Spatio-temporal Prototype-based Hierarchical Learning for OD Demand Prediction

Shilu Yuan[1] , **Xiaoyu Li**[1] , **Wenqian Mu**[1] , **Ji Zhong**[2] , **Meng Chen**[1] , **Haoliang Sun**[1] and **Yongshun Gong**[1]*

[1]School of Software, Shandong University

[2]Shandong Yunhai Guochuang Cloud Computing Equipment Industry Innovation Co., Ltd

{shiluyuan, xyuli, mwq_bella}@mail.sdu.edu.cn, 13671364664@163.com, mchen@sdu.edu.cn, haolsun.cn@gmail.com, yongshun2512@hotmail.com

Abstract

Origin-Destination (OD) demand prediction is a pivotal yet challenging task in intelligent transportation systems, aiming to accurately forecast cross-station ridership flows within urban networks. While previous studies have focused on modeling node-to-node relationships, most of them neglect the fact that nodes (stations/regions) exhibit similar spatio-temporal (ST) patterns, which are termed as spatio-temporal prototypes. Capturing these prototypes is crucial for understanding the unified ST dependencies across the traffic network. To bridge this gap, we propose STPro, an ST prototype-based hierarchical model with a dual-branch structure that extracts ST features from the micro and macro perspectives. At the micro level, our model learns unified ST features of individual nodes, while at the macro level, it employs dynamic clustering to identify city-wide ST prototypes, thereby uncovering latent patterns of urban mobility. Besides, we leverage different roles of nodes as origins and destinations by constructing dual O and D branches and learn the mutual information to model their intricate interactions and correlations. Extensive experiments on two public datasets demonstrate that our STPro outperforms recent state-of-the-art baselines, achieving remarkable predictive improvements in OD demand prediction.

1 Introduction

With rapid urbanization and the ever-growing demand for transportation, urban mobility systems face significant challenges, including traffic congestion [Gong *et al.*, 2020b; Ji *et al.*, 2023], resource inefficiency [Gross *et al.*, 2006], and prolonged passenger wait times [Gong *et al.*, 2020a]. Understanding passenger movement patterns is a critical pathway to optimizing urban transportation systems. Origin-Destination (OD) demand prediction, which focuses on forecasting the number of passengers traveling from specific origins to destinations, has emerged as a critical research area in urban computing [Rong *et al.*, 2024]. OD demand prediction can identify the underlying spatio-temporal patterns of passenger movements and forecast their mobility intentions, thereby offering essential insights into the dynamics of urban mobility.

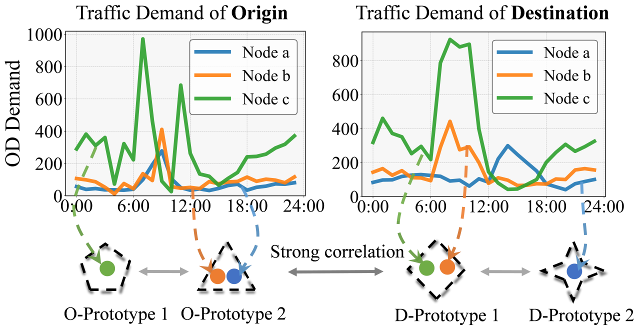


Figure 1: An illustration of OD demand and spatio-temporal prototypes.

Due to its significance in practical applications, OD demand prediction has recently attracted extensive attention in both academic and industrial communities [Wang *et al.*, 2019; Han *et al.*, 2022; Ye *et al.*, 2024; Yu *et al.*, 2025]. Some studies transform passenger flow data into grid-based representations and utilize CNNs to capture spatial dependencies [Wang *et al.*, 2019; Jiang, 2023]. To model more complex spatial relationships, other approaches represent passenger flow data as graphs and employ GNNs to extract intricate spatio-temporal patterns [Liu *et al.*, 2022; Ye *et al.*, 2024; Liu *et al.*, 2024]. However, despite these advancements, existing methods still face several limitations. One significant issue is the failure to adequately capture the distinct characteristics of a station when it functions as an origin (O) compared to when it serves as a destination (D). As illustrated in Figure 1, the demand patterns of Node c as an origin can differ substantially from its traffic demand distribution patterns as a destination. A clear real-world example of this phenomenon is the contrasting traffic patterns observed in workplace districts during morning and evening peak hours. Previous studies often model origin and destination features jointly, which overlooks their inherent semantic differences.

The second challenge lies in capturing the implicit spatio-temporal patterns from a global perspective. As illustrated in

[*] Corresponding author.

Figure 1, different nodes may exhibit similar spatio-temporal evolution patterns regardless of their geographical proximity. We refer to these shared patterns, which are applicable across multiple nodes, as **spatio-temporal prototypes**. The prototypes often represent the traffic demand evolution features of a specific category of nodes, providing a higher-level understanding of urban mobility dynamics. Effectively capturing the spatio-temporal prototypes and modeling their intricate correlations is a crucial yet unexplored step in OD prediction.

To address these challenges, we propose STPro, a spatio-temporal prototype based dual-branch hierarchical framework for OD demand prediction. Our framework works at two levels to comprehensively capture spatio-temporal dynamics. At the macro level, we focus on learning spatio-temporal prototypes to capture the shared mobility patterns across nodes. These prototypes are constructed using dynamic clustering, which adaptively groups nodes with similar spatio-temporal evolution trends. At the micro level, we extract the fine-grained spatio-temporal features for individual nodes with a hypervariate graph that fully connects all nodes. To address the inherent semantic differences between origin and destination roles, we design a dual-branch architecture that separately models the characteristics of nodes as origins (O-branch) and destinations (D-branch). At each branch, spatio-temproal features learned from the micro and macro levels are fused to generate the unified embeddings. To enable effective interaction between the two branches, we employ a cross-attention mechanism to integrate mutual information between the O-branch and D-branch, which is then used to make final OD demand predictions. Extensive experiments conducted on two real-world datasets demonstrate our method, as it consistently outperforms baseline methods. In summary, our work makes the following key contributions:

- We propose STPro, a hierarchical framework for OD demand prediction. We utilize dynamic clustering to construct spatio-temporal prototypes at the macro level and leverage the fine-grained node features at the micro level. This hierarchical integration enriches spatio-temporal semantics and enhances prediction accuracy.
- We design a dual-branch architecture to separately model the semantic characteristics of nodes as origins (O-branch) and destinations (D-branch). To model the dynamic interactions between origin and destination roles, a cross-attention mechanism is employed to extract and integrate mutual information between the two branches.
- Extensive experiments are conducted on two widely used real-world datasets to evaluate the STPro. The experimental results demonstrate that our approach can build effective spatio-temporal prototypes and achieve state-of-the-art performance in both single-step and multi-step OD demand prediction.

