REPST: Language Model Empowered Spatio-Temporal Forecasting via Semantic-Oriented Reprogramming

Hao Wang[1] , **Jindong Han**[3] , **Wei Fan**[4] , **Leilei Sun**[5] , **Hao Liu**[1,2†]
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]The Hong Kong University of Science and Technology
[3]Shandong University
[4]University of Auckland
[5]Beihang University
{figerhaowang, hanjindong01, vegazhang3}@gmail.com,
wei.fan@auckland.ac.nz, leileisun@buaa.edu.cn, liuh@ust.hk

Abstract

Spatio-temporal forecasting is pivotal in numerous real-world applications, including transportation planning, energy management, and climate monitoring. In this work, we aim to harness the reasoning and generalization abilities of Pre-trained Language Models (PLMs) for more effective spatio-temporal forecasting, particularly in data-scarce scenarios. However, recent studies uncover that PLMs, which are primarily trained on textual data, often falter when tasked with modeling the intricate correlations in numerical time series, thereby limiting their effectiveness in comprehending spatio-temporal data. To bridge the gap, we propose REPST, a semantic-oriented PLM reprogramming framework tailored for spatio-temporal forecasting. Specifically, we first propose a semantic-oriented decomposer that adaptively disentangles spatially correlated time series into interpretable sub-components, which facilitates PLM to understand sophisticated spatio-temporal dynamics via a divide-and-conquer strategy. Moreover, we propose a selective discrete reprogramming scheme, which introduces an expanded spatio-temporal vocabulary space to project spatio-temporal series into discrete representations. This scheme minimizes the information loss during reprogramming and enriches the representations derived by PLMs. Extensive experiments on real-world datasets show that the proposed REPST outperforms twelve state-of-the-art baseline methods, particularly in data-scarce scenarios, highlighting the effectiveness and superior generalization capabilities of PLMs for spatio-temporal forecasting. Codes and Appendix can be found at https://github.com/usail-hkust/REPST.

†Corresponding author

1 Introduction

Spatio-temporal forecasting aims to predict future states of real-world complex systems by simultaneously learning spatial and temporal dependencies of historical observations, which plays a pivotal role in diverse real-world applications, such as traffic management [Li *et al.*, 2018; Wu *et al.*, 2019], environmental monitoring [Han *et al.*, 2023], and resource optimization [Geng *et al.*, 2019]. In the past decade, deep learning has demonstrated great predictive power and led to a surge in deep spatio-temporal forecasting models [Jin *et al.*, 2023a]. For example, Recurrent Neural Networks (RNNs) and Graph Neural Networks (GNNs) are frequently combined to capture complex patterns for spatio-temporal forecasting [Li *et al.*, 2018; Han *et al.*, 2020]. Despite fruitful progress made so far, such approaches are typically confined to the one-task-one-model setting, which lacks general-purpose utility and inevitably falls short in handling widespread data-scarcity issue in real-world scenarios, *e.g.*, newly deployed monitoring services.

In recent years, PLMs like GPT-3 [Brown, 2020] and the LLaMA family [Touvron *et al.*, 2023] have achieved groundbreaking success in the Natural Language Processing (NLP) domain. PLMs exhibit exceptional contextual understanding, reasoning, and few-shot generalization capabilities across a wide range of tasks due to their pre-training on extensive text corpora. Although originally designed for textual data, the versatility and power of PLMs have inspired their application to numerically correlated data [Zhou *et al.*, 2024; Jin *et al.*, 2023b; Jin *et al.*, 2024]. For example, [Zhou *et al.*, 2024] pioneers research in this direction and showcases the promise of fine-tuning PLMs as generic time series feature extractors. Besides, model reprogramming [Jin *et al.*, 2023b] has considered the modality differences between time series and natural language, solving time series forecasting tasks by learning an input transformation function that maps time series to a compressed vocabulary.

However, two significant challenges remain in directly applying the aforementioned reprogramming techniques to spatio-temporal forecasting. The foremost issue lies in the underutilization of PLMs' full potential. Recent work [Tan *et al.*, 2024] suggests that existing PLM-based approaches for time series forecasting fail to leverage the generative and reasoning

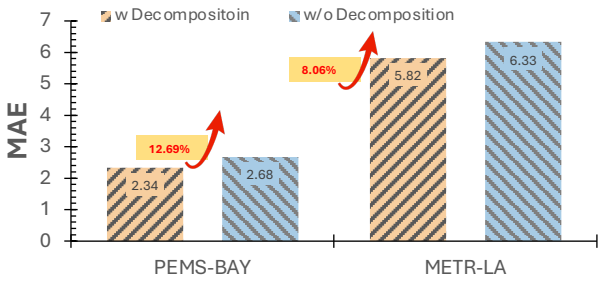

Figure 1: Simple Fourier-based decomposition can improve PLM's understanding of spatio-temporal data. We conduct experiments by applying reprogrammed GPT-2 on widely used PEMS-BAY and METR-LA datasets.

abilities of PLMs. This limitation becomes even more apparent when handling more complex spatially correlated time series data. To this end, a crucial question arises: *how can we better explain this shortcoming and unlock the potential of PLMs for spatio-temporal forecasting?* Another challenge is PLMs' limited capacity to model the intricate correlations present in spatio-temporal data. While PLMs excel at capturing dependencies within one-dimensional sequential data, they fall short in comprehending spatio-temporal data, which often has more complex structures like grids or graphs [Li *et al.*, 2024]. This gap poses a significant obstacle to PLMs' effective use in this domain.

To address these challenges, we argue that the primary limitation of existing approaches lies in their oversimplified treatment of spatio-temporal data, which prevents PLMs from fully understanding the underlying semantics. Instead of merely serving as a one-dimensional encoder, PLMs need a more sophisticated understanding to handle spatio-temporal data effectively. As depicted in Figure 1, our explorative experiments uncover that even applying simple decomposition techniques can significantly facilitate PLMs to better understand spatio-temporal data and lead to improved performance.

