Structural Entropy Based Spatio-temporal Sequence Forecasting

Daliang Liu[1,2], Kun Yue[1,2], Wenjie Liu[1,2], Xiang Chen[1,2], and Liang Duan[1,2] ✉

[1] School of Information Science and Engineering, Yunnan University,
Kunming, China
[2] Yunnan Key Laboratory of Intelligent Systems and Computing, Yunnan University,
Kunming, China
duanl@ynu.edu.cn

Abstract. Spatio-temporal sequence forecasting (STSF) aims to predict the future sequence of spatio-temporal inputs based on previous observations. Most of the existing methods focus on training graph neural networks and Transformers to extract spatial and temporal features of the input data. However, real-world spatio-temporal sequences often contain complex dependencies and lots of irrelevant information for forecasting, which might degrade the model's performance. To tackle these issues, we propose a novel Spatio-Temporal Sequence Conditional Information Bottleneck (STSCIB) approach for STSF based on structural information theory. First, we establish an information bottleneck principle to extract the minimal and sufficient information for STSF, which could maximize the relevant information while minimizing the negative effects by eliminating irrelevant information based on the community dependencies among data. Second, we provide an efficient model to approximately implement the proposed STSCIB upon an encoding tree with the minimal structural entropy. Finally, we design an effective contrastive loss to train the model. Experiments on five datasets show the superiority of our method compared with other state-of-the-art methods.

Keywords: Spatio-temporal Sequence Forecasting · Structural Entropy · Hierarchical Communities · Conditional Information Bottleneck.

1 Introduction

Predicting future sequences based on historical observations is a fundamental task in spatio-temporal sequence forecasting (STSF), and widely used in various domains [12]. In practice, spatio-temporal sequence data often contain complex spatial-temporal dependencies and lots of irrelevant information for forecasting [18, 15], which can significantly impair the model's performance. Therefore, it is essential for STSF methods to capture complex dependencies and minimize the impact of irrelevant information.

Recent works are exploring Transformer-based model for STSF, since the graph structure that indicating spatial dependencies among sequences is often biased, incomplete, or missing in many cases [12]. However, Transformer-based models struggle with short-term sequence data due to limited semantic information from isolated data points [12, 9], resulting in unreliable modeling of the dependency correlation, and are difficult to capture higher-order spatial semantic information. Moreover, these methods lack a theoretical principle of measuring the optimal representation for STSF tasks, which prevents them to learn effective representations that capture both spatial semantics and temporal patterns [15]. Thus, it is essential to develop a novel principle framework for measuring the optimal representation for STSF and extracting the most useful information from time series and graph structures.

According to Information Bottleneck (IB) [18], the optimal representation for STSF should contain useful information as much as possible, while reducing irrelevant information that does not contribute to the task. Inspired by this, we adopt IB to train the Transformer-base model for learning the optimal representations by minimizing the regularization term to discard irrelevant information. However, spatio-temporal sequences exhibit coupled spatial structures and temporal patterns, and the regularization in IB may force the encoder to rely on historical observations from a single sequence or specific time points, making the optimization of the IB difficult [2]. To tackle these challenges, we leverage conditional reconstruction to replace the regularization term in IB, which not only satisfies the minimal and sufficient assumption but also concludes that the uncertainty of spatial structure should be minimized in the STSF task.

As is known that structural information theory [7], centered on structural entropy and encoding tree offers an effective measure of the information embedded in an arbitrary graph and structural diversity. The multilevel semantics of a graph can be abstracted and characterized through an encoding tree, which represents a multi-grained division of graphs into hierarchical communities, providing a richer representation of higher-order spatial semantics. Based on structural information theory, we build an encoding tree by minimizing the uncertainty of input graph structure and make innovative use of it for the instantiation of the conditional reconstruction term, which not only provides high-quality and useful community structure information for our approach, but also addresses the issue of IB optimization.