2 Related Works

2.1 Origin-Destination Demand Prediction

Origin-destination demand prediction is a critical task in urban mobility analysis. It aims to forecast traffic demand between two locations and has attracted significant research attention in recent years. With the rapid development of deep learning techniques, neural network-based methods have emerged as the dominant approach for addressing this task. Shi et al. [2020] used LSTM units to extract temporal characteristics for each pair of OD and separately learned the spatial dependencies of the origins and destinations. Ke et al. [2021] represented OD pairs with multiple OD graphs and developed a spatio-temporal encoder-decoder residual framework to model the spatial dependencies between different OD pairs and the temporal dependencies of OD pairs themselves. Liu et al. [2022] proposed a heterogeneous information aggregation mechanism, which fully utilizes incomplete OD matrices from historical data to jointly learn the evolution patterns of OD and DO rides. Liu et al. [2024] introduced ODMixer, a method for learning traffic evolution by analyzing all OD pairs. They proposed a fine-grained spatio-temporal MLP architecture specifically designed for metro OD prediction.

However, previous studies do not explore the hierarchical spatio-temporal information inherent in OD demand, which is crucial for capturing both global trends and local variations. Furthermore, most of them predict the OD matrix without considering the distinct semantic attributes and intrinsic correlations between origins and destinations

2.2 Prototype Learning

Prototypes have been extensively utilized in various machine learning paradigms, including transfer learning [Quattoni *et al.*, 2008], multi-task learning [Kang *et al.*, 2011], and few-shot learning [Snell *et al.*, 2017; Liu *et al.*, 2019; Li *et al.*, 2021a]. Traditionally, a prototype is defined as the mean feature vector of samples within the same class [Wieting *et al.*, 2015; Babenko and Lempitsky, 2015]. This definition leverages the central tendency of feature distributions to represent class-specific information, which has proven effective in tasks requiring generalization across similar instances. For instance, in distributed machine learning systems, prototypes have been employed to capture task-agnostic information, enabling the fusion of multi-task models for new tasks [Hoang *et al.*, 2020]. Furthermore, the generalization capability of prototypes has been harnessed in federated learning, where they assist local training by aggregating semantic knowledge from distributed data sources [Michieli and Ozay, 2021; Tan *et al.*, 2021; Li *et al.*, 2021b; Mu *et al.*, 2023].

In this paper, the spatio-temporal prototypes are defined as dynamic clustering centers that adaptively capture shared mobility patterns across nodes. Unlike static mean prototypes, our approach leverages dynamic clustering to iteratively update prototypes based on evolving spatio-temporal characteristics. This dynamic nature allows the prototypes to reflect temporal variations and spatial dependencies, which are critical for accurate OD demand prediction.

3 Problem Statement

In this section, we provide the definition of origin-destination demand prediction.

The primary objective of this work is to predict future OD demands based on past observed demand data. Given

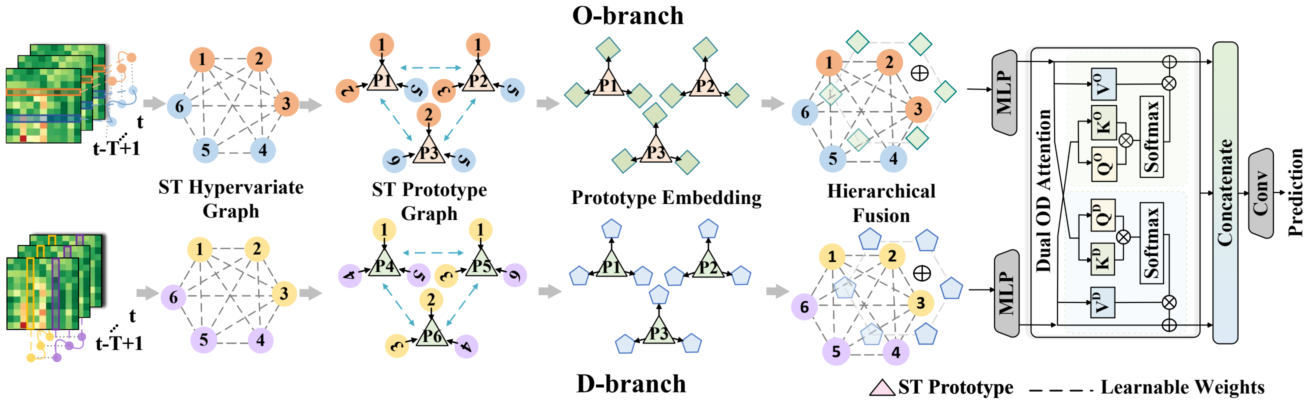


Figure 2: Overview of our proposed STPro. STPro establishes a dual-branch structure based on the distinct attributes of nodes as origins and destinations. Each branch first generates spatio-temporal features of nodes based on a fully connected spatio-temporal graph. Next, OD prototypes are generated using dynamic clustering. Dual-branch interaction is achieved through OD dual attention. Finally, the outputs of dual-branch mutual information are fused to produce the final prediction.