Building on this insight, we propose **REPST**, a reprogramming framework specifically designed for spatio-temporal forecasting using PLMs. Specifically, we first propose a **semantic-oriented spatio-temporal decomposer**, which adaptively disentangles spatio-temporal dynamics into components that represent interpretable sub-processes within the system. This is achieved through a Koopman theory-based evolutionary matrix, which results in decomposed components rich in spatio-temporal semantics that PLMs can more easily comprehend. This decomposition-based approach enables PLMs to capture both spatial and temporal dynamics more effectively. Moreover, we introduce a **selective reprogramming strategy** to tackle the complexity of spatio-temporal structures, which differ fundamentally from the one-dimensional sequence-like structure of textual data. Our strategy constructs an expanded spatio-temporal vocabulary by selecting the most relevant spatio-temporal word tokens from the PLM's vocabulary through a differentiable reparameterization process. Unlike previous works that use compressed vocabularies, which can lead to ambiguous semantics, our approach reconstructs the reprogramming space with a rich, semantically distinct spatio-temporal vocabulary. By leveraging pretrained spatio-temporal correlations, this strategy enables PLMs to focus on relationships among tokens in a 3D spatio-temporal space, significantly enhancing their ability to model complex spatio-temporal dynamics. We evaluate REPST on a variety of spatio-temporal forecasting tasks, including energy management, air quality prediction, and traffic forecasting. Extensive experimental results highlight the framework's superior performance compared to state-of-the-art models, particularly in few-shot and zero-shot learning contexts. Our main contributions are summarized as:

- We identify the underlying reason for the underperformance of existing PLM-based approaches for spatio-temporal forecasting, highlighting the need to decompose spatio-temporal dynamics into interpretable components to fully leverage PLMs' potential.
- We propose REPST, a spatio-temporal forecasting framework that enables PLMs to grasp complex spatio-temporal patterns via semantic-oriented decomposition-based reprogramming. The reprogramming module reconstructs an expanded spatio-temporal vocabulary using a selective strategy, allowing PLMs to model spatio-temporal dynamics without altering their pre-trained parameters.
- We show that REPST consistently achieves superior performance across real-world datasets, particularly in data-scarce settings, demonstrating strong generalization capabilities in few-shot and zero-shot learning scenarios.

2 Preliminaries

Spatio-temporal data can be considered as observations of the state of a dynamical system. It is typically represented as a two-dimensional matrix $\mathbf{X} \in \mathbb{R}^{N \times T}$, which captures the states of a set of N nodes $\mathcal{V}$, where each node in $\mathcal{V}$ corresponds to an entity (*e.g.*, grids, regions, and sensors) in space. Specifically, we denote $\mathbf{x}^i_{t-T+1:t} = [\mathbf{x}^i_{t-T+1}, \mathbf{x}^i_{t-T}, ..., \mathbf{x}^i_t]^\top \in \mathbb{R}^{T \times 1}$ as the observations of node i from time step $t-T+1$ to t, where T represents the look-back window length. The goal of spatio-temporal forecasting problem is to predict future states for all nodes $i \in \mathcal{V}$ over the next τ time steps based on a sequence of historical observations. This involves uncovering the complex spatial and temporal patterns inherent in spatio-temporal data to reveal the hidden principles governing the system's dynamics:

$$\hat{\mathbf{Y}}_{t+1:t+\tau} = f_\theta(\mathbf{X}_{t-T+1:t}), \tag{1}$$

where $\mathbf{X}_{t-T+1:t} = [\mathbf{x}^0_{t-T+1:t}, \mathbf{x}^1_{t-T+1:t}, ..., \mathbf{x}^{N-1}_{t-T+1:t}]^\top \in \mathbb{R}^{N \times T}$ denotes the historical observations in previous T time steps, and $f_\theta(\cdot)$ is the spatio-temporal forecasting model parameterized by θ. $\hat{\mathbf{Y}}_{t+1:t+\tau} = \{\hat{\mathbf{y}}^i_{t+1:t+\tau}\}^N_{i=0}$ and $\mathbf{Y}_{t+1:t+\tau} = \{\mathbf{y}^i_{t+1:t+\tau}\}^N_{i=0}$ denote the estimated future states and the ground truth in the next τ time steps, where $\hat{\mathbf{Y}}_{t+1:t+\tau}, \mathbf{Y}_{t+1:t+\tau} \in \mathbb{R}^{N \times \tau}$. For convenience, we omit the lower corner mark and represent $\mathbf{X}_{t-T+1:t}, \mathbf{Y}_{t+1:t+\tau}$ as $\mathbf{X}, \mathbf{Y}$ and $\mathbf{x}^i_{t-T+1:t}, \mathbf{y}^i_{t+1:t+\tau}$ as $\mathbf{x}^i, \mathbf{y}^i$.

