework [Ho *et al.*, 2020], this objective simplifies to minimizing the denoising error at the anchor steps:

$$\mathcal{L}^g_{\text{guidance}} = E_{\boldsymbol{\epsilon},\mathbf{x}^g,n}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\mathbf{x}^g_n, n\right)\right\|^2\right], \quad (8)$$

where $\mathbf{x}^g_n = \left(\prod_{i=N_*^g}^n \alpha_i^1\right)\mathbf{x}^g + \sqrt{1-\prod_{i=N_*^g}^n \alpha_i^1}\,\boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ (see Appendix C for full derivation).

Defining the Sharing Ratio and Anchor Steps. Determining the optimal anchor step N_*^g is critical to ensure coarse guidance aligns with natural diffusion dynamics. We quantify this via the **Sharing Ratio** $r^g := 1-(N_*^g-1)/N$. This metric formalizes the spectral correspondence between the diffusion process (acting as a progressive low-pass filter) and the data hierarchy. By serving as a normalized coordinate, r^g maps discrete physical scales onto the generative timeline, identifying the precise window where coarse guidance is structurally homologous to the latent state.

To avoid arbitrary hyperparameter tuning, we propose a data-driven heuristic to determine N_*^g. We posit that the optimal anchor step is where the signal complexity of the noisy fine-scale data matches that of the coarse data. Mathematically, we select N_*^g by minimizing the Kullback-Leibler (KL) divergence: $N_*^g := \arg\min_n D_{KL}\left(q(\mathbf{x}^g) \,||\, p_\theta(\mathbf{x}_n^0)\right)$.

In practice, we estimate this by pre-training a baseline model and selecting the step n where the Continuous Ranked Probability Score (CRPS) between the generated sample and the ground-truth coarse data is minimized. This ensures that guidance is applied exactly when the diffusion process is naturally resolving features at that specific resolution.

Multi-Scale Objective. With the anchor points defined, we construct a scale-specific variance schedule α_n^g. For steps $n \le N_*^g$, we set $\alpha_n^g = 1$ (no noise added), effectively treating the coarse data as "clean" relative to that noise level. For $n > N_*^g$, it follows the standard schedule:

$$\alpha_n^g(N_*^g) = \begin{cases} 1 & \text{for } n = 1, \ldots, N_*^g, \\ \alpha_n^1 & \text{for } n = N_*^g + 1, \ldots, N. \end{cases} \quad (9)$$

We assume a strictly increasing hierarchy of anchor steps: $N_*^0 < N_*^1 < \cdots < N_*^M$, ensuring that coarser scales enter the diffusion process later. The final training objective aggregates the losses across all scales, weighted by ω^g:

$$\mathcal{L}_{\text{total}} = \sum_{g=0}^{M} \omega^g E_{\boldsymbol{\epsilon},\mathbf{x}_0^g,n}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\mathbf{x}^g_n, n, \boldsymbol{C}^g\right)\right\|^2\right], \quad (10)$$

where $\boldsymbol{C}^g$ is the robust context representation from the STPM. This formulation ensures that the model simultaneously learns to generate accurate global trends (via high N_*^g layers) and precise local details (via the $g = 0$ layer).

Datasets	Nodes	Edges	Timesteps	Interval	Range
PeMS04	307	340	16992	5min	01/01/18–02/28/18
PeMS07	883	866	28224	5min	05/01/17–08/31/17
PeMS08	170	295	17856	5min	07/01/16–08/31/16
NYCTaxi	75 (15×5)	484	17520	30min	01/01/14–12/31/14
CHIBike	270 (15×18)	1966	4416	30min	07/01/20–09/30/20
T-Drive	1024 (32×32)	7812	3600	60min	02/01/15–06/30/15

Table 1: Data Description.

Inference. During inference, we generate predictions recursively. Starting from the end of the historical context $t_0 - 1$, we sample predictions $\boldsymbol{X}^g_{t_0}$ for each scale g. This process is repeated for the entire horizon $[t_0, T]$, producing a complete set of multi-scale predictions $\boldsymbol{S}^g$.

