InfoST: An Information-Theoretic Framework for Evaluating Spatio-Temporal Prediction

Nishan Khanal
College of Computing
Grand Valley State University
Allendale, MI, USA
khanalni@mail.gvsu.edu

Haoyu Li
College of Computing
Grand Valley State University
Allendale, MI, USA
lihao@gvsu.edu

Yong Zhuang
College of Computing
Grand Valley State University
Allendale, MI, USA
zhuangyo@gvsu.edu

***Abstract*—Performance of spatio-temporal predictors is notoriously dataset-dependent, complicating fair comparison and principled model selection. We address this challenge by introducing an information-theoretic evaluation framework that decomposes dataset structure into two complementary indices: *temporal predictability* and *spatial predictability*. Grounded in mutual-information measurements, these indices quantify how much signal is available along the temporal and non-temporal (e.g., spatial, variable, or channel) axes. Building on this formulation, we construct a controllable synthetic benchmark in which temporal and spatial regularities can be independently tuned, enabling factorial experiments that systematically vary task difficulty and structure. Evaluating representative models shows that rankings vary systematically with the balance of temporal versus spatial signal. On synthetic data, S-Mamba and iTransformer lead when temporal predictability dominates; their advantage narrows as spatial predictability increases. Revisiting widely used real-world suites through our lens reveals the same pattern: datasets with higher spatial mutual information favor models that explicitly exploit inter-variable structure, while low-spatial-MI datasets favor temporal-axis methods.**

***Index Terms*—Synthesized data, Time series prediction, Spatio-temporal data, Transformer, State-space models**

I. Introduction

Spatio-temporal prediction is a core task across numerous domains, including IoT, healthcare, environmental sciences, and criminology, where data exhibit dependencies over both time and space. Deep neural networks have been widely adopted for these problems due to their strong empirical performance. Despite a steady stream of new architectures tailored to specific datasets or application niches, relatively few models function as general-purpose backbones for time-series prediction. Recent surveys [1] identify Transformer-based architectures [2] as particularly promising in this role, while state-space models (SSMs), exemplified by Mamba [3], offer a more computationally efficient alternative to standard Transformers.

As architectures and application areas proliferate, fair and reproducible comparison has become increasingly challenging. Performance is often highly dataset-dependent: a model that excels on one benchmark may be outperformed on another. For example, a recent study [4] shows that Mamba achieves leading results on some forecasting datasets, but underperforms relative to the Transformer variant iTransformer [5] on others. Such variability complicates generalization from single-study findings and can hinder principled evaluation and adoption of new models.

These issues are amplified when spatial structure is explicitly modeled alongside temporal dynamics. While historical temporal patterns are indispensable for forecasting, spatial correlations can materially refine predictions (e.g., learning geographical regularities to forecast crime hotspots). Prior work, including the "inverted" Transformer (iTransformer) [5], demonstrates that the axis along which attention is applied (temporal versus spatial) can substantially affect accuracy. More broadly, the "spatial" dimension may denote any non-temporal axis, such as variables or feature channels in multivariate series. Given these intertwined factors, superiority claims based on a single dataset are insufficient. There is a clear need for a principled framework that disentangles sources of variation from data domains, spatio-temporal structure, and modeling choices to enable fair, comprehensive evaluation of spatio-temporal prediction models.

To address the foregoing challenges, we introduce an *information-theoretic* evaluation framework that decomposes dataset-intrinsic variability into two components: *temporal predictability* and *spatial predictability*. Using these indices, we construct a controllable synthetic benchmark in which the degree of temporal and spatial regularity can be independently tuned, enabling systematic stress-testing of spatio-temporal predictors under diverse structural regimes. We further apply the framework and benchmark to a range of representative models to assess robustness and diagnose failure modes.

Contributions. This paper makes three primary contributions:

- We formalize two complementary measures, temporal predictability and spatial predictability, that

quantify dataset structure along temporal and non-temporal axes within a unified, information-theoretic lens.
- We develop a synthetic spatio-temporal benchmark with independent control of temporal and spatial regularities, supporting reproducible, factorial evaluation across a spectrum of difficulty settings.
- We demonstrate the utility of the framework through empirical studies on diverse prediction architectures, revealing how model performance depends on the interplay between temporal and spatial structure.

II. Related Work

We review prior work most relevant to our goal of fairly assessing spatio-temporal (ST) predictors under varying data structures. We focus on foundation architectures, RNN/GNN hybrids, Transformers, and state-space models (SSMs), and on evaluation protocols and synthetic benchmarks.

