Align Along Time and Space: A Graph Latent Diffusion Model for Traffic Dynamics Prediction

Yuhang Liu*, Yingxue Zhang*, Xin Zhang†, Yu Yang‡, Yiqun Xie§, Sahar Ghanipoor Machiani†, Yanhua Li¶, Jun Luo#
*Binghamton University, Binghamton, New York, USA; †San Diego State University, San Diego, California, USA; ‡Lehigh University, Bethlehem, Pennsylvania, USA; §University of Maryland, College Park, Maryland, USA; ¶Worcester Polytechnic Institute, Worcester, Massachusetts, USA; #Logistics and Supply Chain MultiTech R&D Centre, Hong Kong, China
{yliu41, yzhang42}@binghamton.edu, {xzhang19, sghanipoor}@sdsu.edu, yuyang@lehigh.edu, xie@umd.edu,
yli15@wpi.edu, jluo@lscm.hk

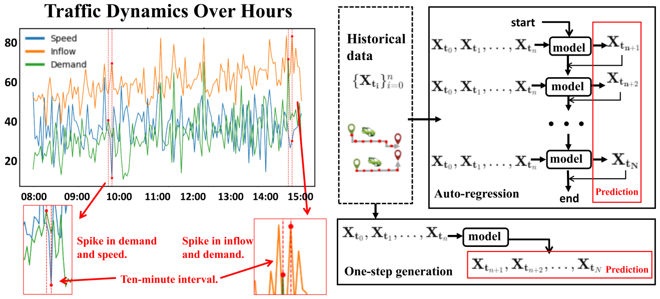
***Abstract*—The problem of traffic dynamics prediction, aiming to capture the complicated patterns of urban dynamics and forecast short-term future traffic status, is essential for managing transportation systems, reducing congestion, enhancing safety, improving commuter efficiency, and supporting urban planning and infrastructure development. Current approaches using machine learning and deep neural networks have advanced traffic prediction but often focus on individual urban dynamic aspects and rely on auto-regressive methods for consecutive predictions, which can be inaccurate and computationally expensive. In this work, we propose the Spatial-Temporal Graph LAtent DIffusion ModeL (STGAIL) to address these limitations. STGAIL views geographical regions as graphs with various traffic features, capturing their interconnections. Operating in a pre-trained latent space, STGAIL uses latent diffusion processes and innovative spatial-temporal graph layers for accurate and efficient multi-step predictions. Fine-tuning with temporal binary masks further enhances its performance, avoiding error accumulation and reducing computational costs. Experiments on real-world datasets demonstrate STGAIL's superior accuracy and efficiency over state-of-the-art methods. We also make our code and dataset available, contributing to ongoing research in traffic dynamics prediction.**

***Index Terms*—urban dynamics prediction, latent diffusion models, spatial-temporal data mining**

I. Introduction

Traffic dynamics refer to the complex patterns of human mobility on road networks, including variations in traffic flow, speed, local travel demand, *etc*. The process of traffic dynamics prediction aims to forecast how the traffic status changes and responds to various influences over time. Effective traffic prediction is essential for managing and optimizing transportation systems and has significant implications for numerous urban applications. Accurate prediction of traffic dynamics can reduce congestion, enhance road safety, and improve route efficiency for commuters. [17], [18] Moreover, these predictions are pivotal in urban planning, emergency response coordination, and infrastructure development. [39] By utilizing predictive insights, cities can develop more resilient and adaptive transportation networks, which are vital for sustainable urban growth and improving the quality of life for residents [15], [31].

State-of-the-art (SOTA) approaches. Various studies have attempted to address the traffic dynamics prediction problem, employing techniques from both traditional machine learning [4], [6], [19], [34] and deep neural networks [8], [20], [30], [42], [43], [58]. For example, the TD^2-DL model [40] predicts high-resolution traffic speed propagation by integrating temporal-spatial dependencies and traffic flow dynamics using deep learning techniques. Approaches including cST-ML [55] and DAC-ML [50] utilize meta-learning to adapt to dynamically changing urban traffic conditions. Furthermore, the methods such as TrafficGAN [51], Curb-GAN [52], C3-GAN [54], STrans-GAN [53], and Mest-GAN [56] have been introduced to estimate traffic under varying urban scenarios. Additionally, the LSGCN model [14] effectively captures complex spatial-temporal features, providing stable predictions for both short and long-term traffic forecasting. Other notable models like STGCN [41] and GMAN [57] have also been developed to enhance the understanding and prediction of traffic flows and related dynamics. However, many of these methods [53], [54] focus on a single traffic dynamics aspect, such as traffic speed, and neglect other important traffic patterns like travel demand and traffic volume. The complicated correlations among various traffic dynamic patterns are omitted. Additionally, most existing works [40], [50], [52],


Traffic Dynamics Over Hours
Speed
Inflow
Demand
80
60
40
20
08:00
09:00
10:00
11:00
12:00
13:00
14:00
15:00
Spike in demand and speed.
Spike in inflow and demand.
Ten-minute interval.
Historical data
$\{\mathbf{X}_{\mathrm{t_i}}\}_{i=0}^{n}$
start
$\mathbf{X}_{\mathrm{t}_0}, \mathbf{X}_{\mathrm{t}_1}, \ldots, \mathbf{X}_{\mathrm{t}_n}$
model
$\mathbf{X}_{\mathrm{t_{n+1}}}$
$\mathbf{X}_{\mathrm{t_{n+2}}}$
$\mathbf{X}_{\mathrm{t_N}}$
Auto-regression
end
Prediction
One-step generation
$\mathbf{X}_{\mathrm{t}_{n+1}}, \mathbf{X}_{\mathrm{t}_{n+2}}, \ldots, \mathbf{X}_{\mathrm{t}_N}$
Prediction

a) Traffic speed, inflow and demand are correlated.

b) Difference between auto-regression and one-step generation.

Fig. 1: Traffic dynamics correlations (a) and the difference of auto-regression methods and one-step generation methods (b). In the left zoomed-in figure in (a), when travel demand begins to decrease, traffic speed increases ten minutes later. In the right zoomed-in figure in (a), the taxi inflow decreases ten minutes after the travel demand decreases.

TABLE I: Notations.

Notations	Descriptions
$\mathcal{R} = \{R_{ij}\}$	Target region.
$\boldsymbol{X}^k_{R,t}$	Traffic feature of type k at time t, region R.
$\boldsymbol{A}^R$	Adjacency matrix at region R.
$V^R_t = \{v_{ij,t}\}$	Set of vertices in region R at time t.
$\boldsymbol{G}^R_t = (V^R_t, E^R, \boldsymbol{A}^R)$	State of the traffic network at region R, time t.

[55] rely on either auto-regression for long-term prediction or can only predict for a fixed period [57]. These limitations hinder the development of more flexible and efficient traffic prediction models. These limitations are elaborated below:

Limition 1. Complex multifaceted traffic dynamics. Most current approaches focus on predicting a single facet of traffic dynamics, such as traffic speed or volume [21], [22], [48], without considering the interconnectedness among other facets. However, traffic dynamics are inherently multifaceted and inter-correlated; overlooking these interconnections can lead to incomplete and sub-optimal performance. For example, a spike in traffic volume at a specific location often impacts traffic speed and congestion levels both upstream and downstream and might lead to a temporary traffic speed decrease at a later time demonstrated in Fig. 1a. When neglecting these interconnections, models may fail to capture the cascading effects throughout the traffic network, ultimately resulting in forecasts that lack holistic accuracy and applicability.