In this paper, we propose an innovative Spatio-Temporal Sequence Conditional Information Bottleneck (STSCIB) framework to address key challenges in STSF tasks, including the lack of theoretical guidance, difficulty in optimizing spatio-temporal sequence representations, and the inability to capture spatial heterogeneity, which collectively leads to low prediction accuracy. Our approach decouples conditional mutual information into temporal and spatial components for efficient optimization. On the temporal dimension, we pre-train a Transformer-based temporal encoder using a random masking strategy to mitigate the issue of limited semantic information from isolated data points. Then, on the spatial dimension, we minimize the uncertainty of the input graph structure by constructing an encoding tree and designing a contrastive loss based on node pairs within the same community, enabling the spatial encoder to capture higher-order spatial semantics. Finally, we train the model using the STSCIB as the objective.

Generally, the contributions of this paper are as follows:

1) We establish a novel information-theoretic principle for STSF by extending the IB theory, demonstrating that optimal representations require constrained conditional mutual information with both graph structures and time series.

2) We propose a novel framework STSCIB for STSF to capture spatial information by incorporating encoding tree with contrastive learning within communities and extract temporal patterns by Transformer-based pre-training.

3) We conduct extensive experiments on real-world datasets, and the results demonstrate that the proposed framework outperforms the state-of-the-art comparison methods.

2 Preliminary

In this section, we briefly introduce the concepts of spatio-temporal sequence forecasting, information bottleneck and structural information theory.

2.1 Spatio-temporal Sequence Forecasting

Spatio-temporal sequence forecasting (STSF) aims to predict future sequences based on historical observations. Given a graph $G = (V, E, \mathbf{A})$, where V and E represent the set of nodes and edges respectively, the adjacency matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ is defined such that $A_{ij} = 1$ if $(v_i, v_j) \in E$ and 0 otherwise, with $N = |V|$ denoting the number of nodes. The spatio-temporal sequence (STS) can be represented as a tensor $\mathbf{X} \in \mathbb{R}^{T\times N\times C}$, comprising T time steps. The snapshot at time step t is denoted $\mathbf{x}_t \in \mathbb{R}^{N\times C}$, where C is the number of features. Then the STSF problem can be formalized as follows: given spatio-temporal data $\mathbf{X}^{1:T_h}$ for the past T_h-steps and the input graph G, the goal is to predict the future sequence of the features on each node with a learning function f, which can be expressed as : $f(\mathbf{X}^{1:T_h}, G) \to [\mathbf{X}^{T_h+1:T_h+T_f}]$, where $[\mathbf{X}^{T_h+1:T_h+T_f}]$ and $[\mathbf{X}^{1:T_h}]$ are the STS with length T_f and T_h respectively.

2.2 Structural Information Theory

Structural entropy is an extension of Shannon entropy for measuring the uncertainty of a graph by hierarchical partitioning [7], aiming to obtain the hierarchical structures of graph compression, The theory is mainly consists of the following two parts [19]:

Encoding Tree. Given a graph $G(V, E, \mathbf{A})$, the encoding tree $\mathcal{T}$ implements encoding and abstracting the graph G into a hierarchical structures with the following properties: i) each node $\alpha \in \mathcal{T}$ is associated with a non-empty node set $\mathcal{T}_\alpha \subseteq V$, and the root node λ, $\mathcal{T}_\lambda$ includes all nodes in G, denoted as $\mathcal{T}_\lambda = V$. ii) For each leaf node α in encoding tree $\mathcal{T}$ has a unique corresponding node v_γ, $\mathcal{T}_\gamma = V_\gamma$. iii) For each non-root and non-leaf node α, $\mathcal{T}_\alpha$ includes some nodes of G. Its ith child node is denoted as $\alpha^{<i>}$, and all $\mathcal{T}_{\alpha^{<i>}}$ are disjointed, i.e., $\mathcal{T}_\alpha = \bigcap_{i=1}^{m} \mathcal{T}_{\alpha^{<i>}} = \emptyset$, where m is the number of α's children.