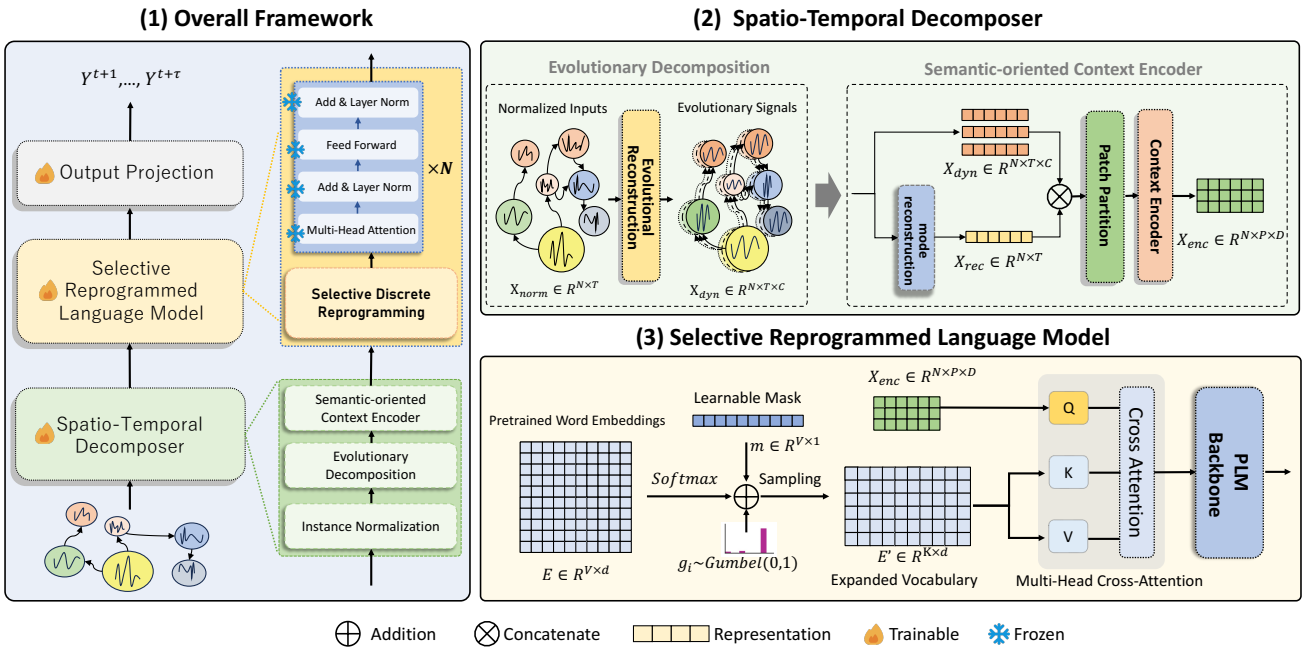

Figure 2: The model framework of REPST. (1) Given a raw input spatio-temporal data, we first perform normalization and then decouple the spatio-temporal data into a set of evolutionary signals. (2) After that, the signals are concatenated and divided into patches and further transformed into embeddings by a semantic-oriented context encoder. (3) Then, the patch embeddings are aligned with natural language by reprogramming with expanded spatio-temporal vocabulary and further processed by the frozen GPT-2 backbone. The output patches of the pre-trained language model are reprocessed by a learnable mapping function to generate the forecasts.

3 Methodology

As illustrated in Figure 2, REPST consists of three components: a semantic-oriented spatio-temporal decomposer, a selective reprogrammed language model, and a learnable mapping function.

3.1 Semantic-Oriented Evolutional Spatio-Temporal Decomposition

Recent studies have revealed that PLMs possess rich spatio-temporal knowledge and reasoning capabilities [Gurnee and Tegmark, 2023; Jin *et al.*, 2024]. However, existing methods failed to fully leverage the capabilities of PLMs, which raises challenges for spatio-temporal data forecasting as well. As aforementioned, reasons for this shortcoming lies in their over simplistic encoding to time series. PLMs requires further process of spatio-temporal data to enhance their comprehensibility to such complex structure. In this section, we address this shortcoming through a carefully designed semantic-oriented spatio-temporal decomposer. Previous works [Liu *et al.*, 2024c; Yi *et al.*, 2024; Shao *et al.*, 2022b] decouple the time series in Fourier space and handle the decoupled signals separately for better use of the hidden information of time series. Simply decomposing time series solely based on frequency intensity is not interpretable and cannot effectively capture the highly coupled spatio-temporal dynamics. Furthermore, this cannot be easily realized by language models as well due to their limited semantics information.

To fully unlock the spatio-temporal knowledge, inspired by dynamic mode decomposition [Schmid, 2010], we propose to capture the underlying dynamic signals in an interpretable manner by leveraging the dynamic system's evolution matrix $\mathcal{A}$. To be specific, considering two state observations $\mathbf{X}_{1:t-1}$ and $\mathbf{X}_{2:t}$, it satisfies $\mathbf{X}_{2:t} = \mathcal{A}\mathbf{X}_{1:t-1}$. This evolution matrix $\mathcal{A}$ is sought in a low-rank setting to capture the modes governing the system's dynamics. By applying a series of mathematical process such as singular value decomposition (SVD) to $\mathbf{X}_{1:t-1}$ and $\mathbf{X}_{2:t}$, we obtain the eigenvectors $\Omega = [\omega_1, \omega_2, ..., \omega_C]$ and corresponding eigenvalues $V = [v_1, v_2, \ldots, v_C]$, which can be leveraged to decompose spatio-temporal dynamic systems into different components. Each ω_i, referred to as a mode of the dynamical system, reflects certain evolution dynamics of the system. We provide detailed description for calculation process in Appendix A.5.

Specifically, we first obtain $\mathbf{X}_{norm}$ by normalizing the input $\mathbf{X}$ for each node to have zero mean and unit standard deviation using reversible instance normalization (RevIN) [Kim *et al.*, 2021]. Then, we disentangle a set of interpretable dynamic components $\mathbf{X}_{dyn} \in \mathbb{R}^{N\times T\times C}$ from intricate spatio-temporal data through reconstructing the system dynamics via modes ω_i from the system's evolution matrix's eigenvectors Ω and the corresponding eigenvalues v_i. By explicitly decoupling the interpretable nature of the spatio-temporal system, our approach is well-suited to capture the various dynamic behaviors of the system, providing PLMs with a series of components enriched with spatio-temporal semantic information that is significantly easier to comprehend compared to the originally densely coupled dynamic signals.

$$\mathbf{X}_{dyn} = \varepsilon_0 e^{\omega_0 t} v_0 \parallel \varepsilon_1 e^{\omega_1 t} v_1 \parallel \cdots \parallel \varepsilon_C e^{\omega_C t} v_C, \quad (2)$$

where $\mathbf{X}_{dyn}$ is a set of spatio-temporal dynamics calculating based on the modes ω_i and eigenvalues v_i. ε_i is based on the input observation (see Appendix A.5). Since the dynamics of the system is disentangled, we can distinguish the noise from the dominant dynamic signals in the original data. If only the most significant information is retained during reconstruction, the reconstruction results can remove noise, thus obtaining a smoother state evolutionary information. Therefore, we further reconstruct the whole system $\mathbf{X}_{rec} \in \mathbb{R}^{N \times T}$ with most dominant modes to enhance prediction:

$$\mathbf{X}_{rec} = \sum_i \varepsilon_i e^{\omega_i t} v_i, i \in \alpha, \tag{3}$$

where α represents a set of indices stands for top-k most dominant modes, constructed based on each mode's contribution to the overall system [Schmid, 2010], which is calculated through the analysis of ω_i and v_i (see Appendix A.5 in supplementary material). Compared to existing Fourier-based methods, the system's evolution matrix Ω is derived from data representing the true dynamics of the system. It can separate modes corresponding to specific intepretable processes, enabling us to capture various aspects of the system's evolution, such as periodic oscillations in traffic flows caused by traffic signals or slow changes in air pollution driven by wind direction [Chen *et al.*, 2012].