4 Experiments

4.1 Experimental Settings

Datasets. MS-STDT is evaluated on three graph-based benchmarks (PeMS04/07/08 [Song *et al.*, 2020]) and three grid-based benchmarks (NYCTaxi [Liu *et al.*, 2020], CHIBike [Wang *et al.*, 2021], and T-Drive [Pan *et al.*, 2019]); dataset details are summarized in Tab. 1. The forecasting horizon is fixed to 24 steps for all datasets.

Evaluation Metrics. We adopt five widely used metrics for evaluation: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) for deterministic accuracy, as well as Continuous Ranked Probability Score (CRPS) for probabilistic prediction.

Baselines. We evaluate the predictive performance of MS-STDT against three categories of strong baselines: (1) Diffusion-based models, including *TimeGrad* [Rasul *et al.*, 2021], *D3VAE* [Li *et al.*, 2022b], *MG-TSD* [Fan *et al.*, 2024], *DiffSTG* [Wen *et al.*, 2023], *USTD* [Hu *et al.*, 2024b], UrbanDiT [Yuan *et al.*, 2025] (2) Pretrained and self-supervised models, such as *ST-SSL* [Ji *et al.*, 2023], *UniST* [Yuan *et al.*, 2024], *STD-MAE* [Gao *et al.*, 2024]; (3) Transformer- and MLP-based models, including *BasisFormer* [Ni *et al.*, 2023], *TimeMixer* [Wang *et al.*, 2024], STDN [Cao *et al.*, 2025].

Parameters Settings. MS-STDT is trained for 100 epochs using the Adam optimizer with a learning rate of 10^{-5}. The batch size is conFig.d to 64 for the CHIBike dataset and 32 for all other datasets. All experiments are conducted on a single NVIDIA A100 GPU (40 GB) with 128 GB of memory using PyTorch 1.10.1 on Ubuntu 18.04.

4.2 Main Results

Table 2 presents the quantitative comparison of MS-STDT against 12 competitive baselines across six diverse benchmarks. The results demonstrate that our framework establishes a new state-of-the-art, consistently achieving the lowest error rates across all datasets and metrics.

Deterministic Forecasting Performance. In terms of point accuracy (MAE and RMSE), MS-STDT exhibits a significant performance margin over both graph-based and grid-based baselines.

Graph-Structured Data (PeMS04/07/08): Our model effectively captures the complex topology of traffic networks.

Method	PeMS04			PeMS07			PeMS08		
	MAE↓	RMSE↓	CRPS↓	MAE↓	RMSE↓	CRPS↓	MAE↓	RMSE↓	CRPS↓
TimeGrad	22.841±0.229	35.709±0.396	0.78±0.039	21.380±0.252	36.247±0.383	1.08±0.066	15.033±0.159	29.179±0.288	0.40±0.013
D3VAE	21.103±0.158	33.827±0.291	0.88±0.199	20.439±0.141	35.195±0.289	1.43±0.307	14.915±0.169	27.329±0.239	0.59±0.189
MG-TSD	20.347±0.136	31.012±0.259	0.67±0.198	19.853±0.131	33.061±0.247	0.95±0.144	13.812±0.126	25.323±0.204	0.54±0.172
DiffSTG	20.459±0.121	32.615±0.226	0.42±0.138	20.852±0.133	32.092±0.242	1.10±0.142	14.479±0.149	26.180±0.197	0.50±0.114
USTD	18.746±0.140	30.659±0.216	0.38±0.174	19.662±0.129	31.967±0.231	0.92±0.113	13.293±0.154	21.445±0.263	0.46±0.106
UrbanDiT	*17.974±0.138	*29.057±0.214	*0.36±0.104	*18.712±0.139	*31.072±0.307	*0.91±0.122	*11.962±0.144	*20.844±0.197	*0.44±0.127
ST-SSL	21.561±0.018	33.715±0.021	-	20.875±0.011	33.971±0.023	-	14.327±0.019	26.923±0.021	-
UniST	18.362±0.016	29.253±0.021	-	18.753±0.012	31.440±0.025	-	13.446±0.018	22.912±0.023	-
STD-MAE	20.639±0.019	31.314±0.031	-	20.318±0.016	34.011±0.027	-	14.109±0.018	24.715±0.024	-
BasisFormer	21.423±0.012	31.314±0.021	-	19.293±0.016	32.098±0.020	-	13.008±0.018	22.473±0.024	-
TimeMixer	18.839±0.015	30.611±0.024	-	19.838±0.017	33.329±0.022	-	12.053±0.014	21.159±0.026	-
STDN	18.092±0.002	29.827±0.014	-	18.824±0.007	31.334±0.020	-	12.003±0.004	21.017±0.017	-
MS-STDT(Ours)	**16.726**±0.129	**27.063**±0.173	**0.30**±0.017	**17.362**±0.131	**29.537**±0.204	**0.87**±0.064	**10.391**±0.127	**18.096**±0.152	**0.32**±0.006