the passenger flow between stations (or regions) at a specific time, we formulate this information into an OD matrix $\boldsymbol{X}_t \in R^{N\times N}$, where N denotes the number of stations. We treat these stations as nodes in the urban network in the following context. Here, we represent the node features from two perspectives and generate the input for the O-branch and D-branch. The i-th row of X_t represents the flow from the node i to all other nodes. From the origin perspective, we consider the origin feature of the node $\boldsymbol{v}_t^{Oi} \in R^N$ and the set of all nodes is $\boldsymbol{V}_t^O = \{\boldsymbol{v}_t^{O1}, ..., \boldsymbol{v}_t^{Oi}, ..., \boldsymbol{v}_t^{ON}\} \in R^{N\times N}$. Similarly, the j-th column of the $\boldsymbol{X}_t$ represents the flow from all nodes to node j. The destination feature of the node is $\boldsymbol{v}_t^{Dj}$, and the set of all nodes is $\boldsymbol{V}_t^D = \{\boldsymbol{v}_t^{D1}, ..., \boldsymbol{v}_t^{Di}, ..., \boldsymbol{v}_t^{DN}\} \in R^{N\times N}$. In this work, we utilize the passenger flow from previous T time steps to forecast the future OD demand matrix. The origin feature of T time steps $\boldsymbol{V}^O$ will be the input for the O-branch and the destination feature $\boldsymbol{V}^D$ will be the input for the O-branch. Finally, the prediction task is formulated as:

$$[\boldsymbol{X}_{t-T+1}, ..., \boldsymbol{X}_t] \xrightarrow{f} [\hat{\boldsymbol{Y}}_{t+1}, ..., \hat{\boldsymbol{Y}}_{t+m}]. \tag{1}$$

4 Proposed Method

In this paper, we propose a dual-branch hierarchical model capable of learning the multi-level information of OD to effectively model the spatio-temporal distribution of OD. As shown in Figure. 2, our model consists of an O-branch hierarchical module, a D-branch hierarchical module, and a cross-attention mechanism for O-D interaction modeling. Specifically, at the micro level of each branch, an OD hypervariate graph is constructed to learn spatio-temporal unified node information. Meanwhile, dynamic clustering is applied to generate OD prototypes at the macro level, exploring typical spatio-temporal patterns in the city. Subsequently, the dual-branch information is interacted to enhance the modeling. We develop a hierarchical OD framework to jointly predict multi-level complete OD demand for the future.

4.1 Hierarchical OD Modeling

We propose a hierarchical model for OD demand forecasting that integrates macro and micro levels to capture both global and local spatio-temporal patterns. The micro level captures spatio-temporal local features via fully connected nodes, while the macro level extracts global OD flow patterns by dynamic clustering. Hierarchical interaction combines two levels, providing multi-level representations in the OD system. However, the origin and destination OD flow of a node is affected by different factors that we should consider separately. Therefore, we model the hierarchical structure in both branches (O-branch and D-branch).

OD Hypervariate Graph

Urban OD flow is essentially a spatio-temporal traffic system with a non-Euclidean topology. Recently, GCN has been proven to be effective for spatio-temporal non-Euclidean data embedding [Yi *et al.*, 2024]. However, the unified spatio-temporal graph structure is usually unknown in OD flow scenarios [Zhang *et al.*, 2022; Han *et al.*, 2022; Liu *et al.*, 2022]. Inspired by these works, we construct spatio-temporal OD hypervariate graphs with two branches and utilize graph convolutional units to learn the spatio-temporal representations for OD prediction.

To capture long-term spatio-temporal information, we choose NT nodes from T time steps for the OD hypervariate graph. In the O-branch, we define the set of nodes for T time steps as $\boldsymbol{U}_t^O = \{\boldsymbol{V}_{t-T+1}^O, ..., \boldsymbol{V}_t^O\} \in R^{NT\times N}$, and construct an OD hypervariate graph $G_t^O = \{\boldsymbol{U}_t^O, \boldsymbol{A}_t^O\}$. Since the prior graph structure is unknown, and nodes are spatially and temporally correlated with each other because of the time lag effect [Wei, 1990], we assume all nodes in the graph G_t^O are fully connected. Therefore, $\boldsymbol{A}_t^O \in R^{NT\times NT}$ is the adjacency matrix initialized to make G_t^O as a spatio-temporal fully-connected graph. In the same way, in the D-branch, we construct the graph $G_t^D = \{\boldsymbol{U}_t^D, \boldsymbol{A}_t^D\}$.

Based on the fully connected graphs, we utilize the standard graph convolution [Kipf and Welling, 2016] to learn spatio-temporal representations $\boldsymbol{H}_t^{O/D} \in R^{NT \times d}$. In detail, for each branch, we update its representation by aggregating the state embedding vectors of all its fully connected neighbors including itself. In formulation, the representation of $\boldsymbol{H}_t^{O/D}$ follows:

$$\boldsymbol{H}_t^{O/D} = \boldsymbol{A}_t^{O/D}(\boldsymbol{U}_t^{O/D}\boldsymbol{W}_t^{O/D}), \tag{2}$$

where $\boldsymbol{W}_t^{O/D} \in R^{N \times d}$ is a weight matrix; the superscript O and D denotes the spatio-temporal representation generation processes of the O-branch and D-branch, respectively.

OD Prototypes Embedding

The traffic system naturally exhibits a spatio-temporal hierarchical structure [Guo *et al.*, 2021]. OD system not only includes the basic micro layer of spatio-temporal nodes but also has the macro layer of prototypes. While focusing on the local spatio-temporal behavior of individual OD pairs, we also aim to capture the global spatio-temporal OD flow patterns.