Structural Entropy. Structural information theory also known as Structural entropy, which is determined by the encoding tree and the graph together.The encoding tree generated by minimizing the graph's structural entropy compresses the most knowledge and is also optimal to represent graph hierarchical community structure [19]. The structural entropy can be formulated :

$$H^K(G) = \min_{\forall \mathcal{T}: height(\mathcal{T}) \leq K} H^{\mathcal{T}}(G) = \sum_{\alpha \in T, \alpha \neq \lambda} -\frac{g_\alpha}{2m} log \frac{V_\alpha}{V_{\alpha^-}} \tag{1}$$

where g_α is the sum of weights of edges from the nodes in $\mathcal{T}_\alpha$ to those outside of $\mathcal{T}_\alpha$, $V_\alpha = \sum_{v \in T_\alpha} d_v$ is the volume of $\mathcal{T}_\alpha$, and α^- is the parent node of α, $height(\mathcal{T})$ refers to the height of the encoding tree, and the value range of $height(\mathcal{T})$ is all encoding trees whose height does not exceed K. Given an encoding tree $\mathcal{T}$, its structural entropy is the sum of the entropy of all non-root nodes, i.e. $H^{\mathcal{T}}(G) = \sum_{\alpha \in \mathcal{T}, \alpha \neq \lambda} H^{\mathcal{T}}(G; \alpha)$.

3 Methodology

In this section, we elaborate on proposed spatio-temporal sequence conditional information bottleneck (STSCIB). We first derive the principle of the STSCIB, then instantiate it based on structural information theory and Transformer model to achieve more robust and accurate spatio-temporal sequence forecasting (STSF). The overall architecture of STSCIB is illustrated in Fig.1.

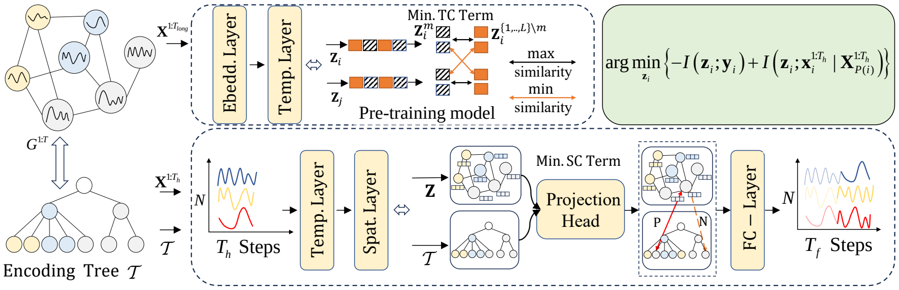

Fig. 1. The framework of our method.

3.1 Spatio-temporal Sequence Conditional Information Bottleneck

Our aim is to learn the optimal representation $\mathbf{Z}$ for STSF from historical observations $\mathbf{X}^{1:T_h}$, which maximizes relevant information about future observations

$\mathbf{Y}$ while eliminating irrelevant information as much as possible. To achieve this goal, we adapt the IB framework to capture spatio-temporal dependencies from both spatial structures and temporal patterns perspectives. Specifically, we focus on the impact of spatio-temporal sequences within community (STSC), defined as $\mathbf{X}_{P(i)}^{1:T_h} = \{\mathbf{X}_j^{1:T_h} : v_j \in V, v_i v_j \in E, v_j \neq v_i\}$, where spatial connections and temporal characteristics are jointly considered. By treating the STSC as a known condition for the conditional reconstruction term, we alleviate the problem of contextual information loss caused by the IB regularization, and propose a novel information-theoretic approach for STSF as follow proposition 1.

Proposition 1. *Give the T_h-step historical spatio-temporal sequence $\mathbf{x}_i^{1:T_h}$ and $\mathbf{X}_{P(i)}^{1:T_h}$, the optimal representation that minimally and sufficiently captures the relevant information for spatio-temporal sequence forecasting can be formulated as the following optimization problem*:

$$\mathbf{z}_i = \arg\min_{\theta_s,\theta_t} -I\left(\mathbf{y}_i;\mathbf{z}_i\right) + \beta I\left(\mathbf{z}_i;\mathbf{x}_i^{1:T_h}\middle|\mathbf{X}_{P(i)}^{1:T_h}\right) \tag{2}$$

where $\beta(\beta > 0)$ is a hyper-parameter. For simplicity, we set $\mathbf{x}_i^{T_h+1:T_h+T_f} = \mathbf{y}_i$. Conditioning on $\mathbf{X}_{P(i)}^{1:T_h}$, Eq.2 guides us to find latent representation $\mathbf{z}_i$ and the corresponding model inference parameters θ_s, θ_t, which capture the relevant spatio-temporal information needed for predicting the next T_f steps, while $\mathbf{z}_i$ discards information unrelated to the target sequence. Next, we instantiate Proposition 1.