Method	NYCTaxi			T-Drive			CHIBike		
	MAE↓	RMSE↓	CRPS↓	MAE↓	RMSE↓	CRPS↓	MAE↓	RMSE↓	CRPS↓
TimeGrad	14.990±0.168	24.484±0.229	0.38±0.097	19.826±0.170	35.471±0.397	0.43±0.016	4.074±0.132	5.876±0.221	0.58±0.118
D3VAE	13.745±0.154	22.179±0.277	0.86±0.160	18.791±0.149	33.876±0.285	0.84±0.375	4.061±0.130	5.818±0.159	0.53±0.040
MG-TSD	12.446±0.138	21.365±0.216	0.37±0.178	16.562±0.123	32.092±0.242	0.44±0.116	3.893±0.109	5.456±0.126	0.40±0.163
DiffSTG	13.464±0.148	22.145±0.201	0.42±0.094	17.647±0.181	32.417±0.255	0.37±0.086	4.013±0.100	5.362±0.108	0.39±0.121
USTD	12.132±0.150	21.103±0.239	0.36±0.067	16.214±0.104	31.766±0.211	0.35±0.126	3.899±0.117	5.473±0.156	*0.34±0.097
UrbanDiT	*10.169±0.140	18.019±0.223	*0.33±0.174	*14.623±0.129	*29.812±0.219	*0.34±0.161	*3.698±0.101	*5.357±0.124	0.36±0.122
ST-SSL	13.054±0.014	22.821±0.027	-	19.191±0.019	34.772±0.025	-	4.054±0.005	5.820±0.007	-
UniST	10.291±0.022	17.937±0.032	-	15.173±0.017	30.237±0.024	-	3.879±0.006	5.426±0.009	-
STD-MAE	12.411±0.010	20.813±0.019	-	17.184±0.015	32.143±0.026	-	3.903±0.009	5.586±0.018	-
BasisFormer	11.426±0.010	18.536±0.029	-	16.293±0.013	31.098±0.027	-	3.993±0.009	5.586±0.026	-
TimeMixer	11.327±0.013	18.693±0.027	-	16.119±0.019	30.972±0.025	-	3.761±0.012	5.431±0.023	-
STDN	10.216±0.003	*17.921±0.011	-	15.010±0.002	29.994±0.021	-	3.816±0.001	5.446±0.003	-
MS-STDT(Ours)	**9.132**±0.122	**16.627**±0.181	**0.26**±0.088	**12.681**±0.134	**27.053**±0.196	**0.30**±0.038	**3.347**±0.080	**5.041**±0.102	**0.33**±0.075

Table 2: Performance comparison of MAE, RMSE, and CRPS on six datasets. Best results are bolded, second-best are marked with *.

On the PeMS08 dataset, for instance, MS-STDT achieves an MAE of **10.391**, representing a reduction of **13.1%** compared to the strongest diffusion baseline, UrbanDiT (11.962), and **13.4%** compared to the advanced transformer STDN (12.003). This indicates that our multi-scale guidance strategy allows the model to retain fine-grained spatial dependencies that standard graph diffusion often smooths over.

Grid-Based Data (NYC, T-Drive, CHIBike): The superiority of MS-STDT extends to grid-based inputs, demonstrating its versatility. On the highly dynamic T-Drive dataset, we surpass the second-best method (UrbanDiT) by reducing the MAE from 14.623 to **12.681** (↓**13.3%**). Notably, while non-generative baselines like UniST and TimeMixer perform competitively on simpler metrics, they struggle to match the peak performance of MS-STDT on high-variance datasets (e.g., NYCTaxi), confirming the necessity of our guided generative approach.