A. Architectures for Spatio-Temporal Prediction

Early neural approaches to spatiotemporal prediction extended recurrent models with explicit spatial structure. For instance, crime forecasting work used recurrent networks to capture temporal dependencies but treated space coarsely (e.g., fixed cells) [6]. Subsequent research introduced spatiotemporal graphs and applied message-passing methods (e.g., GCN, GAT) to jointly model spatial dependencies with temporal dynamics [7]–[9].

Transformer-based models introduced attention to capture long-range temporal and spatial interactions. GMAN, for example, uses attention to model both axes and improves multi-step forecasting [2], [10]. More recently, selective SSMs (e.g., Mamba) offer linear-time sequence modeling with scan/prefix-sum parallelization and hardware-aware kernels, yielding favorable accuracy–efficiency trade-offs relative to standard Transformers [3]. Comparative studies report that Mamba can be state-of-the-art on some datasets yet lose on others, underscoring dataset dependence [4], [5]. Integration of SSM blocks with graph modules (e.g., STG-Mamba) has shown similar benefits on ST graphs [11].

B. Evaluation Protocols and Synthetic Benchmarks

Prevailing evaluation practice relies on heterogeneous real-world datasets and aggregate metrics, which obscures how model rankings shift with underlying data structure (cf. survey discussions in [1]). Existing synthetic settings typically vary noise, sparsity, or graph topology, but seldom *isolate* temporal versus spatial signal strength. Our work is complementary: we introduce information-theoretic indices that separately quantify temporal and spatial predictability and a controllable benchmark that factorially varies these axes. This design enables systematic stress-testing and clearer attribution of performance differences across model families (RNN/GNN, Transformer, SSM).

III. Methodology

In this section, we (i) detail the synthetic data generation process and its controllable parameters, and (ii) introduce an information-theoretic framework with two indices, *temporal predictability* and *spatial predictability*, for quantifying dataset structure.

A. Dataset Generation

We generate two families of synthetic datasets: fields on a 2D spatial grid and sequences with a 1D spatial axis. Here, "dimensionality" refers to the spatial dimension only; the full tensors are thus $2\mathrm{D} \times \mathrm{time}$ (3D) and $1\mathrm{D} \times \mathrm{time}$ (2D), respectively. In both cases, spatial and temporal patterns are synthesized independently and then mixed via a parameter $\alpha \in [0, 1]$, where larger α increases spatial regularity and decreases temporal regularity.

1) Core Design Principle: Let $\mathbf{S}$ denote the spatial pattern and $\mathbf{T}(t)$ denote the temporal pattern at time step t. For each dataset, we generate the samples as:

$$\mathbf{X}(t) = \alpha \cdot \mathbf{S} + (1 - \alpha) \cdot \mathbf{T}(t), \tag{1}$$

where $\alpha \in [0, 1]$ controls the contribution of temporal and spatial information. After the construction of the full dataset $\mathbf{X} \in \mathbb{R}^{T \times N}$, we add Gaussian noise as:

$$\mathbf{X} = \mathbf{X} + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{2}$$

where ϵ has the same shape as $\mathbf{X}$.

2) Synthetic 2D Spatial Dataset: For the spatial component, a Gaussian random field is generated over an $H \times W$ grid to produce smoothly varying values capturing the spatial correlations between the neighbouring nodes. For this, we used the gstools [12] library, which uses the variogram γ given by:

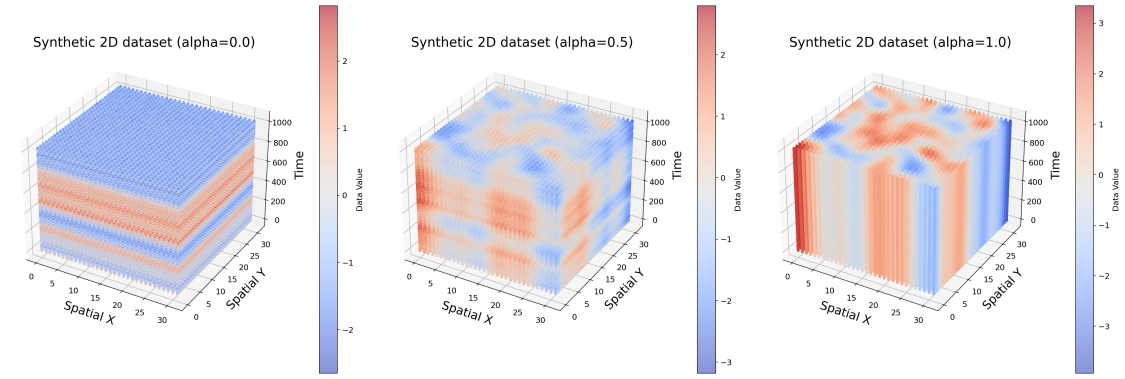
$$\gamma(r) = \sigma^2 \left(1 - \exp\left(-\left(s \cdot \frac{r}{\ell}\right)^2\right)\right) + n, s = \frac{\sqrt{\pi}}{2} \tag{3}$$