3.2 Instantiating STSCIB

We now introduce a framework based on structural entropy and Transformer as an instance of STSCIB, illustrated in Fig.1.

3.2.1 Instantiating Prediction Term. Since the mutual information are intractable, we follow the derivation introduced in some work [2, 15], finding a lower bound of the prediction term as the following:

$$I\left(\mathbf{z}_i;\mathbf{y}_i\right) = \mathbb{E}_{\mathbf{z}_i,\mathbf{y}_i}[\log(\frac{\mathbb{P}(\mathbf{y}_i|\mathbf{z}_i)}{\mathbb{P}(\mathbf{y}_i)})] \geq \mathbb{E}_{\mathbf{z}_i,\mathbf{y}_i}[\log \mathbb{Q}_1(\mathbf{y}_i|\mathbf{z}_i)] \tag{3}$$

where $\log \mathbb{Q}_1(\mathbf{y}_i|\mathbf{z}_i)$ represents the variational approximation of $\mathbb{P}(\mathbf{y}_i|\mathbf{z}_i)$, which can be optimized by prediction loss [2, 4].

3.2.2 Instantiating Conditional Reconstruction Term. To obtain optimal task-relevant representations, we build on insights from previous work [18, 15], designing different modules to extract spatio-temporal features.

(1) Minimizing Spatial Conditional Reconstruction Term. We employ the mutual information chain rule to decompose the spatial conditional reconstruction term as follow:

$$\min I(\mathbf{z}_i;\mathbf{x}_i^{1:T_h}|\mathbf{X}_{P(i)}^{1:T_h}) = \min I(\mathbf{z}_i;\mathbf{X}_P^{1:T_h}) - I(\mathbf{z}_i;\mathbf{X}_{P(i)}^{1:T_h}) \tag{4}$$

Intuitively, the spatial conditional reconstruction term is decomposed into two components: (i) minimizing the mutual information between latent representation $\mathbf{z}_i$ and the community spatio-temporal sequence $\mathbf{X}_P^{1:T_h}$ centered on sequence $\mathbf{x}_i^{1:T_h}$, and (ii) maximizing the information from $\mathbf{X}_{P(i)}^{1:T_h}$ to capture spatio-temporal dependencies. This addresses the loss of useful information, while directing the model's focus to the STSC, thus solving the challenges posed by the IB.

Minimizing $I(\mathbf{z}_i; \mathbf{X}_P^{1:T_h})$. We convert the input STS $\mathbf{X}^{1:T_h}$ into a spatio-temporal graph sequence $G^{1:T_h}$. The community spatio-temporal sequence is then generated by sampling a subgraph G_s from the input graph G. We assume that G_s retains the sufficient spatial structure information [4], as it derives from G. Thus, we have:

$$\min I(\mathbf{z}_i; \mathbf{X}_P^{1:T_h}) = \min I(\mathbf{z}_i; G_s^{1:T_h}) \Longleftrightarrow T_h \min H(G_s) \tag{5}$$