Limitation 2. Inaccurate and inflexible long-term prediction. Most approaches predict longer-term traffic patterns in an auto-regressive manner, which utilizes previously predicted data as inputs for future predictions. This approach encounters two primary issues: *(i) Error accumulation* is observed when these models make further predictions based on the assumption that data predicted in prior steps is sufficiently accurate. However, any initial inaccuracies tend to accumulate progressively with each prediction step, exacerbating errors as the process unfolds and potentially leading to significant deviations in long-term forecasts [1], [35]. (ii) *Excessive computation* is required when auto-regressive models predict traffic dynamics sequentially, as the model must be run repeatedly to generate predictions for all future time steps as is shown in Fig. 1b. This step-by-step nature not only incurs high computational costs but also introduces delays that are untenable in real-time traffic management scenarios. Moreover, the inherent rigidity of this approach limits its ability to adapt swiftly to sudden or short-term changes in traffic conditions, thus diminishing its effectiveness in dynamic urban settings.

Our approach. To address the aforementioned challenges and solve the traffic dynamics prediction problem, this paper introduces the Spatial-Temporal Graph LAtent DIffusion ModeL (STGAIL). This model views geographical regions as graphs, with various traffic dynamics as features. This allows STGAIL to effectively capture the intricate interconnections among diverse traffic dynamics, as well as the complex road networks and spatial-temporal correlations within these regions. Drawing inspiration from latent diffusion models, which are capable of generating multiple data samples simultaneously, STGAIL is designed to predict traffic dynamics both accurately and efficiently for multiple future time steps at once, thereby circumventing the limitations associated with auto-regressive prediction methods. Our key contributions are detailed below:

- We propose a novel spatial-temporal graph latent diffusion model – STGAIL, specifically designed for capturing the complicated traffic patterns. STGAIL operates in a pre-trained graph latent space, effectively reducing computational complexity by projecting traffic dynamics graphs into latent vectors using a pre-trained autoencoder and a discriminator. Moreover, STGAIL integrates a latent diffusion model equipped with innovative spatial and temporal graph layers, which are adept at learning interconnections among various traffic dynamics and capturing their complex spatial-temporal dependencies (Sec. III-A and III-B).
- We introduce a conditional generation approach, adapting our STGAIL model to predict traffic dynamics for a flexible period. During this phase, temporal binary masks and masked traffic dynamics graphs are used to condition the well-trained STGAIL model, enabling it to effectively produce traffic predictions for multiple time steps concurrently. This conditioning method prevents the accumulation of prediction errors and reduces computational costs, significantly enhancing both the accuracy and efficiency of the predictions (Sec. III-C).
- We conducted extensive experiments with real-world traffic dynamics datasets to evaluate the effectiveness of our proposed STGAIL model. The results demonstrate that STGAIL significantly improves traffic prediction performance in terms of accuracy and efficiency and surpasses state-of-the-art baseline methods. *As a contribution to the research community, we have made our code and unique dataset available through an anonymous GitHub link* [1] (Sec. IV).

II. Overview

In this section, we formally define the traffic dynamics prediction problem.To facilitate understanding and clarity, the notations used in this paper are provided in Tab. I.

A. Definitions & Notations

Definition 1 (Target region R). We partition a city into $P \times Q$ grid cells with equal side-length in latitude and longitude as shown in Fig. 2. We denote the set of grid cells as $\mathcal{S} = \{s_{ij}\}$, where $1 \leq i \leq P, 1 \leq j \leq Q$. A target region R formed by $n \times n$ grid cells is a square geographic region within the city, where $n < \min(P, Q)$. The whole city can be split into many overlapping regions denoted as $\mathcal{R} = \{R_{ij}\}$, where $R_{ij} = \langle s_{ij}, n \rangle$ is uniquely defined by the grid cell s_{ij} in the top-left corner of the region and a number n indicating the side-length of the region.

Definition 2 (Traffic dynamics $\boldsymbol{X}^k_{R,t}$). Traffic dynamics, such as speed, volume, and density, are essential for characterizing the traffic status of the road network. We denote a type k

[1] Code: *https://github.com/Yliu1111/STGAIL.git*

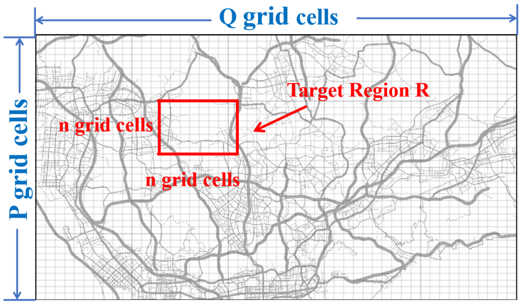


Fig. 2: Illustration example of a region with $n \times n$ grid cells in the roadmap.

dynamics for the target region R at time t by a matrix $\boldsymbol{X}^k_{R,t} \in \mathbb{R}^{n \times n}$, where each entry $x^k_{ij,t} \in \mathbb{R}$ corresponds to the type k dynamics for a specific grid cell s_{ij} within the target region R at time t. There are in total K distinct types of urban dynamics, thereby $1 \leq k \leq K$.

In our study, we have access to the historical traffic dynamics over a span of time denoted as $\{\boldsymbol{X}^k_{R,t}\}^T_{t=1}$ for dynamics type k and $1 \leq k \leq K$, where T denotes the series length of $\boldsymbol{X}^k$.

Definition 3 (Adjacency Matrix $\boldsymbol{A}^R$). The adjacency matrix $\boldsymbol{A}^R$ corresponds to a specific region R, indicating the traffic correlations and geographical patterns along the road networks, and thus describing the connections among grid cells s_{ij} within the target region R. $\boldsymbol{A}^R$ is a square matrix $\boldsymbol{A}^R \in \mathbb{R}^{n^2 \times n^2}$, and each entry $a_{(i-1)n+j,(i'-1)n+j'}$ is the traffic correlation between grid cells s_{ij} and $s_{i'j'}$ in region R.

Definition 4 (Traffic Dynamics Graph $\boldsymbol{G}^R_t$). A traffic dynamics graph $\boldsymbol{G}^R_t$, denoted as $\boldsymbol{G}^R_t = (V^R_t, E^R, \boldsymbol{A}^R)$, represents the traffic status along road networks at time t for region R. Here, $V^R_t = \{v_{ij,t}\}$ comprises a set of vertices, where each vertex $v_{ij,t}$ corresponds to a specific grid cell s_{ij} within region R at time t, and is associated with K types of traffic dynamic features, *i.e.*, $v_{ij,t} = [x^1_{ij,t}, \cdots, x^K_{ij,t}]^\intercal$. The set of edges E^R, representing connections among the nodes, is quantified by the adjacency matrix $\boldsymbol{A}^R$ indicating complicated road connectivity and geographical patterns.

As we have the traffic dynamics $\{\boldsymbol{X}^k_{R,t}\}^T_{t=1}$ for any specific region R over the time span where $1 < t < T$ for dynamic types $1 \leq k \leq K$ to represent vertices, and road network data to derive the edges and adjacency matrices, the corresponding traffic dynamics graphs at region R are obtained and denoted as $\boldsymbol{G}^R = \{\boldsymbol{G}^R_t\}^T_{t=1}$. We also denote $\mathcal{G} = \{\boldsymbol{G}^R\}_{R \in \mathcal{R}}$ as the set of all graph sequences.

B. Preliminaries: Latent Diffusion Models

Latent Diffusion Models (LDMs) [25] are an advanced variant of diffusion models [13] designed to improve computational and memory efficiency. By first compressing input data into a lower-dimensional latent space using a trained autoencoder, LDMs can focus on generating high-fidelity reconstructions from this reduced latent space. LDMs involves an encoder $E(\boldsymbol{x}) = \boldsymbol{z}$ that maps the input data $\boldsymbol{x} \sim \mathbb{P}_{data}$ to a latent representation $\boldsymbol{z}$, and a decoder $Q(\boldsymbol{z}) = \hat{\boldsymbol{x}}$ that reconstructs the data. The diffusion process in this latent space is designed to model the latent distribution via iterative denoising with denoising score matching [25]. It involves a forward diffusion process to gradually add Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ to latent samples $\boldsymbol{z}_\tau = \alpha_\tau \boldsymbol{z} + \sigma_\tau \epsilon$ where α_τ and σ_τ are a noise schedule parameterized by a diffusion-time τ, such that the logarithmic signal-to-noise ratio $\lambda_\tau = \log(\alpha^2_\tau / \sigma^2_\tau)$ decreases monotonically. It then trains a denoiser model f_θ parameterized by θ to denoise the diffused $\boldsymbol{z}_\tau$ by minimizing the denoising score matching objective:

$$\min_\theta \mathbb{E}_{\boldsymbol{x} \sim \mathbb{P}_{\text{data}}, \tau \sim p_\tau, \epsilon \sim \mathcal{N}(0,I)} \left[\left\| \epsilon - f_\theta(\boldsymbol{z}_\tau, \tau) \right\|^2_2 \right].$$

Here, the diffusion time τ is sampled from a uniform distribution $p_\tau \sim U\{1, O\}$, utilizing a cosine noise schedule to define α_τ and σ_τ with S being the diffusion time limit. Specifically, the relationship $\sigma^2_\tau = 1 - \alpha^2_\tau$ is maintained to ensure the preservation of variance throughout the diffusion process.