To explore the dynamic hierarchical relationships in the OD network, we use a learnable matrix $\boldsymbol{\Theta}^{O/D}$ to dynamically cluster spatio-temporal nodes into different spatio-temporal prototypes, where $\boldsymbol{\Theta}^{O/D} \in R^{d \times I}$ and I is the number of the prototypes. In detail, the incidence probability matrix $\boldsymbol{\Lambda}_t^{O/D} \in R^{NT \times I}$ between spatio-temporal nodes and prototypes is formalized as:

$$\boldsymbol{\Lambda}_t^{O/D} = Softmax(\boldsymbol{H}_t^{O/D}\boldsymbol{\Theta}^{O/D}). \tag{3}$$

In OD systems, there are complex dependencies between prototypes. Therefore, we introduce a learnable matrix $\boldsymbol{P}^{O/D} \in R^{I \times I}$ to capture the connection relationships of the prototypes. Here, $\boldsymbol{P}^{O/D}$ is regarded as the spatio-temporal prototype graph used for graph convolution. Then, we utilize the graph convolution [Zhao *et al.*, 2023] to learn spatio-temporal prototype embeddings, which have spatio-temporal global information from the hierarchical OD demand network. Specifically, we generate each prototype's embedding $\boldsymbol{E}_t^{O/D} \in R^{I \times d}$ by aggregating the embeddings of all spatio-temporal nodes and the spatio-temporal prototype graph:

$$\boldsymbol{E}_t^{O/D} = \phi\left(\boldsymbol{P}^{O/D}\left(\boldsymbol{\Lambda}_t^{O/D}\right)^T\boldsymbol{H}_t^{O/D}\right) + \left(\boldsymbol{\Lambda}_t^{O/D}\right)^T\boldsymbol{H}_t^{O/D}, \tag{4}$$

where ϕ is a non-linear activation function.

OD Hierarchical Fusion

To realize the interaction between the macro and micro layers, we propose an OD hierarchical fusion to integrate local spatio-temporal node features and global spatio-temporal prototype features. First, we define the transformation matrix $\boldsymbol{Tran}_t^{O/D} \in R^{NT \times I}$ based on the incidence probability matrices in the previous section:

$$\boldsymbol{Tran}_t^{O/D} = \boldsymbol{\Lambda}^{O/D}. \tag{5}$$

We transform the obtained local node features and the global prototype features into the same dimension using the transformation matrix:

$$\boldsymbol{E}_{tran}^{O/D} = \boldsymbol{Tran}_t^{O/D}\boldsymbol{E}_t^{O/D}, \tag{6}$$

where $\boldsymbol{E}_{tran}^{O/D} \in R^{NT \times d}$ and each variable in $\boldsymbol{E}_{tran}^{O/D}$ contains information about all nodes of the same prototype. To integrate features from different levels, we design the hierarchical fusion function as follows:

$$\boldsymbol{F}_t^{O/D} = MLP\left(ReadOut\left(\boldsymbol{E}_{tran}^{O/D}, \boldsymbol{H}_t^{O/D}\right)\right), \tag{7}$$

where the readout is a summation operation.

4.2 Dual OD Attention

After obtaining clear multi-level representations from the two branches, we explicitly model the correlation between the two branches to enhance the overall representation of OD demand. Inspired by Liu et al. [2022], we introduce a dual OD attention mechanism to model the O and D distributions jointly by propagating their mutual information in a dual manner. Following this interaction, the multi-level O and D features become informative for OD demand flow prediction.

Our dual OD attention is implemented with cross-branch cross-attention, where the bidirectional transformer propagates information from the O-branch to the D-branch, as well as from the D-branch to the O-branch. Specifically, $\boldsymbol{F}_t^O$ and $\boldsymbol{F}_t^D$ are first respectively fed into three linear layers for query, key, and value embedding:

$$\begin{aligned}
\boldsymbol{Q}_t^O &= Conv(\boldsymbol{F}_t^O, \boldsymbol{W}_q^O) + \boldsymbol{b}_q^O, \boldsymbol{Q}_t^D = Conv(\boldsymbol{F}_t^D, \boldsymbol{W}_q^D) + \boldsymbol{b}_q^D, \\
\boldsymbol{K}_t^O &= Conv(\boldsymbol{F}_t^O, \boldsymbol{W}_k^O) + \boldsymbol{b}_k^O, \boldsymbol{K}_t^D = Conv(\boldsymbol{F}_t^D, \boldsymbol{W}_k^D) + \boldsymbol{b}_k^D, \\
\boldsymbol{V}_t^O &= Conv(\boldsymbol{F}_t^O, \boldsymbol{V}_v^O) + \boldsymbol{b}_v^O, \boldsymbol{V}_t^D = Conv(\boldsymbol{F}_t^D, \boldsymbol{W}_v^D) + \boldsymbol{b}_v^D,
\end{aligned} \tag{8}$$

where all convolution kernel sizes are $1 * 1$ with individual parameters, and the biases for the query, key, and value have dimensions of $NT \times d$. Same as $\boldsymbol{F}_t^O$ and $\boldsymbol{F}_t^D$, these query, key, value features also have a dimension of $NT \times d$.

Based on attention mechanisms, the cross-branch information propagation is performed with the following formulations:

$$\begin{aligned}
\boldsymbol{Z}_t^{O2D} &= \boldsymbol{F}_t^O + softmax(\boldsymbol{Q}_t^O(\boldsymbol{K}_t^D)^\top)\boldsymbol{V}_t^D, \\
\boldsymbol{Z}_t^{D2O} &= \boldsymbol{F}_t^D + softmax(\boldsymbol{Q}_t^D(\boldsymbol{K}_t^O)^\top)\boldsymbol{V}_t^O,
\end{aligned} \tag{9}$$

where $softmax(\boldsymbol{Q}_t^O(\boldsymbol{K}_t^D)^\top)$ and $softmax(\boldsymbol{Q}_t^D(\boldsymbol{K}_t^O)^\top)$ are two propagation coefficients that dynamically determine the amount of information propagated between the features of the O-branch and the D-branch. Through this process, the representations of the O-branch and the D-branch not only reinforce each other but also better capture the intrinsic interaction between the origin and destination attributes of nodes in the OD system.