Therefore, the goal of Eq.(5) is to generate a subgraph with sufficient information, while minimizing its uncertainty. In structural information theory, the minimization of graph uncertainty can be achieved using an encoding tree, where the encoding tree $\mathcal{T}$ is a lossless representation of graph structure data. Thus, we use the encoding tree $\mathcal{T}$ as the sampled graph structure G_s, constructed according to the approach in [4]. Each layer of the encoding tree corresponds to a partition of the graph's node set, representing potential communities at different spatial scales. This naturally provides a theoretically grounded community spatio-temporal sequence, enabling an effective realization of $I(\mathbf{z}_i; \mathbf{X}_P^{1:T_h}) = I(\mathbf{z}_i; \mathcal{T}_{i^-}^{1:T_h})$. Consequently, the first term in Eq.(5) can be constrained as follows (derivation process in [2]):

$$I(\mathbf{z}_i; \mathbf{X}_P^{1:T_h}) = I(\mathbf{z}_i; \mathcal{T}_{i^-}^{1:T_h}) \leq \mathbb{E}_{\mathcal{T}_{i^-}^{1:T_h}} \left[KL(\mathbb{P}(\mathbf{z}_i | \mathcal{T}_{i^-}^{1:T_h}) \| \mathbb{Q}_2(\mathbf{z}_i) \right] \tag{6}$$

where $\mathbb{Q}_2(\mathbf{z}_i)$ is variational approximation for the marginal distribution $\mathbb{P}(\mathbf{z}_i)$, and $KL(\cdot)$ denotes the Kullback-Leibler divergence function.

Maximizing $I(\mathbf{z}_i; \mathbf{X}_{P(i)}^{1:T_h})$. Building on the proof of the first term, we incorporate the hierarchical communities of the encoding tree as spatial constraints for STSCIB framework. Additionally, the minimization of InfoNEC through contrastive learning on latent representations can approximately maximize the corresponding mutual information [2]. Therefore, we combine the encoding tree with contrastive learning and define a novel contrastive learning loss to maximize $I(\mathbf{z}_i; \mathbf{X}_{P(i)}^{1:T_h})$, as follows:

$$I\left(\mathbf{z}_i; \mathcal{T}_{i^- \backslash i}^{1:T_h}\right) \geq \mathbb{E}_{\mathcal{T}_{i^- \backslash i}} \left[\sum_{k=1}^{K} \varepsilon_k \log \left(\frac{I(\mathbf{z}_i; \mathcal{T}_{i^- \backslash i}^{k})}{\sum_{j=1}^{N} I(\mathbf{z}_i; \mathcal{T}_j)} \right) \right] \tag{7}$$

where $\mathcal{T}_{i^- \backslash i}^{k}$ denotes the set of nodes in the community at the k-th level of the encoding tree, excluding node i itself, the coefficient $\varepsilon_k = \epsilon(1-\epsilon)^k, 0 < \epsilon < 1$ is associated with the height of the encoding tree.

(2) Minimizing Temporal Conditional Reconstruction Term. Similar to (1) we omit repeated details for brevity. To overcome the difficulty of identifying meaningful trends when simply masking a single value, we introduce long-term spatio-temporal sequence $\mathbf{X}^{T_{long}} \in \mathbb{R}^{T_{long} \times N \times C}$ to pre-train the temporal encoder, enabling it to capture long-term temporal patterns. Thus, we have:

$$\begin{aligned} I\left(\mathbf{Z}_i^m; \mathbf{X}_i^m \middle| \mathbf{X}_i^{\{1,\ldots,L\}\backslash m}\right) \leq \mathbb{E}_{\mathbf{X}_i^{\{1,\ldots,L\}\backslash m}} \left[KL\left(\mathbb{P}(\mathbf{Z}_i^m | \mathbf{X}_i^{\{1,\ldots,L\}\backslash m}) \| \mathbb{Q}_3(\mathbf{Z}_i^m)\right)\right] \\ - \mathbb{E}_{\mathbf{X}_i^{\{1,\ldots,L\}\backslash m}} \log\left(\frac{I\left(\mathbf{Z}_i^m; \mathbf{Z}_i^{um}\right)}{\sum_{\mathbf{Z}_j \in \mathbf{X}_j} I\left(\mathbf{Z}_i^m; \mathbf{Z}_j\right)}\right) \end{aligned} \tag{8}$$

where $\mathbf{Z}_i^{um}$ and $\mathbf{Z}_i^m$ represent the unmasked positive samples and the masked representation of sequence $\mathbf{x}_i$, respectively, while $\mathbf{Z}_j$ denotes negative samples from other sequences $\{\mathbf{x}_j, j \neq i, j \in N\}$. The notation $\mathbf{X}_i^{\{1,\ldots,L\}\backslash m} \in \mathbb{R}^{(L-m) \times T_h \times C}$ denotes the unmasked features of $\mathbf{x}_i$.