where r is the euclidean distance between two nodes, l is the (main) correlation length, s is a rescaling factor to adjust model representation, σ^2 is the variance, and n is the nugget or sub-scale variance.

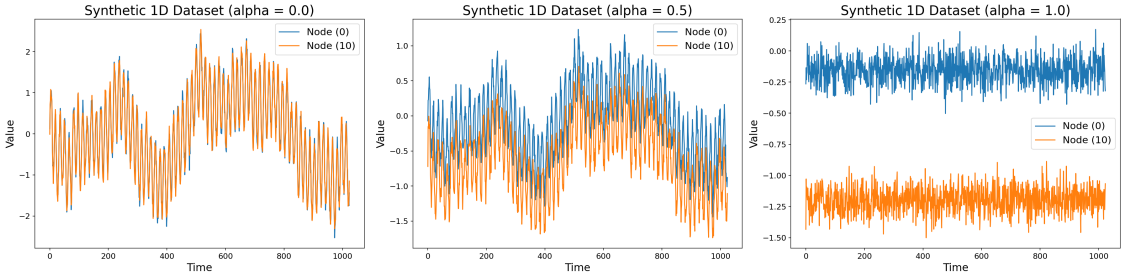
The temporal signal is defined as the combination of a sinusoidal pattern and a stochastic cumulative noise process as:

$$\mathbf{T}(t) = sin(\frac{2\pi \cdot nc \cdot t}{T}) + \sum_{k=1}^{t} \eta_k, \eta_k \sim \mathcal{N}(0, \sigma^2). \tag{4}$$

The sinusoidal part is used to define a periodic trend, where nc specifies the number of oscillations of the



(a) Synthetic 2D spatial dataset



(b) Synthetic 1D spatial dataset

Fig. 1: Synthetic datasets with different values of α. Here, $\alpha = 0$ corresponds to a purely temporal signal, while $\alpha = 1$ corresponds to a purely spatial signal with no temporal variation along the time dimension. (a) The 2D dataset shows spatial maps under varying α. (b) The 1D dataset illustrates temporal signals for two representative nodes ($n = 0$ and $n = 10$), where each curve represents the temporal signal at a fixed spatial position.

sine wave over the entire sequence. For example, with $nc = 66$, there are 66 sine waves across T time steps. The cumulative Gaussian noise part is used to introduce randomness that evolves over time. As the noise is accumulated step by step, the fluctuations are not independent but temporally correlated. This resembles a random walk process. Therefore, the temporal component is the combination of predictable periodic structure and long-term correlated noise, which makes the signal more realistic as compared to a pure sine wave or independent noise alone. For each time step t, the temporal signal is blended with the fixed spatial map through α as described in equation 1. The final dataset is then obtained by adding Gaussian noise.

3) Synthetic 1D Spatial Dataset: We also generate a 1D "spatial" dataset to model cases where the non-temporal axis is either a 1D space or non-physical space, e.g., variables or feature channels in a multivariate series.

The temporal component is generated using a sinusoidal signal with cumulative Gaussian noise evolving across time exactly as described by equation 4. The spatial component is similar to the temporal one. We consider N nodes arranged along a one-dimensional lattice. For each node $n \in \{1, \ldots, N\}$, the spatial signal is generated analogously by

$$S_t(n) = \sin\Big(\tfrac{2\pi \cdot nc \cdot n}{N}\Big) + \sum_{k=1}^{n} \eta_k, \eta_k \sim \mathcal{N}(0, \sigma^2) \quad (5)$$