Limitations of LDMs in Traffic Dynamics Prediction. Recent literature has extensively explored the use of various latent diffusion models to learn diverse data distributions [3], [26], [27], [29], [33]. However, these models face significant challenges when applied directly to traffic dynamics prediction for the following reasons: *(i)* standard LDMs struggle to capture the spatial-temporal dependencies critical to accurately modeling traffic dynamics. The underlying mechanism of LDMs, based on the stochastic generation process of diffusion, inherently overlooks the temporal correlations among data points generated sequentially. This omission can lead to forecasts that fail to reflect real-world temporal dynamics in traffic patterns. *(ii)* To the best of our knowledge, no prior works of LDMs study the complex spatial patterns of traffic and the interconnections among various aspects of traffic dynamics. Traffic prediction requires not only predicting the status of traffic at individual locations but also understanding how these geographical locations influence each other. These limitations highlight the need for enhancements in LDM approaches to address the spatial-temporal intricacies of traffic dynamics.

C. Problem Definition & Challenges

Unlike prior works [50], [55] that either predicts short-term traffic dynamics or fixed-length long-term traffic dynamics, we target the problem that predicts traffic dynamics for a flexible period given an arbitrary length of historical observations all at once. Therefore, we formally define our problem below:

Problem Definition (Flexible Horizon Graph-based Traffic Dynamics Prediction). Given the historical traffic dynamics graphs $\boldsymbol{G}^R = \{\boldsymbol{G}^R_t\}^m_{t=1}$ for any region $R \in \mathcal{R}$ over a period of m time steps, our objective is to train a model f_θ, parameterized by θ, that can effectively capture the spatial and temporal dependencies inherent in the traffic data. This model should be able to predict future traffic dynamics graphs $\{\hat{\boldsymbol{G}}^R_t\}^{m+\ell}_{t=m+1}$ over a period of ℓ time steps for any region $R \in \mathcal{R}$.

Challenges. As illustrated in the introduction, the flexible horizon graph-based traffic dynamics prediction problem is

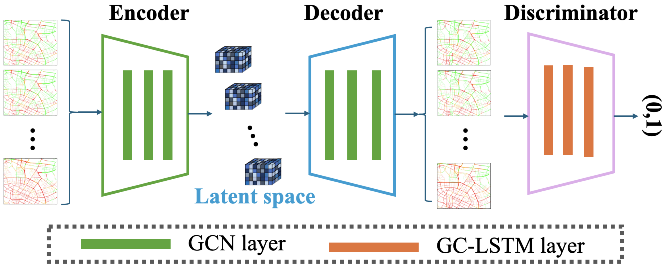


Fig. 3: Autoencoder and discriminator for pre-training the latent space.

challenging in the following aspects: (**C1**) How to capture and characterize the complicated spatial-temporal variation and flow in traffic dynamics graph sequences, and compress them into a lower-dimensional space? (**C2**) How to leverage latent diffusion models to learn the distribution of the complicated spatial-temporal graph sequences? (**C3**) How to adapt the learned knowledge to help predict future traffic dynamics of a flexible period at one time?

III. Methodology

Inspired by the LDMs (see Sec. II-B), we aim to address the traffic dynamics prediction problem using a novel spatial-temporal graph latent diffusion model, *i.e.*, STGAIL. STGAIL features innovative designs that address the inherent limitations of standard LDMs when applied in the spatial-temporal domain, and are specifically tailored to overcome the challenges outlined in Sec. II-C. These novel features include: *(i)* designing an autoencoder to compress traffic dynamics graph sequences into a lower-dimensional latent space and leveraging adversarial training to enhance the recognition of traffic pattern change over time (addressing challenge **C1**, see Sec. III-A); *(ii)* integrating LDMs and designing spatial and temporal graph layers to capture and enhance the spatial-temporal correlations captured in the compressed latent space (addressing challenge **C2**, see Sec. III-B); *(iii)* designing a masking approach to enable the learned latent diffusion model to function as a predictive model for generating flexible horizon predictions given an arbitrary length of historical observations (addressing challenge **C3**, see Sec. III-C).

A. Traffic Graph Compression with Adversarial Autoencoders

To address challenge **C1** – characterizing and compressing the complex spatial-temporal variations in traffic dynamics graph sequences into a lower-dimensional latent space – we pre-train an autoencoder inspired by prior works [10], [12]. Unlike standard autoencoders that primarily handle stochastic data samples, our autoencoder is specifically designed to process and encode sequential data. To better capture the sequential patterns in traffic dynamics graph sequences, we incorporate adversarial training with a discriminator.

The encoding process maps the input historical traffic dynamics graph sequence $\{\boldsymbol{G}_t\}_{t=1}^{T}$ into latent representations $\{\boldsymbol{z}_t\}_{t=1}^{T}$ using an encoder E_ϕ parameterized by ϕ implemented with multiple layers of Graph Convolutional Networks (GCNs) [16] to capture the complicated graph information. The decoder Q_ψ parameterized by ψ then reconstructs these graph sequences from the latent representations, as illustrated in Fig. 3. The objective for training the encoder-decoder is:

$$\min_{\phi,\psi} \mathcal{L}_{E,Q} = \mathbb{E}_{\{\boldsymbol{G}_t\}_{t=1}^{T} \sim \mathcal{G}} \left[\frac{1}{T} \sum_{t=1}^{T} \|\boldsymbol{G}_t - \hat{\boldsymbol{G}}_t\|_2^2 \right], \tag{1}$$

where $\hat{\boldsymbol{G}}_t = Q_\phi(E_\psi(\boldsymbol{G}_t))$ represents the graph reconstructed at time t. This objective function compels the encoder to minimize discrepancies between the actual and reconstructed graphs, thereby effectively learning a latent representation of sequential traffic dynamics.

However, creating a robust latent space involves more than accurate graph reconstruction; it also requires verifying the temporal coherence of the reconstructed graphs. To enhance the quality of the latent space, we incorporate a discriminator D_ω parameterized by ω with adversarial training. As depicted in Fig. 3, the discriminator D_ω utilizes Graph Convolutional Long Short-Term Memory (GC-LSTM) networks [5], which are particularly effective at capturing both spatial and temporal dependencies in traffic data. The discriminator's role is to assess the reliability and temporal consistency of the reconstructed traffic sequences, assigning an output of 1 for real graph sequences that exhibit strong temporal coherence, and 0 for graphs generated by the encoder. Both the autoencoder and discriminator work in an adversarial way, where the former aiming to bypass detection by producing increasingly accurate encodings, while the latter strives to detect discrepancies. The objective function of the discriminator is formulated as follows:

$$\begin{aligned} \max_{\omega} \mathcal{L}_D = \mathbb{E}_{\{\boldsymbol{G}_t\}_{t=1}^{T} \sim \mathcal{G}} \Big[& \log\big(1 - D_\omega(\{\boldsymbol{G}_t\}_{t=1}^{T})\big) \\ & + \log\big(D_\omega(\{\hat{\boldsymbol{G}}_t\}_{t=1}^{T})\big)\Big]. \end{aligned} \tag{2}$$

Here, the discriminator D_ω tries to distinguish between real traffic dynamics graph sequences $\{\boldsymbol{G}_t\}_{t=1}^{T}$ and generated ones $\{\hat{\boldsymbol{G}}_t\}_{t=1}^{T}$. The overall objective function for pre-training combines the contributions of both the autoencoder (E_ϕ, Q_ψ) and the discriminator D_ω, *i.e.*,

$$\min_{\phi,\psi} \max_{\omega} \mathcal{L}_{\text{pretrain}} = \mathcal{L}_{E,Q} + \mathcal{L}_D. \tag{3}$$

The autoencoder and the discriminator are trained iteratively under this final objective in Eq. (3), with $\mathcal{L}_D$ also influencing the parameter updates for the encoder E_ϕ and decoder Q_ψ. This integrated adversarial training approach enables the learning of a high-quality, reliable latent space, crucial for building our STGAIL in the next steps.