4 Experimental Study

4.1 Experimental Settings

Datasets. We evaluate the performance of STSF methods using five real-world traffic flow datasets, including: (1) PEMS04, PEMS07, and PEMS08, constructed by STSGCN [13], and (2) PEMS-BAY and METR-LA, collected by DCRNN [8]. These datasets vary in terms of the number of nodes, with data recorded at 5-minute intervals, covering different time periods and geographic regions. For the first set of datasets, the input graph G is derived from the connectivity between nodes. In contrast, for the second set, the input graphs are constructed based on the road network distance, using a thresholded Gaussian kernel method. Detailed information about the datasets is provided in Table 1.

Comparison Methods. We compare our method with the state-of-the-art STSF methods, including FC-LSTM [14], GWNet [17], STID [11], STPGNN [6], STNorm [3], PDFormer [5], STAEformer [9], MultiSPANS [19], Timemixer [16] and STEP [12]. These comparison methods are employed not only to evaluate the effectiveness of our framework, but also to validate the capability of our theoretic guidance in capturing useful information.

Table 1. Dataset statistics.

Dataset	Nodes	Edges	Time Interval	Time Steps
PEMS04	307	209	5 min	16,992
PEMS07	883	866	5 min	28,224
PEMS08	170	137	5 min	17,856
PEMS-BAY	325	2694	5 min	52,116
METR-LA	207	1722	5 min	34,272

Evaluation Metrics. We adopt three widely used metrics Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) to evaluate the effectiveness of STSF methods. For a fair comparison, we follow the settings from [12], we divide the PEMS04, PEMS07, PEMS08, datasets into training, validation, and test sets according to a $6:2:2$ ratio. For METR-LA and PEMS-BAY datasets, the training, validation, and test ratio is

Table 2. Effectiveness comparison with other STSF methods over all datasets. The best and second best results are indicated in **bold** and underline, respectively.

Method	PEMS04			PEMS07			PEMS08			PEMS-BAY			METR-LA		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
FC-LSTM	27.15	41.52	18.33	30.02	45.88	13.15	22.24	34.11	14.21	2.16	4.51	5.08	3.86	7.41	11.23
GWNet	22.46	35.58	15.42	23.30	37.59	10.15	17.82	27.95	11.78	1.79	4.06	4.13	3.58	7.16	10.23
Timemixer	19.88	32.00	12.22	21.84	34.68	9.03	16.23	25.36	9.98	1.84	3.93	4.13	5.24	12.19	9.66
MultiSPANS	19.07	30.46	13.36	-	-	-	14.37	23.87	9.92	1.65	3.58	3.84	3.32	6.41	9.40
STNorm	18.96	30.98	13.05	20.59	34.86	8.61	15.37	24.80	9.91	1.63	3.69	3.72	3.19	6.56	8.70
STPGNN	18.56	31.43	12.83	20.75	33.91	9.38	14.15	24.31	9.47	1.68	3.74	3.81	3.15	6.24	8.69
STID	18.41	29.95	12.68	19.64	32.72	8.35	14.24	23.37	9.40	1.62	3.61	3.62	3.18	6.48	9.07
PDFormer	18.83	30.07	12.61	20.06	32.96	8.46	13.69	23.71	9.21	1.62	3.68	3.63	3.50	8.00	8.74
STAEformer	18.30	30.37	12.04	19.23	32.61	8.04	13.49	23.36	8.91	1.61	3.58	3.60	3.01	6.04	8.22
STEP	18.30	29.89	12.50	19.33	**32.25**	8.21	14.00	23.39	9.56	1.59	3.52	3.44	3.01	6.01	8.08
Ours	**18.18**	**29.89**	**11.95**	**19.01**	32.33	**7.95**	**13.41**	**23.05**	**8.75**	**1.53**	**3.48**	**3.39**	**2.89**	**5.95**	**7.98**

set to 7 : 1 : 2, and set the lengths of both the input T_h and output T_f to 12 time steps in all datasets. To prevent information leakage during pre-training, all datasets share the same pre-trained weight parameters, with an input length of 4032 time steps and a masking rate of 75%.