Probabilistic Forecasting Quality. Beyond simple accuracy, the CRPS metric reveals the model's ability to model the full distribution of traffic states. As shown in Table 2, MS-STDT achieves the lowest CRPS across all scenarios. On PeMS04, our model reduces CRPS to **0.30**, significantly outperforming the standard diffusion baseline DiffSTG (0.42). This suggests that the *Guided Diffusion Process Module* (GDPM) does not merely memorize the mean trajectory but accurately calibrates the predictive uncertainty, producing tighter and more reliable confidence intervals—a critical attribute for safety-critical ITS applications.

Comparative Analysis. Comparing across categories, we observe that while self-supervised methods (e.g., UniST, STD-MAE) offer strong stability, they often lack the granular resolution provided by generative models. Conversely, standard diffusion models (e.g., TimeGrad, MG-TSD) often suffer from stochastic instability, leading to higher RMSE. MS-STDT effectively bridges this gap: by anchoring the diffusion process with coarse-scale guidance, it combines the stability of deterministic encoders with the generative expressiveness of diffusion.

4.3 Ablation Study

To validate the contribution of each component, we conducted ablation experiments across all six datasets (Table 3). The removal of the Guided Diffusion Process Module (**w/o GDPM**) resulted in the most significant performance degradation, confirming its pivotal role in stabilizing generation via multi-scale supervision. Similarly, excluding the Multi-Scale Fusion (**w/o MSFO**) or Multi-Scale Sampling (**w/o MSS**) consistently impaired accuracy, verifying that both hierarchical feature extraction and adaptive cross-scale aggregation are essential for capturing complex dependencies. Furthermore, the performance drop in w/o SMA highlights the necessity of explicit spatial modeling. Finally, replacing our decoupled attention with a unified variant (**w/ STMA**) yielded inferior results, supporting the design choice of separating spatial and temporal processing for more precise feature learning.

4.4 Effect of Parameters

We further investigate the impact of two critical parameters on model performance: the number of diffusion steps and the masking ratio used in training. Specifically, we varied the diffusion step from 20 to 100 and the masking ratio from 0.05 to 0.25, evaluating their effects on MAE across benchmark datasets. The results reveal that model accuracy generally

Dataset	Metric	w/o GDPM	w/o MSFO	w/o MSS	w/o SMA	w/ STMA	MS-STDT
PeMS04	MAE↓	18.767	17.847	17.780	17.612	18.065	**16.726**
	RMSE↓	30.202	29.066	29.147	28.669	29.224	**27.063**
PeMS07	MAE↓	19.254	18.594	18.557	18.253	18.716	**17.362**
	RMSE↓	32.816	31.723	31.669	31.312	32.065	**29.537**
PeMS08	MAE↓	11.492	11.035	11.068	10.931	11.167	**10.391**
	RMSE↓	20.148	19.400	19.472	19.235	19.659	**18.096**
NYCTaxi	MAE↓	10.109	9.808	9.685	9.597	9.824	**9.132**
	RMSE↓	18.607	17.941	17.792	17.651	18.031	**16.627**
T-Drive	MAE↓	13.936	13.683	13.559	13.357	13.597	**12.681**
	RMSE↓	29.948	29.271	28.994	28.878	29.428	**27.053**
CHIBike	MAE↓	3.712	3.551	3.544	3.515	3.653	**3.347**
	RMSE↓	5.631	5.374	5.398	5.329	5.470	**5.041**

Table 3: Results of ablation study. ↓ indicates lower is better.

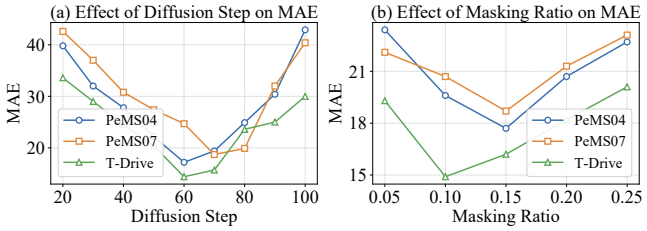

Figure 3: Impact of diffusion step and masking ratio on MAE.

improves with an increased diffusion step up to an optimal point, beyond which performance degrades due to potential over-smoothing. Similarly, a moderate masking ratio yields the best results, suggesting a balance between model robustness and information retention. These findings underscore the importance of careful hyperparameter tuning for optimal spatial-temporal forecasting performance.