B. Spatial-Temporal Graph Latent Diffusion Model

To address challenge **C2**, we design the Spatial-Temporal Graph Latent Diffusion Model, in short, STGAIL. STGAIL aims to learn the distribution of traffic dynamics graph sequences more effectively within the latent space $\boldsymbol{\mathcal{Z}} = \{\{\boldsymbol{z}_t\}_{t=1}^{T} | \boldsymbol{z}_t = E(\boldsymbol{G}_t),\ \text{for}\ \boldsymbol{G}_t \in \{\boldsymbol{G}_t\}_{t=1}^{T}, \{\boldsymbol{G}_t\}_{t=1}^{T} \in \mathcal{G}\}$,

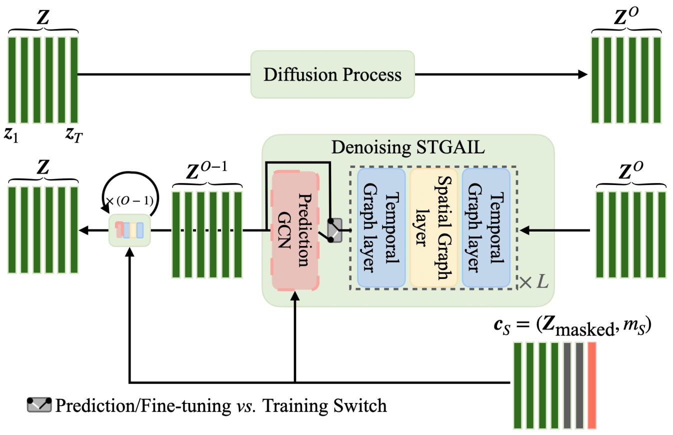


Fig. 4: STGAIL diffusion model.

and perform traffic dynamics graph generation. STGAIL consists of two novel designs: (1) a *STGAIL diffusion design* to learn the distribution of latent graph sequences, and (2) a *STGAIL diffusion backbone* to better capture complicated spatial-temporal correlations of traffic dynamics in STGAIL.

STGAIL diffusion design. To better learn the distribution of traffic dynamics graph latent sequences $\mathcal{Z}$, we design diffusion models tailored for sequential data. First, we stack each traffic dynamics graph latent sequence $\{z_t\}_{t=1}^{T}$ into a latent matrix $\boldsymbol{Z} = [z_1, \cdots, z_T]$, and apply diffusion on these latent matrices via iterative denoising with denoising score matching. STGAIL adds Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ to latent samples $\boldsymbol{Z}_\tau = \alpha_\tau \boldsymbol{Z} + \sigma_\tau \epsilon$, where α_τ and σ_τ are parameters in a noise schedule governed by diffusion time τ, such that the logarithmic signal-to-noise ratio $\lambda_\tau = \log(\alpha_\tau^2/\sigma_\tau^2)$ decreases monotonically. The model then trains a denoiser f_θ, parameterized by θ, to denoise $\boldsymbol{Z}_\tau$ by minimizing the denoising score matching objective:

$$\min_{\theta} \mathbb{E}_{\boldsymbol{Z} \sim \mathbb{P}_{\text{data}}, \tau \sim p_\tau, \epsilon \sim \mathcal{N}(0,I)} \left[\| \epsilon - f_\theta(\boldsymbol{Z}_\tau, \tau) \|_2^2 \right]. \tag{4}$$

Here, the diffusion time τ is sampled from a uniform distribution $p_\tau \sim U\{1, O\}$, using a cosine noise schedule to define α_τ and σ_τ. Specifically, $\sigma_\tau^2 = 1 - \alpha_\tau^2$ is maintained to ensure the preservation of variance throughout the diffusion process.

STGAIL diffusion backbone. STGAIL's architecture extracts spatial and temporal features for traffic dynamics prediction. It integrates graph convolutions and gated temporal convolutions within a denoiser model f_θ to capture dependencies in traffic networks. Each spatial-temporal block consists of two gated temporal graph layers and a central spatial graph layer, efficiently processing graph-structured time series data. *Temporal Graph Layers* in STGAIL employ 1-D causal convolutions followed by gated linear units (GLU), leveraging the rapid training and simpler structure of CNNs over RNNs. Specifically, the output of the temporal graph layer is:

$$\mathbf{H}_{\text{temporal}} = \text{GLU}(\mathbf{W}_{\text{temp}} * \mathbf{Z} + \mathbf{b}_{\text{temp}}),$$

where $\mathbf{W}_{\text{temp}}$ and $\mathbf{b}_{\text{temp}}$ are the weights and biases of the temporal graph layer, and $*$ denotes the convolution operation. The GLU activation function is $\text{GLU}(\mathbf{A}, \mathbf{B}) = \mathbf{A} \otimes \sigma(\mathbf{B})$ where σ is the sigmoid function and $\otimes$ denotes element-wise multiplication.

Spatial Graph Layers operates directly on graph-structured data. These convolutions employ Chebyshev polynomials and first-order approximation strategies [11], [45], [46] to extract significant spatial patterns. The spatial graph convolution is:

$$\mathbf{H}_{\text{spatial}} = \sum_{m=0}^{M} \mathbf{T}_m(\tilde{\mathbf{L}}) \mathbf{Z} \mathbf{W}_m,$$

where $\mathbf{T}_m(\tilde{\mathbf{L}})$ are the Chebyshev polynomials of the scaled Laplacian $\tilde{\mathbf{L}}$, $\mathbf{Z}$ is the input feature matrix, $\mathbf{W}_m$ are the trainable weights, and M is the order of the polynomial.

Each spatial-temporal block in STGAIL combines these layers to capture both spatial and temporal dependencies. The output of a spatial-temporal block can be expressed as: $\mathbf{H}_{\text{block}} = g_\phi(\mathbf{H}_{\text{temporal}}, \mathbf{H}_{\text{spatial}})$ where g_ϕ denotes the integration function of the temporal and spatial outputs. With this design, the model efficiently handles traffic graph sequence and captures crucial spatial and temporal features.

C. STGAIL Flexible Horizon Prediction

After getting a well-trained STGAIL, capable of generating sequential graphs of traffic dynamics for specific regions from random noise, we introduce our prediction solution to address challenge **C3**. Notice that directly using the diffusion model cannot make predictions. Our solution enables STGAIL to function as a predictive model that generates future traffic predictions based on historical traffic data rather than solely on random inputs.

We employ a temporal binary mask m_S to manage varying lengths of traffic sequences that need prediction, adapting STGAIL to perform traffic prediction tailored to specific regions. This process utilizes the pre-trained latent space, where the temporal binary mask, m_S, masks the $T - S$ latent traffic dynamics graphs that STGAIL is tasked to predict. Here, T represents the total sequence length of the traffic dynamics, and S is a variable of the historical graph sequence lengths that the prediction builds upon.