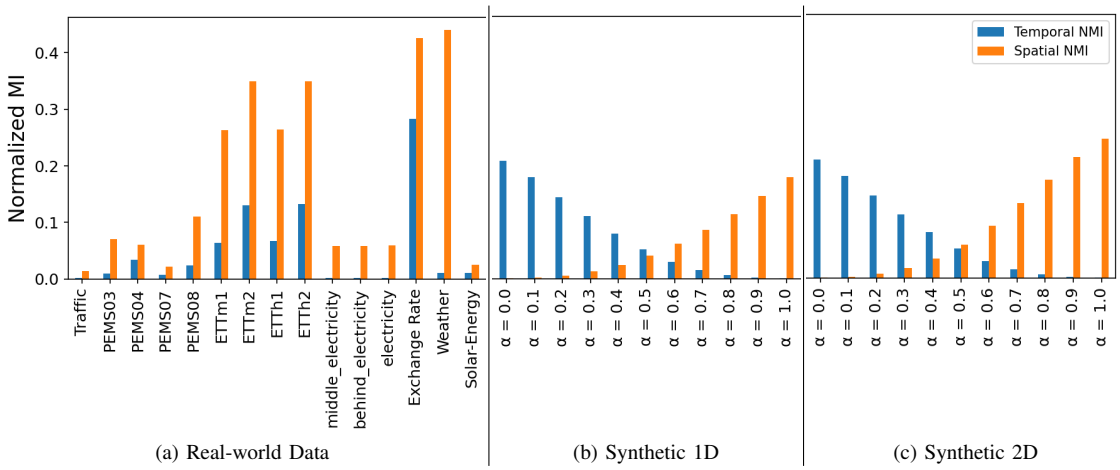
which introduces smoothly varying spatial correlations with stochastic fluctuations. Here, n denotes the spatial index of a node. For each time slice t, the spatial signal and the temporal signal at the node level are blended together using the parameter α. Gaussian noise is then added to form the final dataset $\mathbf{X} \in \mathbb{R}^{T \times N}$.

B. Information-Theoretic Analysis

To evaluate the spatial and temporal dependencies present in our synthetic datasets, we compute information-theoretic measures based on entropy and mutual information. These measures provide a way of evaluating how much information is shared between different components of the data.

1) Entropy: The entropy $H(X)$ measures the uncertainty of a random variable X. For a discrete distribution with probabilities p_i, entropy is given as:

$$H(X) = -\sum_i p_i \log_2 p_i \quad (6)$$



(a) Real-world Data (b) Synthetic 1D (c) Synthetic 2D

Fig. 2: Normalized Mutual Information for real-world and synthetic datasets

In our implementation, we first discretize the continuous-valued data into a fixed number of bins, and then compute the histogram-based estimate of entropy.

2) Mutual Information: Mutual information (MI) measures the reduction in uncertainty of one variable given the knowledge of another. For two variables A and B:

$$I(A;B) = H(A) + H(B) - H(A,B) \tag{7}$$

where $H(A,B)$ is the joint entropy of A and B.

In our implementation, given $X \in \mathbb{R}^{T\times N}$ denotes the dataset with T time steps and N nodes, where each entry in the dataset can be indexed by (t,n), where $t \in \{0,1,\ldots,T-1\}$ and $n \in \{0,1,\ldots,N-1\}$. To align the data with their indices, we flatten X into a 1D vector of length $T \cdot N$ and construct the corresponding index arrays:

- Time index array (I_T) is created by repeating each time step t exactly N times. For instance, with $T = 3, N = 4$: $I_T = [0,0,0,0,1,1,1,1,2,2,2,2]$
- Node index array (I_N) is created by tiling the node indices $\{0,\ldots,N-1\}$ across all the T time steps. For example, with $T = 3, N = 4$: $I_N = [0,1,2,3,0,1,2,3,0,1,2,3]$

For 2D spatial data $X \in \mathbb{R}^{T\times H\times W}$, we first flatten the spatial dimensions into $N = H \cdot W$ using the *Morton order* (also known as the Z-order curve). Morton order is a locality-preserving mapping

$$\pi : \{0,\ldots,H-1\}\times\{0,\ldots,W-1\} \to \{0,\ldots,N-1\},$$

defined by interleaving the binary digits of the row i and column j coordinates:

$$\pi(i,j) = \sum_{k=0}^{\max(\lfloor \log_2 H \rfloor, \lfloor \log_2 W \rfloor)} \left(i_k 2^{2k} + j_k 2^{2k+1}\right),$$

where i_k, j_k denote the k-th bits of i and j. This ensures that points close in 2D remain close in the 1D representation, allowing MI to better reflect spatial correlations.