4.5 Robustness and Efficiency Analysis

Zero-Shot Resilience. We evaluated model resilience against sensor failures by testing on data with 0–30% random missing rates without retraining. As shown in Fig. 4(a), MS-STDT demonstrates superior stability compared to baselines. While the deterministic *STDN* degrades rapidly (MAE rising from ∼18 to >24) and generative models like *UniST* and *UrbanDiT* exhibit linear error increases, MS-STDT maintains a significantly flatter degradation curve. Notably, at 30% sparsity, our model matches the performance of baselines at only 0–10% missing rates, validating the *Spatial-Temporal Process Module*'s ability to impute missing global context.

Computational Efficiency. Fig. 4(b) details the computational overhead. Despite its multi-scale architecture, MS-STDT achieves a highly favorable efficiency trade-off. It requires only ∼83 ms/iter for training, significantly faster than *UrbanDiT* (∼730 ms) and comparable to lightweight baselines due to our parallelized encoder. Inference is similarly accelerated (∼908 ms/iter vs. ∼2046 ms for UrbanDiT) via our guided diffusion strategy. Furthermore, MS-STDT maintains a moderate memory footprint (∼7303 MB), 21% lower than competitive diffusion models, ensuring deployment feasibility.

5 Related Work

Diffusion Models for Traffic Flow Forecasting. Diffusion models have become an important paradigm for traffic forecasting due to their ability to model complex predictive distributions and uncertainty [Zhang *et al.*, 2021; Rühling Cachay *et al.*, 2023; Wen *et al.*, 2023; Hu *et al.*, 2024a]. Early work such as TimeGrad [Rasul *et al.*, 2021] introduced autoregressive denoising for temporal modeling, while DiffSTG [Wen *et al.*, 2023] further incorporated graph structure into diffusion-based forecasting. More recent studies have explored generation stability and scalability, e.g., MG-TSD [Fan *et al.*, 2024] via multi-granularity guidance and UrbanDiT [Yuan *et al.*, 2025] via Diffusion Transformers. However, most existing methods still treat diffusion as a largely monolithic denoising process. In contrast, MS-STDT explicitly aligns the reverse process with the intrinsic coarse-to-fine hierarchy of traffic data, using stable coarse-scale trends to guide fine-grained generation.

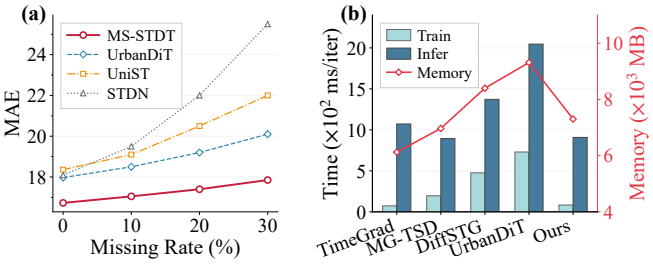

Figure 4: Robustness (a) and Efficiency (b) Analysis.

Transformers for Traffic Flow Forecasting. Transformers dominate long-range spatial-temporal modeling [Wu *et al.*, 2021; Jiang *et al.*, 2023; Jin *et al.*, 2023]. Recent advancements have increasingly focused on pattern disentanglement and generalization. For instance, STDN [Cao *et al.*, 2025] proposes a decomposition network to separate traffic data into distinct trend and seasonality components, and UniST [Yuan *et al.*, 2024] utilizes masked pre-training to learn robust representations for cross-city transfer. However, while architectures like STDN excel at isolating deterministic macro-trends, they often lack the probabilistic capacity to model the stochastic uncertainty inherent in dynamic traffic flows. MS-STDT addresses this limitation by synthesizing the structural decomposition benefits of these Transformers with the generative expressiveness of diffusion models, unifying them within a hierarchical multi-scale framework that is both physically consistent and computationally efficient.

6 Conclusion

In this paper, we proposed the **Coarse-to-Fine Latent Guidance** strategy via **MS-STDT** to address the challenge of multi-scale traffic heterogeneity. By reformulating diffusion as a hierarchical generative process, our method utilizes intrinsic data structures as structural anchors to stabilize prediction without external labels. Extensive experiments confirm that MS-STDT consistently outperforms existing baselines in accuracy while demonstrating superior zero-shot robustness and computational efficiency. In future work, we plan to explore its applicability to broader spatial-temporal domains [Lian *et al.*, 2026].

