Maintaining the status quo: Capturing invariant relations for OOD spatiotemporal learning

Zhengyang Zhou[1,2], Qihe Huang[1,3], Kuo Yang[1], Kun Wang[1], Xu Wang[1],Yudong Zhang[1]
Yuxuan Liang[5], Yang Wang[1,2,3,4†]
[1] University of Science and Technology of China (USTC), Hefei, China.
[2] Suzhou Institute for Advanced Research, USTC. [3] School of Software Engineering, USTC
[4] Key Laboratory of Precision and Intelligent Chemistry, USTC.
[5]Hong Kong University of Science and Technology, Guangzhou, China
{zzy0929,hqh,yangkuo,wk520529,wx309,zyd2020}@mail.ustc.edu.cn,yuxliang@outlook.com,angyan@ustc.edu.cn

ABSTRACT

Spatiotemporal (ST) learning has become a crucial technique for urban digitalization. Due to expansions and dynamics of cities, current spatiotemporal models are inclined to suffer distribution shifts between training and testing sets, leading to the OOD delimma of ST learning. However, very few studies focus on such OOD problem of temporal regressions, let alone spatiotemporal learning. Spatiotemporal data usually reveals segment-level heterogeneity within periodicity and complex spatial dependencies, posing challenges to invariance extraction. In this paper, we find that ST relations make sense for generalization and devise a causal ST learning framework, CauSTG, which enables invariant relation transferred to OOD scenarios. Specifically, we take temporal steps as environments, and transform spatial-temporal relations into learnable parameters. To tackle heterogeneity in periodicity, we partition temporal steps into sub-environments by identifying distinctive trend patterns, enabling re-organized samples trained separately. To extract invariance within ST observations, we propose a spatiotemporal consistency learner and a hierarchical invariance explorer to jointly filter out stable relations. Our spatiotemporal learner quantifies bi-directional spatial consistency and extracts disentangled seasonal-trend patterns via relations reflected by trainable parameters. Further, the hierarchical invariance explorer constructs variation-based filter to achieve both local and global invariances. Experiments on three OOD scenarios reveal that CauSTG can increase at most 10.26% performance against best baselines, and visualized invariant relations can well interpret the physical rationales.

CCS CONCEPTS

• **Information systems** → **Spatial-temporal systems**; **Data mining**; • **Computing methodologies** → *Knowledge representation and reasoning*; *Artificial intelligence*.

†Prof. Yang Wang is the corresponding author.

KDD'23, August 6-10, 2023, Long Beach, U.S.

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

KEYWORDS

Spatiotemproal data mining, causal perspective, invariant learning, dynamic graph

ACM Reference Format:
Zhengyang Zhou[1,2], Qihe Huang[1,3], Kuo Yang[1], Kun Wang[1], Xu Wang[1],Yudong Zhang[1], Yuxuan Liang[5], Yang Wang[1,2,3,4†]. 2023. Maintaining the status quo: Capturing invariant relations for OOD spatiotemporal learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'23), August 6-10, 2023, Long Beach, U.S.*. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Nowadays, many researches reveal the vulnerability of machine learning-based models when they are exposed to data with different distributions, induing the necessity of inference on out-of-distribution (OOD) instances [33, 38]. To this end, surge of literature generalizes models to OOD scenarios via capturing invariances between features and targeted labels. Despite their prosperity, most of them focus on addressing covariate shifts of images [22] and static graphs [19, 33, 35], but few of them focus on spatiotemporal data (ST data). The ST data is an emerging data structure with dynamic observations in both spatial and temporal domains, where it can accommodate diverse urban applications, such as traffics [29, 47], smart grids [34] and air quality [8, 39].

In learning tasks, OOD scenarios refer to the existence of distribution shifts between training and testing sets. Actually, compared with static images or graphs, distributions of ST data are more inclined to change over time due to increasing urban populations, urban constructions and seasonal factors [5, 9]. Thus, covariate distribution shift becomes a core obstacle to OOD generalization. The early practice of ST learning transfers interval-level data into image-like grids and formulates element-wise regression [43]. Recent works devise dynamic graph-based architectures to capture spatiotemporal relations, where each sensing point is seen as node and the inter-sensor correlations are described as edges [2, 3, 40]. Even so, they still ignore the critical issue of covariate shift. On the other hand, OOD learning usually assumes the existence of virtual environments within observations and the distributions of variates are varying across environments. To this end, invariant learning captures the invariance across environments via minimizing prediction error variances, namely IRM objective [1].