Using these arrays, we compute Temporal MI as:

$$I(I_T;X) = H(I_T) + H(X) - H(I_T,X) \tag{8}$$

and Spatial MI as:

$$I(I_N;X) = H(I_N) + H(X) - H(I_N,X) \tag{9}$$

MI provides a quantitative understanding of how the spatial-temporal trade-off parameter α influences the information content. For example, as $\alpha \to 0$, temporal MI dominates, and as $\alpha \to 1$, spatial MI dominates, as shown in figure 2 for the synthetic datasets.

As the raw MI values depend upon the scale of entropy, we normalize them for comparability using the average method as:

$$NMI(A;B) = \frac{2I(A;B)}{H(A) + H(B)} \tag{10}$$

IV. Experiments

We designed forecasting experiments to evaluate how well different models capture the spatial and temporal dependencies present in the synthetic datasets. The Synthetic 1D dataset was created with parameters $T = 1024, N = 1024, nc = 66$ (Figure 1b), and the Synthetic 2D dataset with parameters $T = 1024, H = 32, W = 32, nc = 66$ (Figure 1a). For consistency in experiments, both datasets are represented in the shape (T,N), with $T = 1024$ time steps and $N = 1024$ nodes. We use a context window of 24 time steps and a prediction horizon of 4 time steps, meaning the models observe the previous 24 time steps and forecast the next 4. The noise scale for all synthetic datasets is set to 0.1, i.e., for equations 2, 4, and 5, $\sigma = 0.1$.

To discretize the data for entropy calculation, we experimented with several values for the number of bins. Using too few bins can lead to underestimation of entropy due to oversmoothing, and using too many bins can inflate the entropy. Because the downstream analysis depends on normalized mutual information, which rescales entropy-based measures, the absolute value of entropy is less critical. We therefore selected a fixed, reasonable value of 128 bins, which adapts well to the data range while maintaining consistency across datasets.

A. Compared Models

We benchmarked a variety of deep learning forecasting models, including recent state-of-the-art architectures. To provide a point of reference, we also added a simple baseline model that copies the last observed time step in the context window as the prediction for all future steps. For models that are not designed to directly handle 2D spatial inputs, we apply Morton-order transformation to flatten the spatial dimensions into a 1D sequence while preserving locality.

- **Transformer:** The standard attention-based sequence model [2].
- **Autoformer:** Uses decomposition and auto-correlation mechanisms for long-term forecasting [13].
- **Flashformer:** Optimizes attention with efficient kernels for faster training [14].
- **Flowformer:** Models temporal dynamics with normalizing flows [15].
- **Informer:** Employs ProbSparse attention for efficient long-sequence forecasting [16].
- **Reformer:** Reduces Transformer complexity with locality-sensitive hashing [17].
- **iTransformer:** Incorporates instance-dependent tokenization for time series forecasting [5].
- **S-Mamba:** A selective state space model for time series, designed to capture long-range dependencies efficiently with lower memory and computation costs [4].
- **Copy Model(Baseline)**: Copies the last observed time step in the context window and repeats it for all prediction steps.

B. Findings on the synthetic benchmarks.

We evaluate representative time-series predictors on our synthetic data and summarize the main observations.

Trend with controllable structure. In the 1D setting (fig. 3), S-Mamba and iTransformer achieve the lowest MSE and are closely matched. As the control parameter α increases from 0 to 1—raising spatial predictability while lowering temporal predictability—MSE generally decreases across models. The same pattern holds in the 2D setting (fig. 4).

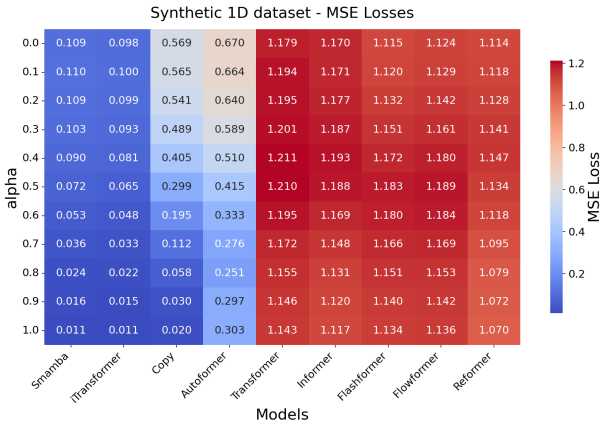


Fig. 3: MSE losses on synthetic 1D dataset. Context length $L = 24$, forecast horizon $T = 4$. Results are reported for $\alpha \in [0, 1]$ with step size 0.1.