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (Nos.T2293771 and 42361144718). The funders had no role in the study design, data collection, analysis, decision to publish, or preparation of the manuscript.

References

[Belkin and Niyogi, 2003] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[Cai *et al.*, 2024] Wanlin Cai, Yuxuan Liang, Xianggen Liu, Jianshuai Feng, and Yuankai Wu. Msgnet: Learning multi-scale inter-series correlations for multivariate time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11141–11149, 2024.

[Cao *et al.*, 2025] Lingxiao Cao, Bin Wang, Guiyuan Jiang, Yanwei Yu, and Junyu Dong. Spatiotemporal-aware trend-seasonality decomposition network for traffic flow forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 11463–11471, 2025.

[Carianni and Gemma, 2025] Annarita Carianni and Andrea Gemma. Overview of traffic flow forecasting techniques. *IEEE Open Journal of Intelligent Transportation Systems*, 6:848–882, 2025.

[Choi *et al.*, 2022] Jeongwhan Choi, Hwangyong Choi, Jeehyun Hwang, and Noseong Park. Graph neural controlled differential equations for traffic forecasting. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 36, pages 6367–6374, 2022.

[Ding *et al.*, 2023] Jiaqi Ding, Chao Yang, Yueyao Wang, Pengfei Li, Fulin Wang, Yuhao Kang, Haoyang Wang, Ze Liang, Jiawei Zhang, Peien Han, et al. Influential factors of intercity patient mobility and its network structure in china. *Cities*, 132:103975, 2023.

[Fan *et al.*, 2024] Xinyao Fan, Yueying Wu, Chang Xu, Yuhao Huang, Weiqing Liu, and Jiang Bian. MG-TSD: Multi-granularity time series diffusion models with guided learning process. In *The Twelfth International Conference on Learning Representations*, 2024.

[Feichtenhofer *et al.*, 2022] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Advances in Neural Information Processing Systems*, 35:35946–35958, 2022.

[Gao *et al.*, 2024] Haotian Gao, Renhe Jiang, Zheng Dong, Jinliang Deng, Yuxin Ma, and Xuan Song. Spatial-temporal-decoupled masked pre-training for spatiotemporal forecasting. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 3998–4006, 8 2024.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[Hu *et al.*, 2024a] 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*, pages 135–146, 2024.

[Hu *et al.*, 2024b] 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*, page 135–146, New York, NY, USA, 2024. Association for Computing Machinery.

[Ji *et al.*, 2023] Jiahao Ji, Jingyuan Wang, Chao Huang, Junjie Wu, Boren Xu, Zhenhe Wu, Junbo Zhang, and Yu Zheng. Spatio-temporal self-supervised learning for traffic flow prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 4356–4364, 2023.

[Jiang *et al.*, 2023] Jiawei Jiang, Chengkai Han, Wayne Xin Zhao, and Jingyuan Wang. Pdformer: Propagation delay-aware dynamic long-range transformer for traffic flow prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 4365–4373, 2023.

[Jin *et al.*, 2023] Di Jin, Jiayi Shi, Rui Wang, Yawen Li, Yuxiao Huang, and Yu-Bin Yang. Trafformer: Unify time and space in traffic prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8114–8122, 2023.

[Kong *et al.*, 2024] Weiyang Kong, Ziyu Guo, and Yubao Liu. Spatio-temporal pivotal graph neural networks for traffic flow forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8627–8635, 2024.

[Lana *et al.*, 2018] Ibai Lana, Javier Del Ser, Manuel Velez, and Eleni I Vlahogianni. Road traffic forecasting: Recent advances and new challenges. *IEEE Intelligent Transportation Systems Magazine*, 10(2):93–109, 2018.

[Li *et al.*, 2022a] Gang Li, Heliang Zheng, Daqing Liu, Chaoyue Wang, Bing Su, and Changwen Zheng. Semmae: Semantic-guided masking for learning masked autoencoders. *Advances in Neural Information Processing Systems*, 35:14290–14302, 2022.