Following above studies, we classify OOD works into graph-based and series-based. For graphs, the idea of invariant learning has been adopted in both graph classification [4, 18, 19, 35] and

node classification [33] where local graph topology is considered as environments. For series learning, AdaRNN first defines covariate shifts in time series and exploits distribution characterization to realize weighted prediction [9]. In addition, CoST [31] interprets the disentangled seasonal knowledge and trend information from causal perspective to enhance anti-noise robustness. Nevertheless, given inherent dynamic and heterogeneous observations, existing solutions are still incapable of dealing with OOD tasks on ST data. Concretely, graph-based causal learning takes local topology as environments but fails to model the invariance of temporal evolution, while series-based OOD learning cannot capture the invariance with spatial dependencies. Moreover, IRM fails to interpret which relation is exactly invariant for generalization, and traditional IRM requires heavy modification to adapt regression tasks. Actually, in causal spatiotemporal learning, temporal steps should be naturally considered as environments, thus the essence of its OOD learning is to find invariant spatiotemporal relations across temporal environments. However, we argue that spatiotemporal OOD learning is more challenging than classification tasks due to two critical issues.

Segment-level heterogeneity entangles temporal environments. In Fig. 1(a), each sequence segment in the periodicity reveals various patterns, which induces diversity of node-wise correlation patterns. Since we aim to capture invariant mappings among observations, such heterogeneity complicates the identification of distinctive segments and exacerbates the sparsity of common invariance across steps. Thus, *how to construct temporal environment partition with segment heterogeneity* is the first challenge.

IRM fails on heterogeneous spatiotemporal observations. IRM learns invariant representations via minimizing deviated risks from expected ones across diverse environments. Instead of using logistic classification risks, IRM tends to bring in unfairness for learning relations on different nodes due to heterogeneous observations. Therefore, accompany with lacked interpretation, *how to devise an interpretable invariance explorer that eliminates effects of observation intensities* becomes the second challenge.

Fortunately, we discover that quantified spatiotemporal relations can determine predictive observations. Generalizations on classification usually capture invariant substructures or feature representations [4, 18, 19, 35] for transferring, but in ST learning, it is the amount of transitions between nodes that matters. As illustrated in Fig. 1(b), ST-based invariant learning cares more about consistent quantified relations across steps, where stable spatial relations can be interpreted as consistent direction of temporal evolution between nodes, while invariant temporal relation can be seen as consistent seasonality and fluctuations. Therefore, we argue that capturing invariant relations within ST data is of great importance, and *how to design a scalable spatial-temporal learner for easily discovering quantified relational invariance* remains an open problem to the community.

In this paper, we propose a Causal Spatiotemporal Graph Learning framework (CauSTG) to discover invariance in spatiotemporal data for OOD generalization. First, *to tackle segment-level temporal heterogeneity*, we partition periodicity into multiple segments as sub-environments by progressively identifying distinctive segments. Then re-organized samples can be trained respectively to capture both segment-level local and global invariance for avoiding invariance sparsity. Second, *to cooperatively remedy IRM failure*, we design a spatiotemporal consistency learner and a hierarchical invariance explorer to transform the spatial-temporal correlation into learnable parameters, and capture the invariance that eliminates heterogeneous risk effects on relation extraction. In particular, the ST consistency learner is specifically designed for relational quantification and easier invariance extraction. As motivated in Fig. 1(b), we propose a bi-directional spatial learner by factorizing spatial relations into negative and positive causal correlations, while devise a decomposed temporal pattern extractor for seasonality and trend abstraction, where all these relations are extracted by learnable parameters. Next, our hierarchical invariance explorer, which aims to filter and ensemble stable relations, separately trains diverse models within and across sub-environments to encourage parameter diversity. Our invariance explorer is instantiated with a proposed MinVar Pooling module that highlights stable weights across environments, and averages selected stable weights across models. We embed MinVar within and across environments, and design a stability-based ensemble to hierarchically integrate segment-level local invariance and global invariance. The fine-tuning is imposed to adapt the parameters to new neural structures with unstable relation removed, which facilitates better OOD generalization.

Contributions. (1) We investigate OOD generalization on spatiotemporal data from an invariance perspective, which enables causal spatial-temporal relations reflected by learnable parameters. (2) To facilitate the invariance extraction, we partition periodicity into temporal sub-environments to enable hierarchical invariance preservation, and embed disentangled relations into trainable parameters for stable relation filtering. We theoretically justify our invariance explorer via deriving a smaller approximated error. (3) We design three OOD scenarios on ST data for experiments. Our CauSTG outperforms non-invariance learning baselines by 0.9%∼10.26% while beats invariant learning by 0.5%∼7.35%.