4.3 OD Demand Prediction

Finally, the extracted features are fused and used for OD demand prediction. The model is capable of predicting both single-step and multi-step OD demands. For multi-step OD demand prediction, the model directly outputs predictions for multiple future time steps to avoid error accumulation. The model uses Mask OD Loss to measure prediction errors and continuously improve its performance.

To predict the OD demand at timestamp $t+1$, the model utilizes the interaction features from previous timestamps, including O2D features and D2O features. First, we combine the two features and multiply them by the transformation matrix $\boldsymbol{W}_z$ resulting in a matrix with dimensions $NT \times N$. Then, we reshape the matrix into $T \times N \times N$. Finally, we sum along the temporal dimension to obtain a combined matrix with dimension of $1 \times N \times N$, aggregating the information across T time steps. The combination feature can be expressed as follows:

$$\boldsymbol{Com}_t = \sum_{T} (\boldsymbol{Z}_t^{O2D} \oplus \boldsymbol{Z}_t^{D2O}) \boldsymbol{W}_z. \tag{10}$$

After combining spatio-temporal features, a convolution layer is employed to make the prediction. The OD demand prediction result at the time $t+m$ can be expressed as follows:

$$\hat{\boldsymbol{Y}}_{t+m} = Conv(\boldsymbol{Com}_t, \boldsymbol{Ker}_{1*1}(m)), \tag{11}$$

where $\hat{\boldsymbol{Y}}_{t+m}$ is the prediction results for future m time steps and $\boldsymbol{Ker}$ denotes the convolution kernel, with the number of kernels corresponding to the predicted the next m time steps, and the size of each kernel being $1 * 1$.

5 Experiments

5.1 Experimental Setup

Datasets

The two OD datasets used in our experiments are NYC-TOD2018 and NYC-TOD2019 [Yu *et al.*, 2025]. New York City taxi datasets were provided by the New York City Taxi and Limousine Commission (TLC). We selected the yellow taxi data for this study. According to the division of TLC, Manhattan is divided into 69 nodes. The NYC-TOD2018 dataset collected passenger demand data for 69 areas in Manhattan from January 1,2018 to December 31, 2018, and the NYC-TOD2019 dataset collected passenger demand data for 69 areas in Manhattan from January 1, 2019, to December 31, 2019. The time interval for both datasets is 1 hour.

Evaluation Setting and Metrics

All models take 3 hours (3 time steps) of historical data as input and make predictions for the future 1/2/3 hours (1/2/3 time steps) of OD demand. For both datasets, we use 60% of the data for training, 20% for validation, and the remaining 20% for testing the model's performance. We choose widely used Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) as the evaluation metrics.

Baselines

We conduct a comprehensive comparison of the proposed STPro with a wide range of baselines, spanning from traditional statistical methods to state-of-the-art graph-based approaches. The baselines include: (1) the statistical-based method **HA**; (2) RNN-based machine learning methods, such as **GRU** [Cho *et al.*, 2014] and **LSTM** [Hochreiter and Schmidhuber, 1997]; (3) recent graph-based traffic prediction methods, including **STGCN** [Yu *et al.*, 2017], **GraphWaveNet** [Wu *et al.*, 2019], and **DCRNN** [Li *et al.*, 2017]; and (4) state-of-the-art graph-based OD prediction methods, such as **GEML** [Wang *et al.*, 2019], **CMOD** [Han *et al.*, 2022], **HIAM** [Liu *et al.*, 2022], **C-AHGCSP** [Ye *et al.*, 2024], and **ODMixer** [Liu *et al.*, 2024]. Notably, STGCN, GraphWaveNet, and DCRNN are representative models originally designed for traffic forecasting. To adapt them for OD demand prediction, we re-implemented these models based on their official code.

5.2 Comparison with Advanced Methods

The performance summary of single-step and multi-step OD demand predictions for all comparison methods is presented in Table 1. From the results, we observe that traditional methods exhibit relatively higher MAE and RMSE values across different time intervals, indicating their limitations in capturing complex spatio-temporal patterns. In contrast, graph-based traffic prediction methods, such as STGCN, GraphWaveNet, and DCRNN, demonstrate improved performance over traditional baselines and RNN-based models by leveraging graph convolution to model spatial and temporal dependencies. Among these, GraphWaveNet achieves notable results, particularly in multi-step predictions, due to its use of dilated causal convolution, which effectively captures long-term temporal dependencies. However, this mechanism appears less effective in capturing fine-grained details critical for single-step predictions, which may explain its relatively weaker performance in such scenarios. While graph-based methods have made significant advancements in modeling traffic dynamics, they primarily focus on establishing relationships between individual traffic points, often overlooking the unified spatio-temporal prototypes.

Our proposed method introduces a novel hierarchical spatio-temporal framework that effectively captures mobility patterns at multiple scales, addressing the limitations of existing approaches. By integrating micro-level fine-grained details and macro-level global trends, our model achieves a balanced representation of OD demand dynamics, enabling accurate predictions for both single-step and multi-step forecasting tasks. The experimental results demonstrate the robustness and effectiveness of our approach. On the NYC-TOD2018 dataset, our model achieves a 5.58% reduction in RMSE for 2-hour predictions, showcasing its ability to capture short-term mobility patterns. Similarly, on the NYC-TOD2019 dataset, our model improves RMSE by 1.55% for 2-hour predictions, further validating its generalizability across different datasets. These consistent improvements highlight the superiority of STPro in modeling the hierarchical spatio-temporal dynamics of OD systems.

5.3 Ablation Studies

In this section, we conduct comprehensive ablation studies to evaluate the contribution of each component in the proposed STPro. Table 2 summarizes the average performance of all model variants for the 3-hour prediction task. The key components under investigation include the dynamic clustering module (Prototypes), the O-branch for modeling nodes as origins, and the D-branch for modeling nodes as destinations.