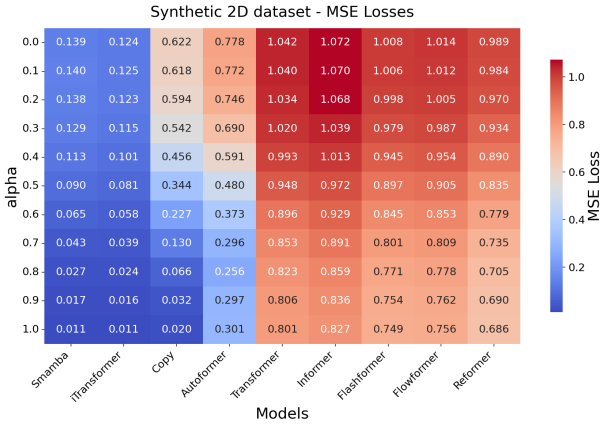


Fig. 4: MSE losses on synthetic 2D dataset. Context length $L = 24$, forecast horizon $T = 4$. Results are reported for $\alpha \in [0, 1]$ with step size 0.1.

Stratified performance by predictability ratio. fig. 5 relates model MSE to α and to the spatial–temporal predictability ratio $r = \frac{I(I_N;X)}{I(I_T;X)}$ computed at each α. Three strata emerge:

1) *Lowest MSE:* S-Mamba and iTransformer. These methods emphasize operations along the temporal axis (e.g., selective SSM scans and/or attention over time) rather than aggressively mixing variables at each step, which helps capture temporal regularities when present.
2) *Middle MSE:* Copy and Autoformer. Autoformer leverages autocorrelation structure [13], and both baselines rely on intermediate history, yielding similar MSE profiles as α varies.
3) *Highest MSE:* Standard Transformer variants (excluding iTransformer) that tokenize by time step and mix non-temporal features early; this tends to

TABLE I: Average MSE/MAE of models on real-world datasets, computed across all forecast horizons.

The lookback length is fixed to $L = 96$. Forecast horizons are $T \in \{12, 24, 48, 96\}$ for the PEMS datasets, and $T \in \{96, 192, 336, 720\}$ for Traffic, ETT, Electricity, Exchange, Weather, and Solar-Energy. Results are reported from [4].

Dataset	S-Mamba	iTransformer	RLinear	PatchTST	Crossformer	TiDE	TimesNet	DLinear	FEDformer	Autoformer
Traffic	**0.414 / 0.276**	0.428 / 0.282	0.626 / 0.378	0.481 / 0.304	0.550 / 0.304	0.760 / 0.473	0.620 / 0.336	0.625 / 0.383	0.610 / 0.376	0.628 / 0.379
PEMS03	0.122 / 0.228	**0.113 / 0.221**	0.495 / 0.472	0.180 / 0.291	0.169 / 0.281	0.326 / 0.419	0.147 / 0.248	0.278 / 0.375	0.213 / 0.327	0.667 / 0.601
PEMS04	**0.103 / 0.211**	0.111 / 0.221	0.526 / 0.491	0.195 / 0.307	0.209 / 0.314	0.353 / 0.437	0.129 / 0.241	0.295 / 0.388	0.231 / 0.337	0.610 / 0.590
PEMS07	**0.089 / 0.188**	0.101 / 0.204	0.504 / 0.478	0.193 / 0.303	0.235 / 0.315	0.308 / 0.425	0.140 / 0.225	0.329 / 0.395	0.165 / 0.283	0.367 / 0.451
PEMS08	0.148 / 0.224	**0.150 / 0.226**	0.529 / 0.487	0.280 / 0.321	0.268 / 0.307	0.441 / 0.464	0.193 / 0.271	0.379 / 0.416	0.286 / 0.358	0.814 / 0.659
ETTm1	0.398 / 0.405	0.407 / 0.410	0.414 / 0.407	**0.387 / 0.400**	0.513 / 0.496	0.419 / 0.419	0.400 / 0.406	0.403 / 0.407	0.448 / 0.452	0.588 / 0.517
ETTm2	0.288 / 0.332	0.288 / 0.332	0.286 / 0.327	**0.281 / 0.326**	0.757 / 0.610	0.358 / 0.404	0.291 / 0.333	0.350 / 0.401	0.305 / 0.349	0.327 / 0.371
ETTh1	0.455 / 0.450	0.454 / 0.447	0.446 / **0.434**	0.469 / 0.454	0.529 / 0.522	0.541 / 0.507	0.458 / 0.450	0.456 / 0.452	**0.440** / 0.460	0.496 / 0.487
ETTh2	0.381 / 0.405	0.383 / 0.407	**0.374 / 0.398**	0.387 / 0.407	0.942 / 0.684	0.611 / 0.550	0.437 / 0.449	0.450 / 0.459	0.450 / 0.467	0.427 / 0.507
Electricity	**0.170 / 0.265**	0.178 / 0.270	0.219 / 0.298	0.205 / 0.290	0.244 / 0.334	0.251 / 0.344	0.192 / 0.295	0.212 / 0.300	0.214 / 0.327	0.227 / 0.338
Exchange	0.367 / 0.408	0.360 / **0.403**	0.378 / 0.417	0.367 / 0.404	0.940 / 0.707	0.370 / 0.413	0.416 / 0.443	**0.354** / 0.414	0.519 / 0.429	0.613 / 0.539
Weather	**0.251 / 0.276**	0.258 / 0.278	0.272 / 0.291	0.259 / 0.281	0.259 / 0.315	0.271 / 0.320	0.259 / 0.287	0.265 / 0.317	0.309 / 0.360	0.338 / 0.382
Solar-Energy	0.240 / 0.273	**0.233 / 0.262**	0.369 / 0.356	0.270 / 0.307	0.641 / 0.639	0.347 / 0.417	0.301 / 0.319	0.330 / 0.401	0.291 / 0.381	0.885 / 0.711