To better prepare STGAIL for prediction tasks, we append a prediction layer parameterized by γ implemented with GCN [16] at the of the diffusion model and fine-tune them. In the fine-tuning process, the masked latent graph sequence $\boldsymbol{Z}_{\text{masked}} = m_S \circ \boldsymbol{Z}$ is concatenated with the mask m_S itself and input into the prediction layer. Formally, let $\boldsymbol{c}_S = (\boldsymbol{Z}_{\text{masked}}, m_S)$ denote the concatenated conditioning of masks and masked latent graphs, the objective for fine-tuning is,

$$\min_{\theta,\gamma} \mathbb{E}_{\boldsymbol{Z} \sim \mathbb{P}_{\text{data}}, m_S \sim p_S, \tau \sim p_\tau, \epsilon \sim \mathcal{N}(0,I)} \left[\| \epsilon - f_{\theta,\gamma}(\boldsymbol{Z}_\tau; \boldsymbol{c}_S, \tau) \|_2^2 \right], \tag{5}$$

where $\boldsymbol{Z}_\tau$ denotes diffused encodings, p_S represents the mask sampling distribution, and ϵ is the Gaussian noise vector $\epsilon \sim \mathcal{N}(0, I)$.

Algorithm 1 Algorithm for STGAIL

```
Input: Traffic dynamics graphs $\mathcal{G}$ and initialized parameters
    for encoder $E_\phi$, decoder $Q_\psi$, discriminator $D_\omega$, and
    denoiser model $f_\theta$; batch size $b$, noise schedule $\alpha_\tau$, $\sigma_\tau$.
Output: Well-trained encoder $E_\phi$, decoder $Q_\psi$, discriminator $D_\omega$, and denoiser model $f_\theta$.
Part 1: Traffic Graph Compression:
for iteration $i = 1, 2, 3, \cdots$ do
    Sample a batch of $b$ graph sequences $\mathcal{G}_i = \{\{\boldsymbol{G}_t\}_{t=1}^T\}$,
        where $\mid \mathcal{G}_i \mid = b$.
    Encode $\mathcal{G}_i$ and decode to get the reconstruction $\hat{\mathcal{G}}_i =
        \{\{\hat{\boldsymbol{G}}_t^R\}_{t=1}^T\}$ and calculate Eq. (1).
    Evaluate real and generated sequences $\mathcal{G}_i$ and $\hat{\mathcal{G}}_i$ respectively and calculate Eq. (2).
    Update $\phi$, and $\psi$ to minimize Eq. (3).
    Update $\omega$ to maximize Eq. (3).
end for
Obtain well-trained $E_\phi$, $Q_\psi$ and $D_\omega$.
Part 2: STGAIL Training:
for iteration $= 1, 2, 3, \cdots$ do
    Sample a batch of $b$ graph sequences $\mathcal{G}_i = \{\{\boldsymbol{G}_t\}_{t=1}^T\}$,
        where $\mid \mathcal{G}_i \mid = b$.
    Get latent graphs $\mathcal{Z}_i = \{\{\boldsymbol{z}_t\}_{t=1}^T\}$ with $E_\phi$, and stack
        each sequence to get $\tilde{\mathcal{Z}}_i = \{\boldsymbol{Z}\}$.
    Sample a batch of $b$ time steps $\{\tau\}$ with $\tau \sim U\{1, I\}$
        and noises $\{\epsilon\}$ with $\epsilon \sim \mathcal{N}(0, I)$ respectively and
        construct noisy latents.
    Update $f_\theta$ using Eq. (4).
end for
A pre-trained model $f_\theta$ is produced.
Part 3: Flexible Horizon Prediction:
for iteration $i = 1, 2, 3, \cdots$ do
    Sample one graph sequence $\{\boldsymbol{G}_t\}_{t=1}^T$, and one $S \sim
        U\{1, T\}$, diffusion time $\tau \sim U\{1, O\}$ and noise $\epsilon \sim
        \mathcal{N}(0, I)$.
    Get latent graphs $\boldsymbol{Z}$ with $E_\phi$, and condition $\boldsymbol{c}_S$.
    Update $f_{\theta,\gamma}$ using Eq. (5).
end for
```

This method allows STGAIL to generate extended traffic prediction sequences from initial historical data, preserving temporal dependencies and consistency. Consequently, STGAIL is transformed into a traffic prediction model that adapts to various regions and prediction lengths, providing accurate and efficient traffic predictions over multiple time steps. The algorithm for our STGAIL is shown in Alg. 1.

IV. Experiment

In our experiments, we aim to evaluate the effectiveness of our STGAIL for traffic prediction tasks. We will answer the following questions with extensive experiments: (1) Can our model accurately predict different traffic dynamics compared to other state-of-the-art methods? (2) Can our STGAIL outperform auto-regressive baselines in traffic dynamics prediction in terms of accuracy and efficiency? (3) Are different components of our STGAIL, including the autoencoder, discriminator, and spatial-temporal graph layer, effective in enhancing prediction accuracy? (4) How do different hyperparameters affect the performance of our STGAIL?

A. Dataset and Experiment Descriptions

Data Description. We evaluate our STGAIL using three real-world urban dynamics datasets from Shenzhen, China, spanning from July 1st to December 31st, 2016. These datasets include: (1) traffic speed, (2) taxi inflow, and (3) travel demand. The urban area of Shenzhen is divided into 40×50 equal-sized grid cells, which are further aggregated into 63 regions. Each region, containing a 10×10 grid, is treated as a graph where each grid cell represents a node. Consequently, each graph comprises 10×10 nodes, with each node associated with different traffic dynamics features including traffic speed, taxi inflow, and travel demand. More details about our urban dynamics datasets are as follows:

- **Traffic Speed:** The dataset includes traffic speed measurements across 63 regions, captured in 4416 one-hour intervals over a six-month period. For each grid cell and time slot, the average traffic speed is determined by dividing the total travel distance by the elapsed time.
- **Taxi Inflow:** This dataset records the number of taxis arriving in each of the 63 regions, also measured in 4416 one-hour intervals over six months. The taxi inflow for each grid cell is quantified by tallying the total arrivals within each hour.
- **Travel Demand:** This dataset tracks hourly taxi pickups and drop-offs in 63 regions over the same time frame. Due to the challenge of collecting comprehensive travel demand data across all transportation modes, it focuses on taxi services. Several studies have confirmed the relevance and effectiveness of using taxi data as a proxy for overall travel demand [51]–[54].

In summary, all three datasets have been aggregated and processed to yield dimensions of $(162 \times 63, 12, 100, 3)$. Here, 162 represents the number of days, 12 indicates the twelve one-hour time slots per day, 63 corresponds to the number of regions or graphs, 100 denotes the number of nodes per graph, and 3 signifies the number of features per node.

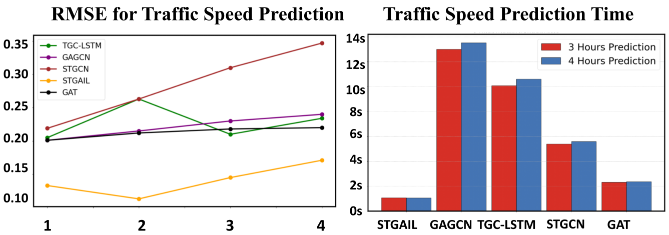
Adjacency Matrix. The adjacency matrices for traffic dynamics graphs are derived using the Pearson correlation coefficient to measure the traffic correlation between nodes. The Pearson correlation coefficient, $\rho_{X,Y}$, for two time series X and Y is given by:

$$\rho_{X,Y} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}\sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}},$$

where X_i and Y_i are the traffic values at time i, and $\bar{X}$ and $\bar{Y}$ are the average values of the time series X and Y. This method creates an adjacency matrix that represents the strength of linear relationships between nodes, forming the foundation of our graph-based model.

TABLE II: Performance of traffic dynamics prediction. One-hour and two-hour prediction results are provided in this table.