[Li *et al.*, 2022b] Yan Li, Xinjiang Lu, Yaqing Wang, and Dejing Dou. Generative time series forecasting with diffusion, denoise, and disentanglement. *Advances in Neural Information Processing Systems*, 35:23009–23022, 2022.

[Lian *et al.*, 2026] Xinglin Lian, Chengtai Cao, Ting Zhong, Yong Wang, Kai Chen, and Fan Zhou. Decompose to understand, fuse to detect: Frequency-decoupled anomaly detection for encrypted network traffic. *arXiv preprint arXiv:2605.02970*, 2026.

[Liu *et al.*, 2020] Lingbo Liu, Jiajie Zhen, Guanbin Li, Geng Zhan, Zhaocheng He, Bowen Du, and Liang Lin. Dynamic spatial-temporal representation learning for traffic

flow prediction. *IEEE Transactions on Intelligent Transportation Systems*, 22(11):7169–7183, 2020.

[Liu *et al.*, 2023] 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.

[Ma *et al.*, 2020] Dongfang Ma, Xiang Song, and Pu Li. Daily traffic flow forecasting through a contextual convolutional recurrent neural network modeling inter-and intra-day traffic patterns. *IEEE Transactions on Intelligent Transportation Systems*, 22(5):2627–2636, 2020.

[Maimaris and Papageorgiou, 2016] Athanasios Maimaris and George Papageorgiou. A review of intelligent transportation systems from a communications technology perspective. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 54–59, 2016.

[Manibardo *et al.*, 2021] Eric L Manibardo, Ibai Laña, and Javier Del Ser. Deep learning for road traffic forecasting: Does it make a difference? *IEEE Transactions on Intelligent Transportation Systems*, 23(7):6164–6188, 2021.

[Ni *et al.*, 2023] Zelin Ni, Hang Yu, Shizhan Liu, Jianguo Li, and Weiyao Lin. Basisformer: Attention-based time series forecasting with learnable and interpretable basis. *Advances in Neural Information Processing Systems*, 36:71222–71241, 2023.

[Pan *et al.*, 2019] Zheyi Pan, Yuxuan Liang, Weifeng Wang, Yong Yu, Yu Zheng, and Junbo Zhang. Urban traffic prediction from spatio-temporal data using deep meta learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1720–1730, 2019.

[Rasul *et al.*, 2021] Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *International Conference on Machine Learning*, pages 8857–8868, 2021.

[Rühling Cachay *et al.*, 2023] Salva Rühling Cachay, Bo Zhao, Hailey Joren, and Rose Yu. Dyffusion: A dynamics-informed diffusion model for spatiotemporal forecasting. *Advances in Neural Information Processing Systems*, 36:45259–45287, 2023.

[Song *et al.*, 2020] Chao Song, Youfang Lin, Shengnan Guo, and Huaiyu Wan. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 914–921, 2020.

[Tan *et al.*, 2023] Qiaoyu Tan, Ninghao Liu, Xiao Huang, Soo-Hyun Choi, Li Li, Rui Chen, and Xia Hu. S2gae: Self-supervised graph autoencoders are generalizable learners with graph masking. In *Proceedings of the sixteenth ACM International Conference on Web Search and Data Mining*, pages 787–795, 2023.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[Wang *et al.*, 2021] Jingyuan Wang, Jiawei Jiang, Wenjun Jiang, Chao Li, and Wayne Xin Zhao. Libcity: An open library for traffic prediction. In *Proceedings of the 29th International Conference on Advances in Geographic Information Systems*, pages 145–148, 2021.

[Wang *et al.*, 2024] Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2024.

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

[Wu *et al.*, 2021] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: decomposition transformers with auto-correlation for long-term series forecasting. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2021.

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

[Yuan *et al.*, 2025] Yuan Yuan, Chonghua Han, Jingtao Ding, Guozhen Zhang, Depeng Jin, and Yong Li. Diffusion transformers as open-world spatiotemporal foundation models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[Zhang *et al.*, 2021] Xiyue Zhang, Chao Huang, Yong Xu, Lianghao Xia, Peng Dai, Liefeng Bo, Junbo Zhang, and Yu Zheng. Traffic flow forecasting with spatial-temporal graph diffusion network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 15008–15015, 2021.