blur the temporal signal and degrades forecasting accuracy on our benchmarks.

Effect of shifting signal from time to space. As spatial predictability increases (and temporal predictability decreases), the gap between group (1) and groups (2)/(3) narrows. Since group (1) models derive their advantage from stronger temporal-axis modeling, that advantage diminishes when the dataset contains less temporally predictable signal.

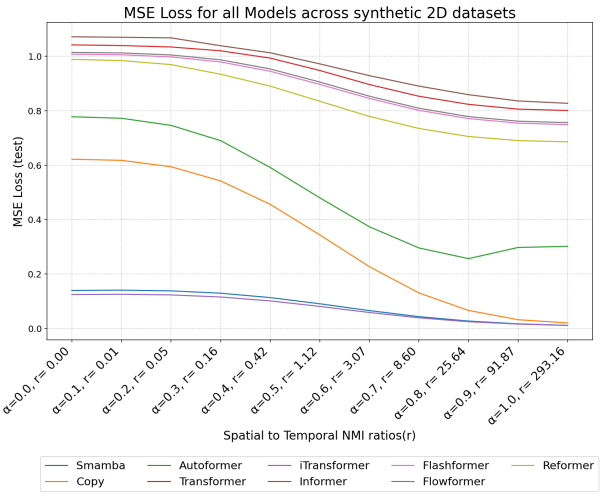


Fig. 5: MSE losses for synthesized datasets in increasing order of the spatial to temporal normalized mutual information ratios

C. Revisiting Real-World Benchmarks via an Information-Theoretic Lens

We apply our framework to the benchmark suites analyzed by [4]. Figure 2 reports the normalized temporal and spatial mutual information for these real-world datasets. In the original results, S-Mamba and iTransformer attain the best performance on most benchmarks; notable exceptions include Exchange-Rate, Weather, and the ETT family (ETTm1, ETTm2, ETTh1, ETTh2), where models such as PatchTST [18], CrossFormer [19], RLinear [20], and DLinear [21] have very close loss value or even perform better. MSE and MAE results are reported in table I, a consolidated summary of Tables 2–4 in [4].

Our measurements indicate that these exception datasets exhibit comparatively higher spatial mutual information (Figure 2). Because S-Mamba and iTransformer operate primarily along the temporal axis (without explicit variable/space mixing), they are less effective when the predictive signal is predominantly non-temporal. Conversely, on datasets with low spatial MI, S-Mamba and iTransformer outperform other models by a clear margin. These observations align with our synthetic-study findings and support the claim that model rankings shift systematically with the balance of temporal versus spatial predictability.

V. Conclusion

We introduced an information-theoretic framework that quantifies dataset structure via two indices, temporal and spatial predictability, and a synthetic spatio-temporal benchmark that independently controls both axes. Experiments across representative predictors reveal systematic shifts in model rankings as the balance of temporal vs. spatial signal varies: S-Mamba and iTransformer excel when temporal predictability dominates, but their advantage diminishes, and can reverse, on datasets with stronger spatial regularities. These findings help reconcile inconsistencies in prior evaluations and motivate benchmarks that report and vary data structure, as well as architectures that jointly model temporal dynamics and inter-variable (spatial) dependencies for high-accuracy forecasting.