Additionally, to enhance the information density of decoupled signals, we employ patching strategy [Nie *et al.*, 2022] to construct patches as the input tokens for PLMs. Given the decoupled signals $\mathbf{X}_{dec} = \mathbf{X}_{rec} \parallel \mathbf{X}_{dyn} \in \mathbb{R}^{N \times T \times (C+1)}$, we divide the observations of each node as a series of non-overlapped patches $\mathbf{X}_{dec}^{P} \in \mathbb{R}^{N \times P \times T_P \times (C+1)}$, where $P = [T/T_P] + 1$ represents the number of the resulting patches, and T_P denotes the patch length. Next, we encode the patched signals as patched embeddings : $\mathbf{X}_{enc} = \mathrm{Conv}(\mathbf{X}_{dec}^{P}, \theta_p) \in \mathbb{R}^{N \times P \times D}$, where N stands for the number of nodes, and D is the embedding dimension. $\mathrm{Conv}(\cdot)$ denotes the patch-wise convolution operator and θ_p represents the learnable parameters of the patch-wise convolution. Unlike previous works [Liu *et al.*, 2024a; Liu *et al.*, 2024b] that simply regard each node as a token, our model treats each patch as one token, allowing to construct fine-grained relationships among both spatial and temporal patterns. By doing so, our model can preserve representations rich in semantic information, allowing PLM's comprehension in both spatial and temporal dynamics more effectively.

3.2 Selective Reprogrammed Language Models

Based on the decoupled signal patches $\mathbf{X}_{enc}$, how to tackle the complexity of spatio-temporal structures raises another question. Compared to directly handling the spatio-temporal embeddings, representations in natural language space are inherently suitable for PLMs. To enrich spatio-temporal semantics and enable more comprehensive modeling of hidden spatio-temporal relationships, as well as unlocking the reasoning capabilities of PLMs, we further reprogram the components into the textual embedding place via an expanded spatio-temporal vocabulary. When handling textual-based components, the rich semantic information can boost the pre-trained knowledge of PLMs, resulting in an adequate modeling of the hidden interactions between disentangled components.

Specifically, we introduce our selective reprogramming strategy, which further constructs an expanded spatio-temporal vocabulary in a differentiable reparameterization process. We begin with $\mathbf{E} \in \mathbb{R}^{V \times d}$, the pretrained vocabulary of the PLMs, where V is the vocabulary size and d is the dimension of the embedding. We introduce a learnable word mask vector $\mathbf{m} \in \mathbb{R}^{V \times 1}$ to adaptively select the most relevant words, where $\mathbf{m}[i] \in \{0, 1\}$. In specific, we first obtain $\mathbf{m}$ through a linear layer followed by a Softmax activation, denoted as $\mathbf{m} = \mathrm{Softmax}(\mathbf{EW})$, where $\mathbf{W}$ is a learnable matrix. Afterward, we sample Top-K word embeddings from $\mathbf{E}$ based on probability $\mathbf{m}[i]$ associated with word i for reprogramming. Since the sampling process is non-differentiable, we employ Gumbel-Softmax trick [Jang *et al.*, 2016] to enable gradient calculation with back-propagation, defined as:

$$\mathbf{m}'[i] = \frac{\exp((\log \mathbf{m}[i] + g_i)/\tau)}{\sum_{j=1}^{V} \exp((\log \mathbf{m}[i] + g_j)/\tau)}, \tag{4}$$

where $\mathbf{m}'$ is a continuous relaxation of binary mask vector $\mathbf{m}$ for word selection, τ is temperature coefficient, g_i and g_j are i.i.d random variables sampled from distribution $\mathrm{Gumbel}(0, 1)$. Concretely, the Gumbel distribution can be derived by first sampling $u \sim \mathrm{Uniform}(0, 1)$ and then computing $g_i = -\log(-\log(u))$. By doing so, we can expand vocabulary space while preserving the semantic meaning of each word.

After obtaining the sampled word embeddings $\mathbf{E}' \in \mathbb{R}^{K \times d}$, we perform modality alignment by using cross-attention. In particular, we define the query matrix $\mathbf{X}_q = \mathbf{X}_{enc}\mathbf{W}_q$, key matrix $\mathbf{X}_k = \mathbf{E}'\mathbf{W}_k$ and value matrix $\mathbf{X}_v = \mathbf{E}'\mathbf{W}_v$, where $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$. After that, we calculate the reprogrammed patch embedding as follows: $\mathbf{Z} = \mathrm{Attn}(\mathbf{X}_q, \mathbf{X}_k, \mathbf{X}_v) = \mathrm{Softmax}(\frac{\mathbf{X}_q \mathbf{X}_k^{\top}}{\sqrt{d}})\mathbf{X}_v$, where $\mathbf{Z} \in \mathbb{R}^{N \times P \times d}$ denotes the aligned textual representations for the input spatio-temporal data.

Based on the aligned representation, we utilize the frozen PLMs as the backbone for further processing. Roughly, PLMs consist of three components: self-attention, Feedforward Neural Networks, and layer normalization layer, which contain most of the learned semantic knowledge from pre-training. The reprogrammed patch embedding $\mathbf{Z}$ is encoded by this frozen language model to further process the semantic information and generates hidden textual representations $\mathbf{Z}_{text}$. A learnable mapping function $\mathrm{Projection}(\cdot)$ is then used to generate the desired target outputs, which map the textual representations into feature prediction: $\hat{\mathbf{Y}} = \mathrm{Projection}(\mathbf{Z}_{text})$.

4 Experiments

4.1 Experimental Settings

Datasets. We conducted experiments on six commonly used real-world datasets [Lai *et al.*, 2018]*, varying in the fields of traffic [Zhang *et al.*, 2017; Li *et al.*, 2018], solar energy, and air quality†. Each dataset comprises tens of thousands of

*https://github.com/LibCity/Bigscity-LibCity

†https://www.biendata.xyz/competition/kdd_2018/data/

time steps and hundreds of nodes, offering a robust foundation for evaluating spatio-temporal forecasting models. The detailed statistics of the datasets are summarized in Appendix in supplementary material.

Baselines. We extensively compare our proposed REPST with the state-of-the-art (sota) forecasting approaches, including (1) the GNN-based methods: [Wu *et al.*, 2019; Shao *et al.*, 2022b; Wu *et al.*, 2020] (2) non-GNN-based sota models: [Shao *et al.*, 2022a; Liu *et al.*, 2023a; Deng *et al.*, 2021] which emphasizes the integration of spatial and temporal identities; (3) sota time series models: [Zhou *et al.*, 2021; Liu *et al.*, 2023b; Nie *et al.*, 2022] (4) PLM-based time series forecasting models: [Zhou *et al.*, 2024]; (5) methods with no trainable parameters: [Cui *et al.*, 2021]. We reproduce all of the baselines based on the original paper or official code.