Effectiveness of Prototypes

From Table 2, it can be observed that the model with prototypes significantly outperforms other variants in OD pre-

Model	NYC-TOD2018						NYC-TOD2019					
	1 hour		2 hour		3 hour		1 hour		2 hour		3 hour	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
HA	1.973	4.944	1.974	4.945	1.974	4.946	1.718	4.419	1.718	4.419	1.718	4.420
LSTM	1.569	3.568	1.596	3.772	1.620	3.996	1.350	3.147	1.369	3.607	1.403	3.635
GRU	1.586	3.613	1.608	3.839	1.627	3.953	1.365	3.160	1.376	3.613	1.418	3.646
STGCN	1.515	3.483	1.586	3.702	1.621	3.923	1.317	3.110	1.370	3.220	1.411	3.777
GraphWaveNet	1.752	4.668	1.668	4.342	1.625	4.179	1.476	3.995	1.406	3.767	1.373	3.646
DCRNN	1.533	3.925	1.570	4.068	1.613	4.199	1.324	3.533	1.353	3.562	1.390	3.669
GEML	1.787	4.071	1.817	4.230	1.855	4.386	1.544	3.592	1.548	3.609	1.583	3.739
CMOD	1.533	3.435	1.542	3.738	1.565	3.785	1.311	3.024	1.318	3.043	1.338	3.082
HIAM	1.424	3.621	1.434	3.632	1.462	3.702	1.215	3.160	1.226	3.167	1.251	3.234
C-AHGCSP	1.425	3.550	1.457	3.699	1.523	3.742	1.217	3.231	1.237	3.260	1.236	3.448
ODMixer	1.454	3.605	1.460	3.726	1.527	3.749	1.243	3.269	1.301	3.431	1.294	3.458
STPro (Ours)	**1.420**	**3.410**	**1.429**	**3.429**	**1.454**	**3.505**	**1.209**	**2.977**	**1.211**	**2.992**	**1.221**	**3.044**

Table 1: Performance of OD prediction on the NYC-TOD2018 and NYC-TOD2019.

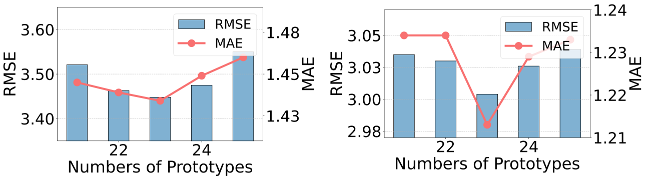
Prototypes	O-branch	D-branch	NYC-TOD2018		NYC-TOD2019	
			MAE	RMSE	MAE	RMSE
	√		1.477	3.695	1.257	3.324
		√	1.475	3.689	1.259	3.332
	√	√	1.457	3.615	1.239	3.194
√	√		1.466	3.661	1.248	3.234
√		√	1.464	3.660	1.245	3.265
√	√	√	**1.434**	**3.448**	**1.213**	**3.004**

Table 2: Ablation studies on NYC-TOD2018 and NYC-TOD2019.

diction, indicating that the dynamic clustering module effectively enhances OD prediction performance. Specifically, the dynamic clustering module generates spatio-temporal prototypes by aggregating features of similar samples into dynamic centers. These prototypes enable the model to better learn the intrinsic patterns of OD demand. For instance, on the NYC-TOD2018 dataset, the STPro model with prototypes achieves a 1.57% reduction in MAE and a 4.61% reduction in RMSE compared to the variant without prototypes. Similar performance improvements were also observed on the NYC-TOD2019 dataset. These experimental results further validate the effectiveness of the dynamic clustering module in capturing spatio-temporal correlations.

Effectiveness of Dual OD Branch

The results in Table 2 also highlight the importance of the dual OD branch structure, which explicitly models the interactions between origins and destinations. The model incorporating OD branch interaction significantly outperforms the variant without interaction, demonstrating its ability to capture the complex relationships between origins and destinations. These correlations are often overlooked or inadequately modeled in traditional approaches, and the OD branch interaction module addresses this limitation.



(a) NYC-TOD2018

(b) NYC-TOD2019

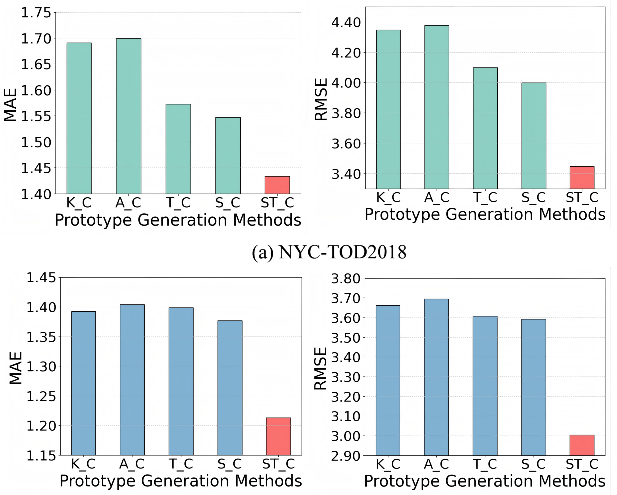
Figure 3: Studies on the number of prototypes.

6 Model Analysis

6.1 Impact of the Number of Prototypes

In this section, we analyze the impact of the number of prototypes I on the prediction performance. As described in the methodology, we cluster $N \times T$ nodes to form I prototypes, which serve as high-level representations of shared spatio-temporal patterns. To determine the value of I, we first conducted a wide-range grid search and identified that models with I around twenty consistently achieved superior performance. Based on this observation, we performed a detailed study of I around twenty to evaluate the average performance for both single-step and multi-step predictions.

As shown in Figure 3, both MAE and RMSE reach the minimum values for single-step and multi-step predictions when I equals 23. When the number of prototypes I deviates from 23—either by being too small or too large—the prediction errors increase. We infer that the number of prototypes should align with the real-world number of macro spatio-temporal patterns inherent in the dataset. The prototypes can reflect the true diversity of real-world traffic patterns captured in the data. An appropriate value of I enables the model to better capture the underlying spatio-temporal correlations in the OD demand data. While the choice of I has an implicit impact on prediction performance, our hierarchical approach consistently achieves competitive results compared to state-of-the-art methods.