	Dataset	Metric	**STGAIL**	DYffusion	GAGCN	TGC-LSTM	STGCN	GCRN	GAT	GMAN	VideoLDM
One-hour	Traffic speed	RMSE	**0.133**	0.215	0.201	0.205	0.219	0.185	0.201	0.513	0.134
		MAPE(%)	**33.1**	83.0	95.3	34.8	51.3	83.7	184.7	214.0	65.32
	Taxi inflow	RMSE	**0.146**	0.239	0.219	0.229	0.515	0.170	0.185	0.623	0.165
		MAPE(%)	**10.5**	42.5	57.7	108.7	152.3	39.0	47.1	85.5	11.9
	Travel demand	RMSE	**0.168**	0.291	0.264	0.254	0.730	0.279	0.250	0.757	0.170
		MAPE(%)	**68.3**	69.1	85.8	104.8	78.5	70.0	70.9	116.4	158.1
Two-hour	Traffic speed	RMSE	**0.113**	0.215	0.215	0.209	0.263	0.197	0.212	0.564	**0.113**
		MAPE(%)	**35.5**	110.4	136.2	36.2	70.0	56.5	104.1	134.2	72.39
	Taxi inflow	RMSE	**0.230**	0.241	0.249	0.284	0.391	0.238	0.242	0.433	0.232
		MAPE(%)	**14.7**	41.5	47.0	126.7	213.0	39.1	46.1	73.0	15.7
	Travel demand	RMSE	**0.215**	0.347	0.313	0.303	0.709	0.288	0.309	0.463	0.238
		MAPE(%)	**69.9**	74.4	74.1	128.6	217.7	108.3	72.3	121.4	167.5



a) Traffic speed prediction from one-hour to four-hour prediction. **b) Time efficiency of STGAIL and auto-regressive baselines.**

Fig. 5: Results of STGAIL and auto-regression models on the performance of traffic speed prediction.

B. Baselines

We use graph-based auto-regressive models and non-auto-regressive models including diffusion models as baselines to evaluate our STGAIL. The *non auto-regressive baselines* include:

- **DYffusion** [3]: DYffusion is an advanced diffusion model for probabilistic spatial-temporal forecasting, designed to generate stable and accurate rollout forecasts. By coupling temporal dynamics with diffusion steps, it is well-suited for traffic dynamics prediction.
- **VideoLDM** [2]: VideoLDM is designed for high-resolution video generation using a diffusion model in a compressed latent space. Incorporating a temporal dimension and fine-tuning on traffic data, it captures spatial-temporal dependencies for traffic predictions.
- **GMAN** [57]: GMAN addresses long-term traffic prediction using an encoder-decoder architecture with attention blocks. A transform attention layer links the encoder and decoder, converting features into future sequences and reducing error propagation.

The *auto-regressive baselines* include:

- **GAGCN** [38]: GAGCN leverages graph attention networks to dynamically capture and adjust spatial associations among nodes over time, effectively modeling complex spatial and temporal dependencies for accurate traffic predictions.
- **TGC-LSTM** [7]: TGC-LSTM captures time-varying patterns and spatial dependencies using traffic and spectral graph convolutions, making it effective for traffic dynamics prediction.
- **STGCN** [41]: STGCN effectively handles the complexities and nonlinearity of traffic flow, making it suitable for mid- and long-term predictions. It formulates the problem on graphs using a fully convolutional architecture, avoiding standard convolutional and recurrent units.
- **GCRN** [28]: GCRN extends RNNs to graph-structured data, modeling spatial and temporal dependencies to effectively predict complex spatial-temporal traffic patterns.
- **GAT** [37]: GATs use masked self-attentional layers to handle graph-structured data, overcoming traditional graph convolutional limitations. We enhance this by sequentially passing data through the model for each hour, allowing it to use previous predictions as inputs for forecasting subsequent hours and capturing temporal dependencies effectively.

C. Evaluation Metrics

We use mean absolute percentage error (MAPE) and root mean square error (RMSE) for evaluation:

$$\text{MAPE} = \frac{1}{N_s N_t} \sum_{s=1}^{N_s} \sum_{t=1}^{N_t} \left| \frac{y_{s,t} - \hat{y}_{s,t}}{y_{s,t}} \right|,$$

$$\text{RMSE} = \sqrt{\frac{1}{N_s N_t} \sum_{s=1}^{N_s} \sum_{t=1}^{N_t} (y_{s,t} - \hat{y}_{s,t})^2},$$

where N_s is the number of nodes for a graph ($N_s = 100$ in this work), N_t is the number of predicted time slots, $y_{s,t}$ represents the ground-truth traffic dynamics observed in the s-th node at the t-th time slot, and $\hat{y}_{s,t}$ is the corresponding predicted value.

D. Experimental Settings

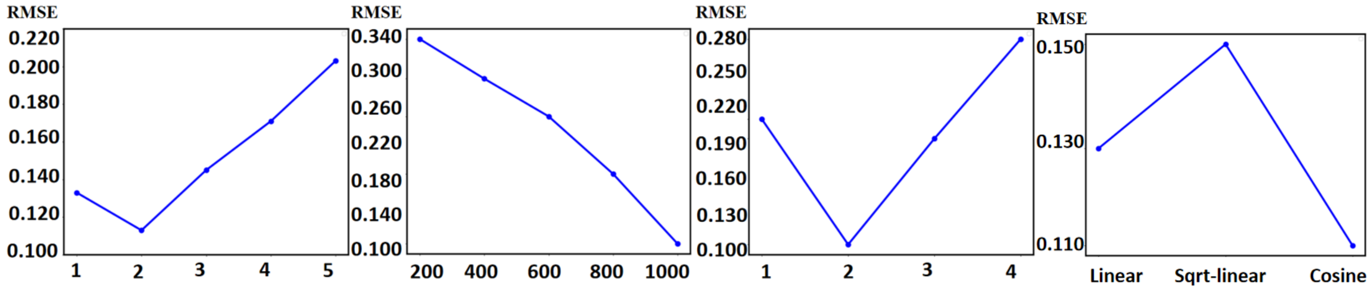
All experiments use the Adam optimizer with an initial learning rate of 0.0001. For prediction tasks, 50 regions are used for training and the remaining 13 for testing, training with one node feature at a time.

E. Empirical Results

Results of Question (1). To assess the performance of STGAIL in traffic dynamics prediction, we conducted experiments on three different urban dynamics datasets: traffic speed, taxi inflow, and travel demand. We performed both one-hour and two-hour predictions and compared the results against state-of-the-art baseline models mentioned in Sec. IV-B. As shown in Tab. II, our model consistently outperforms the baseline models across all datasets and metrics. For one-hour

TABLE III: Ablation study: Impacts of different components in STGAIL. We assess the importance of the Autoencoder, Discriminator, and Temporal graph layer in STGAIL and show the corresponding urban dynamics prediction performance.

	Dataset	Metric	**STGAIL**	STGAIL w/o Autoencoder	STGAIL w/o Discriminator	STGAIL w/o Spatial-temporal graph layer
One-hour	Traffic speed	RMSE	**0.133**	0.206	0.240	0.192
		MAPE(%)	**33.1**	67.7	61.7	54.3
	Taxi inflow	RMSE	**0.146**	0.204	0.268	0.188
		MAPE(%)	**10.5**	13.5	11.0	16.9
	Travel demand	RMSE	**0.168**	0.219	0.180	0.187
		MAPE(%)	**68.2**	80.3	111.3	214.3
Two-hour	Traffic speed	RMSE	**0.113**	0.210	0.256	0.272
		MAPE(%)	**35.5**	51.4	50.0	78.5
	Taxi inflow	RMSE	**0.230**	0.295	0.266	0.266
		MAPE(%)	**14.7**	20.4	18.5	19.7
	Travel demand	RMSE	**0.215**	0.352	0.256	0.266
		MAPE(%)	**69.9**	75.8	117.3	216.6



a) Mask length L_{mask} b) Diffusion time τ c) Layers N_L d) Noise schedule σ_τ and α_τ

Fig. 6: Impacts of hyperparameters on the RMSE performance of traffic speed prediction.

forecasts, our model shows the lowest RMSE and MAPE values in predicting traffic speed, taxi inflow, and travel demand. This trend of superior performance extends to two-hour predictions, where our STGAIL model continues to excel. The effectiveness of STGAIL in producing more accurate predictions than other baseline models stems from its capability to capture both spatial and temporal dependencies in the data. By integrating a latent diffusion model with innovative spatial and temporal graph layers, STGAIL effectively learns the complex relationships inherent in traffic dynamics, ensuring both spatial coherence and temporal consistency. Additionally, the fine-tuning process enables STGAIL to quickly adapt to specific regional conditions, enhancing its functionality as a predictive tool and helping to minimize prediction errors.