4.2 REPST Generalization Performance

Few-shot performance. PLMs were trained using large amounts of data that cover various fields, equipping them with cross-domain knowledge. Therefore, PLMs can utilize specific spatio-temporal related textual representations to unlock their capabilities for spatio-temporal reasoning, which can handle the difficulties caused by data sparsity. To verify this, we further conduct experiments on each field to evaluate the predictive performance of our proposed REPST in data-sparse scenarios. Our evaluation results are listed in Table 1 with the best in **bold** and the second underlined. Concretely, all models are trained on 1-day data from the train datasets and tested on the whole test dataset. REPST consistently outperforms other deep models and PLM-based time series forecasters. This illustrates REPST can perform well on a new downstream dataset and is suitable for spatio-temporal forecasting tasks with the problem of data sparsity.

Specifically, our REPST show competitive performance over other baselines in few-shot experiments, demonstrating that PLMs contain a wealth of spatio-temporal related knowledge from pre-training. Moreover, the capabilities of spatio-temporal reasoning can be enhanced by limited data. This shows a reliable performance of REPST when transferred to data-sparse scenarios.

Zero-shot performance. In this part, we focus on evaluating the zero-shot generalization capabilities of REPST within cross-domain and cross-region scenarios following the experiment setting of [Jin *et al.*, 2023b]. Specifically, we test the performance of a model on dataset A after training under a supervised learning framework on another dataset B, where dataset A and dataset B have no overlapped data samples. We use the similar experiment settings to full training experiments and evaluate on various cross-domain and region datasets. The datasets includes NYC Bike, CHI Bike [Jiang *et al.*, 2023], Solar Energy and Air Quality (NYC, CHI, Solar and Air). We compare our performance with recent works in time series or spatio-temporal data with open-sourced model weights [Das *et al.*, 2023; Li *et al.*, 2024; Ekambaram *et al.*, 2024].

Our results in Figure 3 show that REPST consistently secure top positions on all settings. This outstanding zero-shot prediction performance indicates REPST's versatility and adaptability in handling diverse scenarios. It does obtain transferable knowledge for dynamic systems by unlocking the reasoning capabilities of PLMs. Its excellent adaptation to brand new scenarios significantly reduces the time and computational resources typically required by traditional approaches. Although our REPST falls a little short to OpenCity-base in Solar Energy → CHI_Bike, it is because of the large amount of traffic related datasets included by OpenCity's pretrain datasets. Compared to it, our REPST is trained on Solar Energy dataset which has almost no connection with such traffic datasets. This relatively comparable performance demonstrate REPST's excellent generative capability in cross-domain settings. The numerical results are shown in Appendix B.3 in supplementary material.

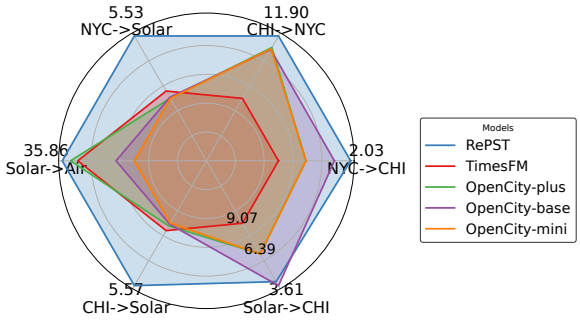

Figure 3: Zero-Shot Performance. We evaluate the zero-shot capability of our REPST in the same setting as few-shot experiments.

Increasing predicted length. In this part, we analyze the model performance across varying prediction horizons $\tau \in \{6, 12, 24, 36, 48\}$, with a fixed input length T = 48. Figure 4 showcases the MAE across two datasets: Air Quality and NYC Bike, for four models. The REPST model demonstrates the most stable performance across both MAE and RMSE metrics, particularly in longer prediction horizons ($\tau = 36, 48$). In contrast, previous state-of-the-art models exhibit notable performance degradation as the prediction horizon increases. The performance of REPST, on the other hand, remains relatively stable and robust, demonstrating its efficacy in leveraging PLM to improve performance over longer-term predictions, which can also be attributed to its ability to capture both spatial and temporal correlations effectively, making it highly suited for few-shot learning tasks in spatio-temporal forecasting.

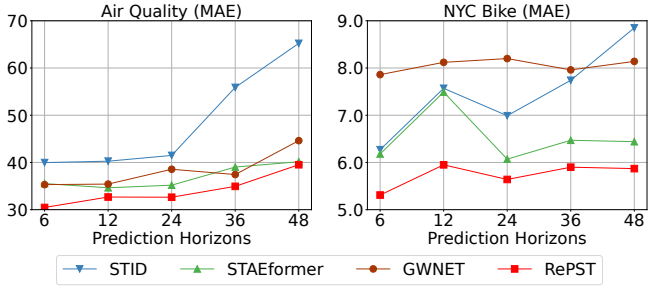

Figure 4: Few-Shot performance with multiple prediction horizons $\tau \in \{6, 12, 24, 36, 48\}$ and fixed input length $T = 48$.

Dataset	METR-LA		PEMS-BAY		Solar Energy		Air Quality		Beijing Taxi				NYC Bike			
									Inflow		Outflow		Inflow		Outflow	
Metric	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Informer	8.19	14.35	5.30	10.43	8.95	11.92	38.02	56.45	29.20	53.52	28.76	52.53	6.99	16.44	6.33	15.62
iTransformer	7.72	15.85	5.20	10.94	4.74	8.27	35.59	52.95	31.56	58.11	32.22	59.93	8.23	16.21	7.46	15.68
PatchTST	7.20	15.56	4.52	8.85	4.65	7.82	35.76	53.80	32.66	61.17	32.58	60.95	7.03	15.32	6.88	14.84
MTGNN	9.62	17.60	5.67	8.91	4.73	8.68	36.51	53.14	28.98	48.72	28.80	46.87	6.51	14.85	6.56	14.90
GWNet	7.04	12.58	5.84	9.42	9.10	11.87	36.26	54.88	29.24	51.68	29.47	50.52	12.55	21.97	12.68	22.27
STNorm	7.93	13.67	5.15	8.92	5.36	9.59	36.38	57.66	28.92	50.59	28.86	49.39	11.69	20.17	12.53	21.84
D2STGNN	6.41	11.57	5.31	9.39	8.80	11.26	40.77	55.07	36.73	58.70	36.06	66.01	10.64	18.96	10.33	18.43
STID	7.26	12.70	6.83	12.88	4.89	9.41	43.21	61.07	32.73	51.77	32.91	51.94	8.94	16.34	8.88	15.77
STAEFormer	6.35	11.38	5.37	9.35	4.66	12.57	37.68	53.39	28.88	49.86	28.06	48.13	12.50	20.77	11.84	20.88
FPT	6.80	11.36	4.55	9.71	10.59	13.92	36.62	51.33	41.66	74.87	43.28	77.84	12.97	20.06	12.72	20.11
REPST	**5.63**	**9.67**	**3.61**	**7.15**	**3.65**	**6.74**	**33.57**	**47.30**	**26.85**	**45.88**	**26.30**	**43.76**	**5.29**	**12.11**	**5.66**	**12.85**