4.2 Experimental Results

Exp-1: Effectiveness Evaluation. In the first set of tests, we evaluate the effectiveness of our method comparing with other STSF methods. The encoding tree height K is fixed to 3, and the results are reported in Table 2.

The result tells us that: Our method achieves the best MAE, RMSE and MAPE on all datasets, except for performing the second best in RMSE on PEMS07 dataset, which shows the superiority of the proposed STSCIB compared with other methods. Note that MultiSPANS cannot be conducted on PEMS07 dataset and is thus marked as '-' (Detailed reasons can be found in its code repository). By extracting the hierarchical community information as spatial features based on the encoding tree and capturing temporal patterns in time series based on conditional information, our method demonstrates significant improvements compared with other STSF methods on three metrics over all datasets. These verify the effectiveness of our proposed method.

Table 3. Ablation study on PEMS04, PEMS08 and METR-LA.

Method	PEMS04			PEMS08			METR-LA		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
w/o T-Trans	18.29	30.47	12.22	13.65	23.22	8.89	2.95	6.02	8.17
w/o S-Trans	18.33	30.01	12.36	13.55	23.18	8.85	2.92	5.96	8.07
w/o ST-Trans	19.38	31.23	12.72	14.67	23.73	9.48	3.04	6.31	8.56
IB	19.20	30.68	12.82	14.52	23.65	9.25	2.98	6.13	8.25
ours	**18.18**	**29.89**	**11.95**	**13.41**	**23.05**	**8.75**	**2.89**	**5.95**	**7.98**

Exp-2: Ablation Study. In the second set of tests, we evaluate the effectiveness of each component in our proposed objective function, comparing with traditional IB method. We denote our model without spatial feature, temporal feature and spatio-temporal feature extraction components as w/o S-Trans, w/o

T-Trans and w/o ST-Trans, respectively. The results on PEMS04, PEMS08 and METR-LA datasets are reported in Table 3.

The result tells us that: (a) the significance of various components on the performance of our model. (b) Whether removing the w/o T-Trans or w/o S-Trans, our method performs better than simply using IB, which fully verifies the advantage of our theory in effectively capturing the underlying spatio-temporal dynamics. (c) When both w/o T-Trans and w/o S-Trans are removed, performance decreases and falls below that of traditional IB methods. We speculate that in the absence of both spatio-temporal components, the model degenerates into a conventional spatio-temporal sequence forecasting method that solely maximizes the information between the latent representation and the target sequence, which of course performs worse than information-theoretic approaches.

Exp-3: Impacts of Encoding Tree. In the third set of tests, we analyze the effectiveness of encoding tree by comparing it with three community discovery methods kNN, Louvain [1] and Infomap [10]. The results are reported in Fig. 2.

The results tell us that: (a) encoding tree outperforms other methods on extracting the community information. (b) Infomap achieves the second best method since it could discover more effective community structure by using information theory while kNN uses only the similarity of node attributes and Louvain uses the degree of nodes.

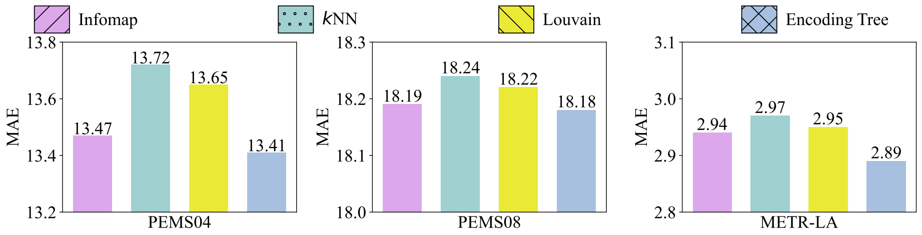

Fig. 2. Comparison of different community discovery methods.