VI. Future Work

This work is an initial step toward structure-aware evaluation of spatio-temporal predictors. We highlight two priorities for extension.

Beyond separating temporal and spatial mutual information (MI), the total entropy of the process also governs difficulty via the effective noise level. In our current generator, total entropy decreases as α increases, which can confound comparisons of MSE across α. We will introduce explicit entropy control, e.g., by calibrating additive noise or rescaling marginals, to keep total entropy approximately constant across settings, yielding more comparable error profiles.

Our study primarily compared time-series predictors; although we treated "spatial" as any non-temporal axis, a fuller assessment requires models that explicitly couple space and time (e.g., ST-GNNs, graph/patch Transformers, SSM–GNN hybrids). We will (i) evaluate such architectures under our indices, (ii) examine design choices for spatial discretization (grids vs. graphs vs. learned topology) and space–time fusion (axis-aligned vs. cross-axis interactions), and (iii) use insights from the synthetic studies to prototype a new spatio-temporal predictor aligned with dataset predictability profiles.

Acknowledgment

This research was supported in part by the College of Computing Seed Funding Program at Grand Valley State University. The authors gratefully acknowledge this support.

References

[1] J. A. Miller, M. Aldosari, F. Saeed, N. H. Barna, S. Rana, I. B. Arpinar, and N. Liu, "A survey of deep learning and foundation models for time series forecasting," *arXiv preprint arXiv:2401.13912*, 2024.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[3] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[4] Z. Wang, F. Kong, S. Feng, M. Wang, X. Yang, H. Zhao, D. Wang, and Y. Zhang, "Is mamba effective for time series forecasting?" *Neurocomputing*, vol. 619, p. 129178, 2025.

[5] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long, "itransformer: Inverted transformers are effective for time series forecasting," *arXiv preprint arXiv:2310.06625*, 2023.

[6] Y. Zhuang, M. Almeida, M. Morabito, and W. Ding, "Crime hot spot forecasting: A recurrent model with spatial and temporal information," in *2017 IEEE International Conference on Big Knowledge (ICBK)*. IEEE, 2017, pp. 143–150.

[7] X. Bresson and T. Laurent, "Residual gated graph convnets," *arXiv preprint arXiv:1711.07553*, 2017.

[8] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio *et al.*, "Graph attention networks," *stat*, vol. 1050, no. 20, pp. 10–48 550, 2017.

[9] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[10] C. Zheng, X. Fan, C. Wang, and J. Qi, "Gman: A graph multi-attention network for traffic prediction," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 01, 2020, pp. 1234–1241.

[11] L. Li, H. Wang, W. Zhang, and A. Coster, "Stg-mamba: Spatial-temporal graph learning via selective state space model," *arXiv preprint arXiv:2403.12418*, 2024.

[12] S. Müller, L. Schüler, A. Zech, and F. Heße, "GSTools v1.3: a toolbox for geostatistical modelling in python," *Geoscientific Model Development*, vol. 15, no. 7, pp. 3161–3182, 2022. [Online]. Available: https://gmd.copernicus.org/articles/15/3161/2022/

[13] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," 2022. [Online]. Available: https://arxiv.org/abs/2106.13008

[14] T. Dao, D. Fu, S. Ermon, A. Rudra, and C. Ré, "Flashattention: Fast and memory-efficient exact attention with io-awareness," *Advances in neural information processing systems*, vol. 35, pp. 16 344–16 359, 2022.

[15] Z. Huang, X. Shi, C. Zhang, Q. Wang, K. C. Cheung, H. Qin, J. Dai, and H. Li, "Flowformer: A transformer architecture for optical flow," 2022. [Online]. Available: https://arxiv.org/abs/2203.16194

[16] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," 2021. [Online]. Available: https://arxiv.org/abs/2012.07436

[17] N. Kitaev, Łukasz Kaiser, and A. Levskaya, "Reformer: The efficient transformer," 2020. [Online]. Available: https://arxiv.org/abs/2001.04451

[18] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," *arXiv preprint arXiv:2211.14730*, 2022.

[19] Y. Zhang and J. Yan, "Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting," in *The eleventh international conference on learning representations*, 2023.

[20] Z. Li, S. Qi, Y. Li, and Z. Xu, "Revisiting long-term time series forecasting: An investigation on linear mapping," *arXiv preprint arXiv:2305.10721*, 2023.

[21] A. Zeng, M. Chen, L. Zhang, and Q. Xu, "Are transformers effective for time series forecasting?" in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 9, 2023, pp. 11 121–11 128.