Table 1: Few-Shot performance comparison on six real-world datasets in terms of MAE and RMSE. We utilize data in one day (less than 1%)for training and the same data as full training settings for validation and test. The input history time steps T and prediction steps τ are both set to 24. We use the average prediction errors over all prediction steps.

4.3 Full Training Performance of REPST

Table 2 reports the overall performance of our proposed REPST as well as baselines in 6 real-world datasets with the best in **bold** and the second underlined. As can be seen, REPST consistently achieves either the best or second-best results in terms of MAE and RMSE.

Notably, REPST surpasses the state-of-the-art PLM-based time series forecaster FPT [Zhou *et al.*, 2024] by a large margin in spatio-temporal forecasting tasks, which can demonstrate that simply leveraging the PLMs cannot handle problems with complex spatial dependencies. Additionally, the performance of our REPST reaches either the best or second-best results in METR-LA and PEMS-BAY datasets. Previous state-of-the-art models, STAEformer and STID, learn global shared embeddings both in spatial structure and temporal patterns tailored for certain datasets, which is harmful to their generalization abilities but benefits their capabilities to handle single datasets. Our spatio-temporal reprogramming block leverages a wide range of vocabulary and sample words that can adequately capture the spatio-temporal patterns, which do make an impact on unlocking the capabilities of PLMs to capture fine-grained spatio-temporal dynamics.

4.4 Ablation Study

To figure out the effectiveness of each component in REPST, we further conduct detailed ablation studies on Air Quality and Solar Energy datasets with three variants as follows: **r/p PLM**: it replaces the PLM backbone with transformer layers, following the setting of [Tan *et al.*, 2024]; **r/p Decomposition**: it replaces the semantic-oriented spatio-temporal decomposer with a transformer encoder; **w/o expanded vocabulary**: it removes our selective spatio-temporal vocabulary and utilizes the dense mapping function to enhance reprogramming.

Figure 5 shows the comparative performance of the variants above on Air Quality, Solar Energy and Beijing Taxi. Based on the results, we can make the conclusions as follows: (1) Our REPST actually leverage the pretrain knowledge and generative capabilities of PLMs. When we replace the PLM backbone with transformer layers, the performance of all the datasets decline obviously, indicating that the pretrain knowledge makes an effect to handle spatio-temporal dynamics. (2) The semantic-oriented spatio-temporal decomposer which adaptively disentangles input spatio-temporal data into interpretable components can actually enable PLMs to better understand spatio-temporal dynamics. When constructing spatio-temporal dependencies by a transformer encoder layer, it is still unclear for PLMs to comprehend. (3) The impressive performance in w/o expanded vocabulary demonstrates that the selectively reconstructed vocabulary achieves accurate reprogramming which enables PLMs to focus on relationships among tokens in 3D spatio-temporal space.

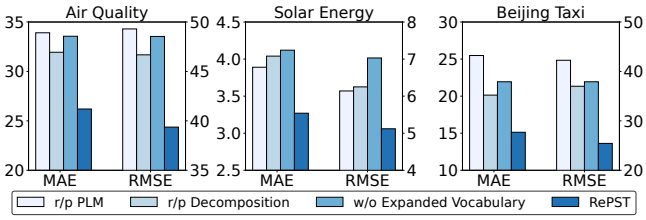

Figure 5: Ablation study.

5 Related Works

5.1 Spatio-Temporal Forecasting

Spatio-temporal forecasting has been playing a critical role in various smart city services, such as traffic flow prediction [Liu *et al.*, 2023a; Shao *et al.*, 2022a; Fang *et al.*, 2023], air quality monitoring [Han *et al.*, 2023; Han *et al.*, 2021], and energy management [Geng *et al.*, 2019]. Unlike traditional time series forecasting, the forecasting challenges associated with spatio-temporal data are often characterized by the unique properties of strong correlation and heterogeneity along the spatial dimension, which are inherently more complex.

Early studies usually capture spatial dependencies through a predefined graph structure [Li *et al.*, 2018; Han *et al.*, 2020;

Dataset	METR-LA		PEMS-BAY		Solar Energy		Air Quality		Beijing Taxi				NYC Bike			
									Inflow		Outflow		Inflow		Outflow	
Metric	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
HI	9.88	16.98	5.51	10.50	9.42	12.53	53.18	67.42	105.55	142.98	105.63	143.08	11.98	19.23	12.18	19.50
Informer	4.68	8.92	2.54	5.30	3.92	5.91	29.38	42.58	16.41	29.03	16.01	26.90	3.49	8.36	3.92	9.52
iTransformer	4.16	9.06	2.51	5.90	3.33	5.41	28.37	44.33	21.72	36.80	22.15	38.63	3.15	7.55	3.28	7.82
PatchTST	4.15	9.07	2.06	4.85	3.49	5.89	28.05	44.81	23.64	43.63	22.71	41.52	3.58	8.83	3.66	8.99
MTGNN	3.76	7.45	1.94	4.40	3.60	5.61	27.07	40.17	15.92	26.15	15.79	26.08	3.31	7.91	3.38	8.24
GWNet	3.93	8.19	2.28	5.06	3.55	5.39	31.57	44.82	15.69	26.82	15.76	26.84	3.13	7.58	3.33	7.64
STNorm	3.98	8.44	2.20	5.02	4.17	5.99	30.73	44.82	15.37	27.50	15.45	27.52	3.14	7.46	3.24	7.63
D2STGNN	3.94	7.68	2.11	4.83	4.36	5.85	27.77	41.87	24.33	45.65	26.86	45.57	3.10	7.43	3.25	7.75
STID	3.68	7.46	1.93	**4.31**	3.70	5.57	26.94	41.01	15.60	27.96	15.81	28.28	3.36	7.91	3.38	8.24
STAEFormer	**3.60**	7.44	1.97	4.33	3.44	5.21	28.12	41.83	15.47	26.45	16.08	26.83	3.03	7.39	3.27	7.56
FPT	6.03	10.85	2.56	5.01	6.02	8.31	32.79	47.55	32.41	55.28	32.77	55.77	7.21	12.76	7.75	13.85
REPST	3.63	**7.43**	**1.92**	4.33	**3.27**	**5.12**	**26.20**	**39.37**	**15.13**	**25.44**	**15.75**	**25.24**	**3.01**	**7.33**	**3.16**	**7.43**

Table 2: Performance comparison of full training on six real-world datasets in terms of MAE and RMSE. The input history time steps T and prediction steps τ are both set to 24. We use the average prediction errors over all prediction steps.