Results of Question (2). As illustrated in Fig. 5, we conducted a comparative analysis to assess the performance of STGAIL against traditional auto-regressive models in predicting traffic speed. The experiment was conducted over one to four-hour traffic speed prediction using auto-regression baselines, including TGC-LSTM, GAGCN, STGCN, and GAT. The results, depicted in Fig. 5, consistently show that STGAIL outperforms these models across all prediction tasks. Additionally, considering STGAIL's capability to generate predictions for multiple time steps simultaneously, we also examined its time efficiency in comparison to auto-regressive baselines. The findings, as shown in Fig. 5, indicate that STGAIL significantly reduces prediction time, especially for three to four-hour forecasts, outperforming auto-regressive models including TGC-LSTM, GAGCN, STGCN, and GAT. These results collectively underscore the superior effectiveness and efficiency of STGAIL over traditional auto-regressive models.

Results of Question (3) (Ablation Study). We conducted an ablation study to assess the significance of each component in our STGAIL model, comparing the full STGAIL model against three variants: **STGAIL w/o Autoencoder** (i.e., excluding the autoencoder), **STGAIL w/o Discriminator** (i.e., excluding the discriminator), and **STGAIL w/o Spatial-Temporal Graph Layer** (i.e., substituting the spatial-temporal graph layer with a GCN). These comparisons elucidate the contribution of each component to the model's efficacy in predicting traffic dynamics. In Tab. III, the omission of any component results in increased RMSE and MAPE, underscoring their collective importance in enhancing prediction accuracy. Notably, the complete STGAIL model consistently outperforms its ablated versions across all metrics and datasets. Detailed analysis shows that while the STGAIL model without the spatial-temporal graph layer has a similar RMSE to the full model for short-term traffic speed predictions, it gets a higher MAPE. This suggests that the spatial-temporal graph layer significantly reduces relative error, improving forecast accuracy, especially for longer horizons and more variable factors like taxi inflow and travel demand. Furthermore, performance comparisons reveal that removing either the autoencoder or the discriminator degrades performance compared to the full STGAIL model. The autoencoder creates a perceptually similar but simpler latent space, while the discriminator assesses the reliability and temporal consistency of the reconstructed traffic sequences. Together, these components enhance the latent space quality and overall predictive performance.

Results of Question (4). To study how different hyperparameters affect the performance, we select four major hyperparameters including the the length of the binary temporal mask L_{MASK}, diffusion time τ, numbers of layers in the STGAIL N_L and the type of noise schedule method for diffusion, and test how different values of these hyperparameters affect the traffic speed prediction performance. As shown in Fig. 6(a), we find that the RMSE performance is optimal when the mask length L_{MASK} is 2. As the mask length increases from 2 to 5, the RMSE progressively worsens. Fig. 6(b) shows that the longer diffusion time τ goes on, the better traffic prediction performs. As shown in Fig. 6(c), if the number of graph layers in the STGAIL, N_L is 2, we can get the best prediction performance. As shown in Fig. 6(d), where we adjusted the noise schedule method, it is evident that with cosine schedule, the model is better capable of predicting accurate traffic speed.

V. Related Work

Traffic dynamics prediction is key for traffic management, urban planning, and intelligent transportation systems. Research ranges from traditional machine learning to advanced deep learning models. Key contributions include integrating temporal-spatial dependencies in models like TD^2-DL [40] for speed predictions and using meta-learning approaches such as cST-ML [55] and DAC-ML [50] to adapt to changing traffic conditions. Generative models like TrafficGAN [51] and STrans-GAN [53] estimate traffic across diverse scenarios, while LSGCN [14] captures complex spatiotemporal features for stable forecasts. Models like STGCN [41] and GMAN [57] enhance traffic flow predictions, and frameworks combining traditional theories with machine learning improve volume and flow forecasts [36], [44]. Advanced methods like ConvLSTM for accidents [43] and CNNs for citywide flow [47] emphasize spatial dependencies and real-time processing. However, focusing on single dynamics and using auto-regressive models often lead to error accumulation and high computation costs. Our model addresses these issues by capturing diverse traffic dynamics for accurate, efficient multi-step predictions.

Diffusion models transform Gaussian noise into detailed data distributions via reverse diffusion, enhancing image synthesis [9]. Sohl-Dickstein *et al.* [32] first modeled this process as a Markov chain. Latent diffusion models, operating in compressed spaces, offer efficiency and adaptability, as demonstrated in high-quality image generation from text [24] and conditional generation [13]. They also excel in trajectory generation [49] and video generation [23]. Despite their success, diffusion models are underexplored in traffic dynamics prediction. This paper introduces a novel latent diffusion model to forecast diverse traffic dynamics, effectively capturing spatial-temporal dependencies.

VI. Conclusion

In conclusion, this paper introduces the Spatial-Temporal Graph Latent Diffusion Model (STGAIL), a novel approach that efficiently predicts future traffic dynamics, overcoming the limitations of traditional auto-regressive methods. STGAIL operates within a pre-trained graph latent space to significantly reduce computational complexity. This model integrates novel spatial and temporal graph layers within a latent diffusion framework, enhancing the learning of complex traffic dynamics. Furthermore, our innovative fine-tuning approach ensures accurate, concurrent predictions across multiple time steps, reducing both error accumulation and computational demands. Extensive experiments with real-world datasets confirm that STGAIL outperforms existing SOTA, significantly enhancing the accuracy and efficiency of traffic dynamics predictions.

VII. Acknowledgments

Yanhua Li was supported in part by NSF grants IIS-1942680 (CAREER), CNS-1952085 and DGE-2021871. Jun Luo is supported by The Innovation and Technology Fund (Ref. ITP/069/23LP).

Disclaimer: Any opinions, findings, conclusions or recommendations expressed in this material/event (or by members of the project team) do not reflect the views of the Government of the Hong Kong Special Administrative Region, the Innovation and Technology Commission or the Innovation and Technology Fund Research Projects Assessment Panel.

References

[1] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, 28, 2015.

[2] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023.

[3] S. R. Cachay, B. Zhao, H. James, and R. Yu. Dyffusion: A dynamics-informed diffusion model for spatiotemporal forecasting. *arXiv preprint arXiv:2306.01984*, 2023.

[4] M. Castro-Neto, Y.-S. Jeong, M.-K. Jeong, and L. D. Han. Online-svr for short-term traffic flow prediction under typical and atypical traffic conditions. *Expert systems with applications*, 36(3):6164–6173, 2009.

[5] J. Chen, X. Wang, and X. Xu. Gc-lstm: Graph convolution embedded lstm for dynamic link prediction, 2021.

[6] Y. Cong, J. Wang, and X. Li. Traffic flow forecasting by a least squares support vector machine with a fruit fly optimization algorithm. *Procedia Engineering*, 137:59–68, 2016.

[7] Z. Cui, K. Henrickson, R. Ke, and Y. Wang. Traffic graph convolutional recurrent neural network: A deep learning framework for network-scale traffic learning and forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 21(11):4883–4894, 2019.

[8] Z. Cui, R. Ke, Z. Pu, and Y. Wang. Deep bidirectional and unidirectional lstm recurrent neural network for network-wide traffic speed prediction. *arXiv preprint arXiv:1801.02143*, 2018.

[9] P. Dhariwal and A. Nichol. Improved denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 34:16925–16938, 2021.