(a) NYC-TOD2018

(b) NYC-TOD2019

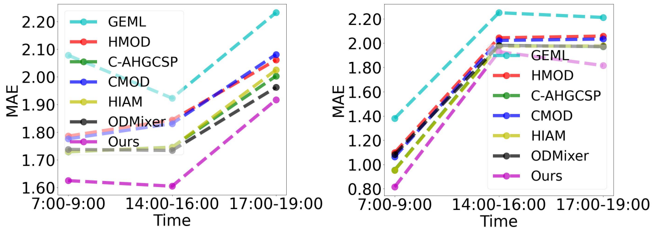
Figure 4: Studies on the prototype generation methods.

6.2 Exploration of Prototype Generation Methods

In this section, we conduct a detailed exploration of various spatio-temporal prototype generation methods to validate the effectiveness of our proposed approach. Specifically, we compare our method with the following approaches: K-Means clustering (K_C), Agglomerative clustering (A_C), Temporal clustering (T_C), and Spatial clustering (S_C). The experimental results, presented in Figure 4, clearly demonstrate that our proposed clustering method ST_C significantly outperforms other approaches. This superior performance stems from its ability to jointly model spatial and temporal information, effectively capturing the complex dependencies inherent in spatio-temporal data. In addition to their suboptimal performance in prediction accuracy, K_C and A_C suffer from high computational complexity and low efficiency. On the other hand, S_C and T_C are limited by their focus on either spatial or temporal features. Our proposed method effectively captures the complex spatial and temporal relationships in OD demand data and ensures a more accurate prediction.

6.3 Experiments on the Stability of the STPro

To further validate the stability of our model, we analyze the prediction performance of STPro across different time periods. This analysis focuses on the model's ability to handle varying levels of OD flow complexity during weekdays and weekends. We provide the comparison with the six best-performing baselines in Figure 5. On weekdays, all methods often exhibit higher prediction errors during the morning and evening periods, primarily due to the increased complexity of traffic patterns during peak hours. Our method not only achieves optimal prediction results across all time periods but also demonstrates exceptional performance during peak hours, outperforming some methods even during off-peak hours. Even though the traffic patterns on weekends differ from those on weekdays, our method consistently achieves the best prediction performance across both scenarios. By maintaining stable and accurate predictions under diverse conditions, our method ensures reliable performance across a wide range of real-world applications, from weekday commutes to weekend leisure travel.



(a) Weekdays

(b) Weekends

Figure 5: Studies on model stability.

Model	OD demand (HZMetro)		OD flow (HZMOD)	
	MAE	RMSE	MAE	RMSE
GEML	1.181	2.385	1.202	2.355
HIAM	1.035	2.011	0.978	1.939
C-AHGCSP	1.072	2.145	1.099	2.143
ODMixer	1.074	2.194	1.101	2.147
STpro	**0.891**	**1.986**	**0.883**	**1.928**

Table 3: Experiments of OD Flow and OD Demand predictions.

6.4 Model Applicability Analysis

To validate the applicability of STPro, we conducted additional experiments on datasets from a new city and extended the model to OD flow prediction. Experimental results presented in Table 3 demonstrate that STPro maintains outstanding prediction performance on the HZMetro dataset [Yu *et al.*, 2025], collected from Hangzhou, China. For OD flow prediction, it focuses on real-time flow estimation based on incomplete trip observations (e.g., subway station entries without exits) [Yu *et al.*, 2025]. The results on HZMOD [Liu *et al.*, 2022] highlight STPro's robustness against sparsity issues in OD flow prediction.

7 Conclusion

In this paper, we present STPro, a dual-branch hierarchical model designed for OD demand prediction. STPro captures multi-level spatio-temporal information, combining fine-grained details at the micro level with city-wide patterns at the macro level. STPro effectively models urban mobility dynamics by leveraging dynamic clustering to identify spatio-temporal prototypes and employing a dual-branch structure to model origin-destination interactions. Experiments on two real-world datasets show that STPro achieves state-of-the-art performance in both single-step and multi-step OD demand predictions. This work advances OD demand prediction and provides a robust framework for spatio-temporal modeling in intelligent transportation systems. Future research will explore extending the framework with additional data sources and applying it to other spatio-temporal tasks.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62476154, 62202270), and Major Basic Research Project of Shandong Provincial Natural Science Foundation (ZR2024ZD03) and the Shandong Excellent Young Scientists Fund (Oversea) (2022HWYQ-044), and the Taishan Scholar Project of Shandong Province (tsqn202306066), and Shandong University Qilu Young Scholars Program.

Contribution Statement

Shilu Yuan and Xiaoyu Li have made equal contributions to this work.

References

[Babenko and Lempitsky, 2015] Artem Babenko and Victor Lempitsky. Aggregating local deep features for image retrieval. In *Proceedings of the IEEE international conference on computer vision*, pages 1269–1277, 2015.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[Gong *et al.*, 2020a] Yongshun Gong, Zhibin Li, Jian Zhang, Wei Liu, and Jinfeng Yi. Potential passenger flow prediction: A novel study for urban transportation development. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 4020–4027, 2020.

[Gong *et al.*, 2020b] Yongshun Gong, Zhibin Li, Jian Zhang, Wei Liu, and Yu Zheng. Online spatio-temporal crowd flow distribution prediction for complex metro system. *IEEE Transactions on knowledge and data engineering*, 34(2):865–880, 2020.

[Gross *et al.*, 2006] Philip Gross, Albert Boulanger, Marta Arias, David L Waltz, Philip M Long, Charles Lawson, Roger Anderson, Matthew Koenig, Mark Mastrocinque, William Fairechio, et al. Predicting electricity distribution feeder failures using machine learning susceptibility analysis. In *AAAI*, pages 1705–1711, 2006.