Shao *et al.*, 2022b; Wu *et al.*, 2020], which decribes the explicit relationships among different spatial locations. In recent years, there is a growing trend toward the utilization of adaptive spatio-temporal graph neural networks, which can automatically capture dynamic spatial graph structures from data [Wu *et al.*, 2019]. Besides, attention mechanism is also widely employed in existing models, as seen in examples like [Guo *et al.*, 2021; Liu *et al.*, 2023a; Fang *et al.*, 2024]. In contrast, [Shao *et al.*, 2022a], a model based on Multi-Layer Perceptrons, achieves state-of-the-art results by utilizing multiple embedding techniques to memorize stable spatial and temporal patterns.

More recently, inspired by the huge success of PLMs in NLP field, there is increasing interest in building pre-trained models for spatio-temporal forecasting tasks. Several studies [Liu *et al.*, 2024a; Liu *et al.*, 2024b; Yan *et al.*, 2023] explore the application of PLMs for handling spatio-temporal data. Furthermore, the strong power of Transformer offers an opportunity to build spatio-temporal foundation models, such as OpenCity [Li *et al.*, 2024] and UniST [Yuan *et al.*, 2024]. Trained on numerous spatio-temporal data, these models demonstrate strong capabilities across diverse forecasting scenarios. However, due to the problems of data-sparsity in multiple spatio-temporal scenarios, it is difficult for these models to gather large amount of data to perform pretraining comprehensively. In addition, the PLM-based spatio-temporal forecasting approaches do not fully leverage PLM's potential. To address these gaps, this paper introduces a new reprogramming framework to leverage PLM's generative and reasoning capabilities for spatio-temporal forecasting, particularly in data-sparse scenarios.

5.2 Pretrained Language Models for Time Series

In recent years, PLMs have demonstrated remarkable performance across various time series analysis tasks [Zhou *et al.*, 2024; Gruver *et al.*, 2024; Sun *et al.*, 2023]. A significant body of research has focused on leveraging PLMs to address these challenges [Gruver *et al.*, 2023]. However, a persistent issue in these efforts is the modality gap between time series data and natural language. To address this challenge, [Jin *et al.*, 2023b] develops a time series reprogramming approach, which can effectively bridges the modality gap between time series and text data. The objective of reprogramming is to learn a trainable transformation function that can be applied to the patched time series data, enabling it to be mapped into the textual embedding space of the PLM.

Nevertheless, [Tan *et al.*, 2024] conducted numerous experiments showing that existing PLM-based approaches do not fully unlock the reasoning or generative capabilities of PLMs. The reason these approaches achieve high performance lies in the similar sequential formulation shared by time series and natural language [Liu *et al.*, 2024d]. Although PLMs can handle one-dimensional sequential data like text and time series, they fall short in capturing dependencies among complex spatio-temporal structure, leading to suboptimal performance for spatio-temporal forecasting tasks. In this work, we propose REPST, which enables PLMs to comprehend complex spatio-temporal dynamics via a semantic-oriented decomposition-based reprogramming strategy.

6 Conclusion

In this paper, we highlight the underlying reason for the poor performance of previous PLM-based approaches in spatio-temporal forecasting, emphasizing the need for the interpretability to fully leverage PLM's potential. To address this problem, we developed REPST, a tailored spatio-temporal forecasting framework that enables PLMs to comprehend the complex spatio-temporal patterns via a semantic-oriented decomposition-based reprogramming strategy. As a result, PLM's potential is full unlocked to handle spatio-temporal forecasting tasks. Extensive experiments demonstrate that our proposed framework, REPST, achieves state-of-the-art performance on real-world datasets and exhibits exceptional capabilities in few-shot and zero-shot scenarios.

Acknowledgments

This work was supported by the National Key R&D Program of China (Grant No.2023YFF0725004), National Nat-

ural Science Foundation of China (Grant No.92370204), the Guangzhou Basic and Applied Basic Research Program under Grant No. 2024A04J3279, Education Bureau of Guangzhou Municipality.

References

[Brown, 2020] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[Chen *et al.*, 2012] Kevin K Chen, Jonathan H Tu, and Clarence W Rowley. Variants of dynamic mode decomposition: boundary condition, koopman, and fourier analyses. *Journal of nonlinear science*, 22:887–915, 2012.

[Cui *et al.*, 2021] Yue Cui, Jiandong Xie, and Kai Zheng. Historical inertia: A neglected but powerful baseline for long sequence time-series forecasting. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 2965–2969, 2021.

[Das *et al.*, 2023] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*, 2023.

[Deng *et al.*, 2021] Jinliang Deng, Xiusi Chen, Renhe Jiang, Xuan Song, and Ivor W Tsang. St-norm: Spatial and temporal normalization for multi-variate time series forecasting. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 269–278, 2021.

[Ekambaram *et al.*, 2024] Vijay Ekambaram, Arindam Jati, Nam H Nguyen, Pankaj Dayama, Chandra Reddy, Wesley M Gifford, and Jayant Kalagnanam. Ttms: Fast multi-level tiny time mixers for improved zero-shot and few-shot forecasting of multivariate time series. *arXiv preprint arXiv:2401.03955*, 2024.

[Fang *et al.*, 2023] Yuchen Fang, Yanjun Qin, Haiyong Luo, Fang Zhao, Bingbing Xu, Liang Zeng, and Chenxing Wang. When spatio-temporal meet wavelets: Disentangled traffic forecasting via efficient spectral graph attention networks. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 517–529. IEEE, 2023.