[10] C. Feichtenhofer, Y. Li, K. He, et al. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35:35946–35958, 2022.

[11] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, page 1126–1135, 2017.

[12] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[13] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[14] R. Huang, C. Huang, Y. Liu, G. Dai, and W. Kong. Lsgcn: Long short-term traffic prediction with graph convolutional networks. In C. Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 2355–2361. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track.

[15] J. Ji, J. Wang, C. Huang, J. Wu, B. Xu, Z. Wu, J. Zhang, and Y. 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.

[16] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

[17] F. Li, J. Feng, H. Yan, G. Jin, F. Yang, F. Sun, D. Jin, and Y. Li. Dynamic graph convolutional recurrent network for traffic prediction: Benchmark and solution. *ACM Transactions on Knowledge Discovery from Data*, 17(1):1–21, 2023.

[18] L. Liu, J. Zhen, G. Li, G. Zhan, Z. He, B. Du, and L. Lin. Dynamic spatial-temporal representation learning for traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*, 22(11):7169–7183, 2020.

[19] X. Luo, L. Niu, and S. Zhang. An algorithm for traffic flow prediction based on improved sarima and ga. *KSCE Journal of Civil Engineering*, 22(10):4107–4115, 2018.

[20] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang. Traffic flow prediction with big data: a deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):865–873, 2014.

[21] C. Ma, G. Dai, and J. Zhou. Short-term traffic flow prediction for urban road sections based on time series analysis and lstm_bilstm method. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):5615–5624, 2021.

[22] X. Ma, Z. Tao, Y. Wang, H. Yu, and Y. Wang. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transportation Research Part C: Emerging Technologies*, 54:187–197, 2015.

[23] A. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models for video generation. *arXiv preprint arXiv:2112.10752*, 2021.

[24] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022.

[25] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. *CoRR*, abs/2112.10752, 2021.

[26] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[27] R. Rombach, A. Blattmann, and B. Ommer. Text-guided synthesis of artistic images with retrieval-augmented diffusion models. *arXiv preprint arXiv:2207.13038*, 2022.

[28] Y. Seo, M. Defferrard, P. Vandergheynst, and X. Bresson. Structured sequence modeling with graph convolutional recurrent networks. In *Neural Information Processing: 25th International Conference, ICONIP 2018, Siem Reap, Cambodia, December 13-16, 2018, Proceedings, Part I 25*, pages 362–373. Springer, 2018.

[29] J. Shi, C. Wu, J. Liang, X. Liu, and N. Duan. Divae: Photorealistic images synthesis with denoising diffusion decoder. *arXiv preprint arXiv:2206.00386*, 2022.

[30] X. SHI, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. WOO. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[31] X. Shi, H. Qi, Y. Shen, G. Wu, and B. Yin. A spatial–temporal attention approach for traffic prediction. *IEEE Transactions on Intelligent Transportation Systems*, 22(8):4909–4918, 2020.

[32] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *International Conference on Machine Learning*, pages 2256–2265, 2015.

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

[34] Y. Sun, B. Leng, and W. Guan. A novel wavelet-svm short-time passenger flow prediction in beijing subway system. *Neurocomputing*, 166:109–121, 2015.

[35] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

[36] E. Toto, E. A. Rundensteiner, Y. Li, R. Jordan, M. Ishutkina, K. Claypool, J. Luo, and F. Zhang. Pulse: A real time system for crowd flow prediction at metropolitan subway stations. In *ECMLPKDD*, 2016.

[37] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[38] M. Xia, D. Jin, and J. Chen. Short-term traffic flow prediction based on graph convolutional networks and federated learning. *IEEE Transactions on Intelligent Transportation Systems*, 24(1):1191–1203, 2022.

[39] P. Xie, T. Li, J. Liu, S. Du, X. Yang, and J. Zhang. Urban flow prediction from spatiotemporal data using machine learning: A survey. *Information Fusion*, 59:1–12, 2020.

[40] H. Yang, L. Du, G. Zhang, and T. Ma. A traffic flow dependency and dynamics based deep learning aided approach for network-wide traffic speed propagation prediction. *Transportation Research Part B: Methodological*, 167:99–117, 2023.

[41] B. Yu, H. Yin, and Z. Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, IJCAI-2018. International Joint Conferences on Artificial Intelligence Organization, July 2018.

[42] H. Yu, Z. Wu, S. Wang, Y. Wang, and X. Ma. Spatiotemporal recurrent convolutional networks for traffic prediction in transportation networks. *Sensors*, 17(7):1501, 2017.

[43] Z. Yuan, X. Zhou, and T. Yang. Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 984–992, 2018.

[44] X. Zhan, Y. Zheng, X. Yi, and S. Ukkusuri. Citywide traffic volume estimation using trajectory data. *TKDE*, 29(2):272–285, 2017.

[45] H. Zhang, L. Cao, P. VanNostrand, S. Madden, and E. A. Rundensteiner. Elite: Robust deep anomaly detection with meta gradient. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2174–2182, 2021.

[46] H. Zhang, B. Yan, L. Cao, S. Madden, and E. Rundensteiner. Metastore: Analyzing deep learning meta-data at scale. *Proceedings of the VLDB Endowment*, 17(6):1446–1459, 2024.

[47] J. Zhang, Y. Zheng, and D. Qi. Deep spatio-temporal residual networks for citywide crowd flows prediction. *CoRR*, 2016.

[48] J. Zhang, Y. Zheng, and D. Qi. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[49] X. Zhang, Y. Li, Z. Zhang, C. G. Brinton, Z. Liu, and Z.-L. Zhang. Distributional cloning for stabilized imitation learning via admm. In *2023 IEEE International Conference on Data Mining (ICDM)*, pages 818–827. IEEE, 2023.

[50] X. Zhang, Y. Li, X. Zhou, O. Mangoubi, Z. Zhang, V. Filardi, and J. Luo. Dac-ml: domain adaptable continuous meta-learning for urban dynamics prediction. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 906–915. IEEE, 2021.

[51] Y. Zhang, Y. Li, X. Zhou, X. Kong, and J. Luo. Trafficgan: Off-deployment traffic estimation with traffic generative adversarial networks. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 1474–1479, 2019.

[52] Y. Zhang, Y. Li, X. Zhou, X. Kong, and J. Luo. Curb-gan: Conditional urban traffic estimation through spatio-temporal generative adversarial networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 842–852, New York, NY, USA, 2020. Association for Computing Machinery.

[53] Y. Zhang, Y. Li, X. Zhou, X. Kong, and J. Luo. Strans-gan: Spatially-transferable generative adversarial networks for urban traffic estimation. In *2022 IEEE International Conference on Data Mining (ICDM)*, pages 743–752. IEEE, 2022.

[54] Y. Zhang, Y. Li, X. Zhou, Z. Liu, and J. Luo. C3-gan: Complex-condition-controlled urban traffic estimation through generative adversarial networks. In *2021 IEEE International Co10.1016/j.knosys.2023.110591nference on Data Mining (ICDM)*, pages 1505–1510, 2021.

[55] Y. Zhang, Y. Li, X. Zhou, and J. Luo. cst-ml: Continuous spatial-temporal meta-learning for traffic dynamics prediction. In *International Conference on Data Mining*, 2020.

[56] Y. Zhang, Y. Li, X. Zhou, and J. Luo. Mest-gan: Cross-city urban traffic estimation with me ta s patial-t emporal g enerative a dversarial n etworks. In *2022 IEEE International Conference on Data Mining (ICDM)*, pages 733–742, 2022.

[57] C. Zheng, X. Fan, C. Wang, and J. Qi. Gman: A graph multi-attention network for traffic prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):1234–1241, Apr. 2020.

[58] A. Zonoozi, J.-j. Kim, X.-L. Li, and G. Cong. Periodic-crn: A convolutional recurrent model for crowd density prediction with recurring periodic patterns. In *IJCAI*, pages 3732–3738, 2018.