[Guo *et al.*, 2021] Kan Guo, Yongli Hu, Yanfeng Sun, Sean Qian, Junbin Gao, and Baocai Yin. Hierarchical graph convolution network for traffic forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 151–159, 2021.

[Han *et al.*, 2022] Liangzhe Han, Xiaojian Ma, Leilei Sun, Bowen Du, Yanjie Fu, Weifeng Lv, and Hui Xiong. Continuous-time and multi-level graph representation learning for origin-destination demand prediction. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 516–524, 2022.

[Hoang *et al.*, 2020] Nghia Hoang, Thanh Lam, Bryan Kian Hsiang Low, and Patrick Jaillet. Learning task-agnostic embedding of multiple black-box experts for multi-task model fusion. In *International Conference on Machine Learning*, pages 4282–4292. PMLR, 2020.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

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

[Jiang, 2023] Wenxiao Jiang. Grid-based origin-destination matrix prediction: a deep learning method with vector graph transformation similarity loss function. In *Handbook of Mobility Data Mining*, pages 135–151. Elsevier, 2023.

[Kang *et al.*, 2011] Zhuoliang Kang, Kristen Grauman, and Fei Sha. Learning with whom to share in multi-task feature learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 521–528, 2011.

[Ke *et al.*, 2021] Jintao Ke, Xiaoran Qin, Hai Yang, Zhengfei Zheng, Zheng Zhu, and Jieping Ye. Predicting origin-destination ride-sourcing demand with a spatio-temporal encoder-decoder residual multi-graph convolutional network. *Transportation Research Part C: Emerging Technologies*, 122:102858, 2021.

[Kipf and Welling, 2016] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *CoRR*, abs/1609.02907, 2016.

[Li *et al.*, 2017] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*, 2017.

[Li *et al.*, 2021a] Bingcong Li, Bo Han, Zhuowei Wang, Jing Jiang, and Guodong Long. Confusable learning for large-class few-shot classification. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part II*, pages 707–723. Springer, 2021.

[Li *et al.*, 2021b] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10713–10722, 2021.

[Liu *et al.*, 2019] Lu Liu, Tianyi Zhou, Guodong Long, Jing Jiang, and Chengqi Zhang. Learning to propagate for graph meta-learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[Liu *et al.*, 2022] Lingbo Liu, Yuying Zhu, Guanbin Li, Ziyi Wu, Lei Bai, and Liang Lin. Online metro origin-destination prediction via heterogeneous information aggregation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3574–3589, 2022.

[Liu *et al.*, 2024] Yang Liu, Binglin Chen, Yongsen Zheng, Guanbin Li, and Liang Lin. Fine-grained spatial-temporal mlp architecture for metro origin-destination prediction. *arXiv preprint arXiv:2404.15734*, 2024.

[Michieli and Ozay, 2021] Umberto Michieli and Mete Ozay. Prototype guided federated learning of visual feature representations. *arXiv preprint arXiv:2105.08982*, 2021.

[Mu *et al.*, 2023] Xutong Mu, Yulong Shen, Ke Cheng, Xueli Geng, Jiaxuan Fu, Tao Zhang, and Zhiwei Zhang. Fedproc: Prototypical contrastive federated learning on non-iid data. *Future Generation Computer Systems*, 143:93–104, 2023.

[Quattoni *et al.*, 2008] Ariadna Quattoni, Michael Collins, and Trevor Darrell. Transfer learning for image classification with sparse prototype representations. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[Rong *et al.*, 2024] Can Rong, Jingtao Ding, and Yong Li. An interdisciplinary survey on origin-destination flows modeling: Theory and techniques. *ACM Computing Surveys*, 57(1):1–49, 2024.

[Shi *et al.*, 2020] Hongzhi Shi, Quanming Yao, Qi Guo, Yaguang Li, Lingyu Zhang, Jieping Ye, Yong Li, and Yan Liu. Predicting origin-destination flow via multi-perspective graph convolutional network. In *2020 IEEE 36th International conference on data engineering (ICDE)*, pages 1818–1821. IEEE, 2020.

[Snell *et al.*, 2017] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[Tan *et al.*, 2021] Y Tan, G Long, L Liu, T Zhou, and J Jiang. Fedproto: federated prototype learning over heterogeneous devices. arxiv. *Learning*, 2021.

[Wang *et al.*, 2019] Yuandong Wang, Hongzhi Yin, Hongxu Chen, Tianyu Wo, Jie Xu, and Kai Zheng. Origin-destination matrix prediction via graph convolution: a new perspective of passenger demand modeling. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1227–1235, 2019.

[Wei, 1990] WS Wei. Time series analysis. univariate and multivariate methods. addison. *Wesley publishingcomp*, 1990.

[Wieting *et al.*, 2015] John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. Towards universal paraphrastic sentence embeddings. *arXiv preprint arXiv:1511.08198*, 2015.

[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.

[Ye *et al.*, 2024] Jiexia Ye, Juanjuan Zhao, Furong Zheng, and Chengzhong Xu. A heterogeneous graph convolution based method for short-term od flow completion and prediction in a metro system. *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[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.

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

[Yu *et al.*, 2025] Piao Yu, Xu Zhang, Yongshun Gong, Jian Zhang, Haoliang Sun, Junjie Zhang, Xinxin Zhang, and Yilong Yin. Enhancing origin–destination flow prediction via bi-directional spatio-temporal inference and interconnected feature evolution. *Expert Systems with Applications*, 264:125679, 2025.

[Zhang *et al.*, 2022] Ruixing Zhang, Liangzhe Han, Boyi Liu, Jiayuan Zeng, and Leilei Sun. Dynamic graph learning based on hierarchical memory for origin-destination demand prediction. *arXiv preprint arXiv:2205.14593*, 2022.

[Zhao *et al.*, 2023] Yusheng Zhao, Xiao Luo, Wei Ju, Chong Chen, Xian-Sheng Hua, and Ming Zhang. Dynamic hypergraph structure learning for traffic flow forecasting. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 2303–2316. IEEE, 2023.