[Fang *et al.*, 2024] Yuchen Fang, Yuxuan Liang, Bo Hui, Zezhi Shao, Liwei Deng, Xu Liu, Xinke Jiang, and Kai Zheng. Efficient large-scale traffic forecasting with transformers: A spatial data management perspective. *arXiv preprint arXiv:2412.09972*, 2024.

[Geng *et al.*, 2019] Xu Geng, Yaguang Li, Leye Wang, Lingyu Zhang, Qiang Yang, Jieping Ye, and Yan Liu. Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3656–3663, 2019.

[Gruver *et al.*, 2023] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. Large language models are zero-shot time series forecasters. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 19622–19635. Curran Associates, Inc., 2023.

[Gruver *et al.*, 2024] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36, 2024.

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

[Gurnee and Tegmark, 2023] Wes Gurnee and Max Tegmark. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*, 2023.

[Han *et al.*, 2020] Haoyu Han, Mengdi Zhang, Min Hou, Fuzheng Zhang, Zhongyuan Wang, Enhong Chen, Hongwei Wang, Jianhui Ma, and Qi Liu. Stgcn: a spatial-temporal aware graph learning method for poi recommendation. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1052–1057. IEEE, 2020.

[Han *et al.*, 2021] Jindong Han, Hao Liu, Hengshu Zhu, Hui Xiong, and Dejing Dou. Joint air quality and weather prediction based on multi-adversarial spatiotemporal networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4081–4089, 2021.

[Han *et al.*, 2023] Jindong Han, Weijia Zhang, Hao Liu, and Hui Xiong. Machine learning for urban air quality analytics: A survey. *arXiv preprint arXiv:2310.09620*, 2023.

[Jang *et al.*, 2016] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[Jiang *et al.*, 2023] Jiawei Jiang, Chengkai Han, Wenjun Jiang, Wayne Xin Zhao, and Jingyuan Wang. Libcity: A unified library towards efficient and comprehensive urban spatial-temporal prediction. *arXiv preprint arXiv:2304.14343*, 2023.

[Jin *et al.*, 2023a] Guangyin Jin, Yuxuan Liang, Yuchen Fang, Zezhi Shao, Jincai Huang, Junbo Zhang, and Yu Zheng. Spatio-temporal graph neural networks for predictive learning in urban computing: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[Jin *et al.*, 2023b] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*, 2023.

[Jin *et al.*, 2024] Ming Jin, Yifan Zhang, Wei Chen, Kexin Zhang, Yuxuan Liang, Bin Yang, Jindong Wang, Shirui Pan, and Qingsong Wen. Position paper: What can large language models tell us about time series analysis. *arXiv preprint arXiv:2402.02713*, 2024.

[Kim *et al.*, 2021] Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*, 2021.

[Lai *et al.*, 2018] Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 95–104, 2018.

[Li *et al.*, 2018] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations*, 2018.

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

[Liu *et al.*, 2023a] Hangchen Liu, Zheng Dong, Renhe Jiang, Jiewen Deng, Jinliang Deng, Quanjun Chen, and Xuan Song. Spatio-temporal adaptive embedding makes vanilla transformer sota for traffic forecasting. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pages 4125–4129, 2023.

[Liu *et al.*, 2023b] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*, 2023.

[Liu *et al.*, 2024a] Chenxi Liu, Sun Yang, Qianxiong Xu, Zhishuai Li, Cheng Long, Ziyue Li, and Rui Zhao. Spatial-temporal large language model for traffic prediction. *arXiv preprint arXiv:2401.10134*, 2024.

[Liu *et al.*, 2024b] Lei Liu, Shuo Yu, Runze Wang, Zhenxun Ma, and Yanming Shen. How can large language models understand spatial-temporal data? *arXiv preprint arXiv:2401.14192*, 2024.

[Liu *et al.*, 2024c] Yong Liu, Chenyu Li, Jianmin Wang, and Mingsheng Long. Koopa: Learning non-stationary time series dynamics with koopman predictors. *Advances in Neural Information Processing Systems*, 36, 2024.

[Liu *et al.*, 2024d] Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Autotimes: Autoregressive time series forecasters via large language models. *arXiv preprint arXiv:2402.02370*, 2024.

[Nie *et al.*, 2022] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

[Schmid, 2010] Peter J Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of fluid mechanics*, 656:5–28, 2010.

[Shao *et al.*, 2022a] Zezhi Shao, Zhao Zhang, Fei Wang, Wei Wei, and Yongjun Xu. Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4454–4458, 2022.

[Shao *et al.*, 2022b] Zezhi Shao, Zhao Zhang, Wei Wei, Fei Wang, Yongjun Xu, Xin Cao, and Christian S Jensen. Decoupled dynamic spatial-temporal graph neural network for traffic forecasting. *arXiv preprint arXiv:2206.09112*, 2022.

[Sun *et al.*, 2023] Chenxi Sun, Yaliang Li, Hongyan Li, and Shenda Hong. Test: Text prototype aligned embedding to activate llm's ability for time series. *arXiv preprint arXiv:2308.08241*, 2023.

[Tan *et al.*, 2024] Mingtian Tan, Mike A Merrill, Vinayak Gupta, Tim Althoff, and Thomas Hartvigsen. Are language models actually useful for time series forecasting? *arXiv preprint arXiv:2406.16964*, 2024.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. corr, abs/2302.13971, 2023. doi: 10.48550. *arXiv preprint arXiv.2302.13971*, 2023.

[Wu *et al.*, 2019] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. *arXiv preprint arXiv:1906.00121*, 2019.

[Wu *et al.*, 2020] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 753–763, 2020.

[Yan *et al.*, 2023] Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. When urban region profiling meets large language models. *arXiv preprint arXiv:2310.18340*, 2023.

[Yi *et al.*, 2024] Kun Yi, Qi Zhang, Wei Fan, Hui He, Liang Hu, Pengyang Wang, Ning An, Longbing Cao, and Zhendong Niu. Fouriergnn: Rethinking multivariate time series forecasting from a pure graph perspective. *Advances in Neural Information Processing Systems*, 36, 2024.

[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.

[Zhang *et al.*, 2017] Junbo Zhang, Yu Zheng, and Dekang Qi. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the AAAI conference on artificial intelligence*, 2017.

[Zhou *et al.*, 2021] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11106–11115, 2021.

[Zhou *et al.*, 2024] Tian Zhou, Peisong Niu, Liang Sun, Rong Jin, et al. One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems*, 36, 2024.