5 Conclusion

In this work, we propose a structural entropy based spatio-temporal sequence forecasting (STSF) model. We first establish a theoretical principle to guide the model in learning optimal representations for STSF. We then utilize community information from the encoding tree to instantiate this theory, which addresses the limitations of the IB principle in capturing underlying spatio-temporal dependencies. Finally, we train the model using the STSCIB as the objective function. Extensive experiments on five datasets prove the effectiveness of our method.

Acknowledgments. This work is supported by the Key Project of Fundamental Research of Yunnan Province (202401AS070138), Program of Yunnan Key Laboratory of Intelligent Systems and Computing (202405AV340009), and Yunnan Fundamental Research Project (202301AT070193).

References

1. Blondel, V.D., Guillaume, J., Lambiotte, R.: Fast unfolding of communities in large networks: 15 years later. CoRR **abs/2311.06047** (2023)
2. Choi, M., Lee, C.: Conditional information bottleneck approach for time series imputation. In: ICLR (2024)
3. Deng, J., Chen, X., Jiang, R., Song, X., Tsang, I.W.: St-norm: Spatial and temporal normalization for multi-variate time series forecasting. In: SIGKDD. pp. 269–278 (2021)
4. Duan, L., Chen, X., Liu, W., Liu, D., Yue, K., Li, A.: Structuarl entropy based graph structure learning for node classification. In: AAAI. pp. 8372–8379 (2024)
5. Jiang, J., Han, C., Zhao, W.X., Wang, J.: Pdformer: Propagation delay-aware dynamic long-range transformer for traffic flow prediction. In: AAAI. pp. 4365–4373 (2023)
6. Kong, W., Guo, Z., Liu, Y.: Spatio-temporal pivotal graph neural networks for traffic flow forecasting. In: AAAI. pp. 8627–8635 (2024)
7. Li, A., Pan, Y.: Structural information and dynamical complexity of networks. IEEE Transactions on Information Theory **62**(6), 3290–3339 (2016)
8. Li, Y., Yu, R., Shahabi, C., Liu, Y.: Diffusion convolutional recurrent neural network: data-driven traffic forecasting. In: ICLR (2018)
9. Liu, H., Dong, Z., Jiang, R., Deng, J., Deng, J., Chen, Q., Song, X.: Staeformer: Spatio-temporal adaptive embedding makes vanilla transformer SOTA for traffic forecasting. CoRR **abs/2308.10425** (2023)
10. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences (PNAS) **105**(4), 1118–1123 (2008)
11. Shao, Z., Zhang, Z., Wang, F., Wei, W., Xu, Y.: Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting. In: CIKM. pp. 4454–4458 (2022)
12. Shao, Z., Zhang, Z., Wang, F., Xu, Y.: Pre-training enhanced spatial-temporal graph neural network for multivariate time series forecasting. In: SIGKDD. pp. 1567–1577 (2022)
13. Song, C., Lin, Y., Guo, S., Wan, H.: Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. In: AAAI. pp. 914–921 (2020)
14. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: NIPS. pp. 3104–3112 (2014)
15. Tang, J., Xia, L., Huang, C.: Explainable spatio-temporal graph neural networks. In: CIKM. pp. 2432–2441 (2023)
16. Wang, S., Wu, H., Shi, X., Hu, T., Luo, H., Ma, L., Zhang, J.Y., Zhou, J.: Timemixer: Decomposable multiscale mixing for time series forecasting. In: ICLR (2024)
17. Wu, Z., Pan, S., Long, G., Jiang, J., Zhang, C.: Graph wavenet for deep spatial-temporal graph modeling. In: IJCAI. pp. 1907–1913 (2019)
18. Yuan, H., Sun, Q., Fu, X., Ji, C., Li, J.: Dynamic graph information bottleneck. In: WWW. pp. 469–480 (2024)
19. Zou, D., Wang, S., Li, X., Peng, H., Wang, Y., Liu, C., Sheng, K., Zhang, B.: Multispans: A multi-range spatial-temporal transformer network for traffic forecast via structural entropy optimization. In: WSDM. pp. 1032–1041 (2024)