2 PRELIMINARIES

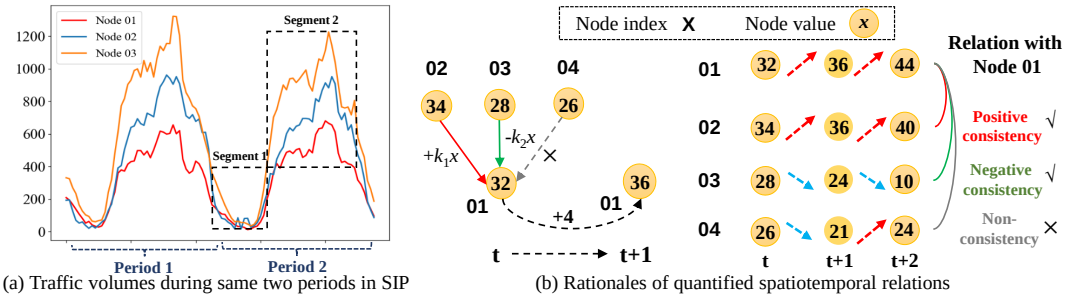
2.1 Problem formulation

Let $\mathbb{G} = \{\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_t, ..., \mathcal{G}_T\}$ be a sequence of dynamic graphs with T steps. Each $\mathcal{G}_t$, is described as $\{\mathcal{V}, \boldsymbol{X}_t, \mathcal{E}\}$ where $\mathcal{V} = \{v_1, v_2, ..., v_N\}$ is the node set, $\boldsymbol{X}_t \in \mathbb{R}^{N \times F}$ is the deterministic observation of $\mathcal{G}_t$, $\mathcal{E}$ describes the graph structure. The spatiotemporal prediction model predicts following τ steps by exploiting previous κ steps, i.e., $\widehat{\boldsymbol{y}} = f(\boldsymbol{x})$ where $(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{X}_{t-\kappa:t}, \boldsymbol{X}_{t+1:t+\tau})$. Given training and testing data $\mathbf{P}_{train}$, $\mathbf{P}_{test}$, OOD scenarios indicate that conditional distributions are identical but marginal probability distributions change. In our task, given the training sequence $\mathbb{G}_t$ and testing sequence $\mathbb{G}_s$, where $\mathbf{P}_{\mathbb{G}_t}(\mathbf{x}, \mathbf{y}) \neq \mathbf{P}_{\mathbb{G}_s}(\mathbf{x}, \mathbf{y})$ but $\mathbf{P}_{\mathbb{G}_t}(\mathbf{x}|\mathbf{y}) = \mathbf{P}_{\mathbb{G}_s}(\mathbf{x}|\mathbf{y})$, the goal of OOD generalization is to refine an invariant relation mapping f^* from $\mathbb{G}_t$, achieving risk minimization on test set $\mathbb{G}_s$,

$$\min_{(\boldsymbol{x},\boldsymbol{y})\in\mathbb{G}_s} \mathcal{R}(\widehat{\boldsymbol{y}}; f^*(\boldsymbol{x})) \tag{1}$$

2.2 Theoretical guarantee

We first illustrate the failure of non-invariance learning on OOD regression, and further dissect an invariant relation-aware solution to alleviating the prediction deviation. In particular, we extend assumptions in static graphs [33] to dynamic spatiotemporal graphs to facilitate our analysis.



(a) Traffic volumes during same two periods in SIP (b) Rationales of quantified spatiotemporal relations

Figure 1: Subfigure (a) conveys two messages. (a.1) Spatiotemporal heterogeneity. Volumes in one region can be with large dispersion across temporal steps while different nodes have heterogeneous intensities. (a.2) Segment-level heterogeneity. Each period can be decomposed into different segments where each segment reveals a specific pattern. Subfigure (b) delivers that, if two nodes are consistently with positive or negative evolution trends across steps, these two nodes are considered with a causal relation that can be transferred to unseen environments.

Assumption 1 (**Environment variation**). *Given sequential spatiotemporal observations $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_N$, we suppose the distributions $\mathbf{P}(\boldsymbol{X})$ are dependent on the virtual environments, and there are total K environments, i.e., $\mathbb{E} = \{e_1, e_2, ..., e_K\}$. In other words, the distribution shift is induced by the changes of virtual environments.*

Assumption 2 (**Invariance property**). *Even though the environments and covariate distributions are changing over time, there must exist some invariant relations, which are induced by either positive or negative physically causal correlations, i.e., $\forall(i,j) \in |\mathcal{V}|$, $\exists(p,q), s.t.\ \mathbf{P}(x_p, x_q|e_i) = \mathbf{P}(x_p, x_q|e_j)$.*

In the dynamic graph regression, given node v_i, let degree of v_i, and proportions of neighbors with causally invarint relations to v_i denote as d_i, p_i with $d_i > 1, 0 < p_i < 1$. Considering the covariate shifts, i.e., we train the model on samples following Gaussian distribution $\mathbb{G}_s \sim N(\mu_0, \sigma_0|e_0)$ and test on samples following $\mathbb{G}_t \sim N(\mu_q, \sigma_q|e_q)$. We have the following two propositions.

Proposition 1. *Let the graph-based ST learner f trained without considering invariant relations, then the upper bound of empirical risk under environment e_0 would be $\varepsilon_0 \sim \frac{2(1-p_i)d_i\mu_0 w_i^s}{1+d_i}$ that is irreducible, where w_i^s is the weight for causal neighbor aggregation. When f is transferred to OOD test set $N(\mu_q, \sigma_q|e_q)$ satisfying $\mu_q = q\mu_0$ where $q \in \mathbb{N}^+$. The OOD risks are amplified to $\varepsilon_q \sim \frac{2(1-p_i)d_i q\mu_0(\mu_w \pm 3\sigma_{ws})}{1+d_i}$, where μ_w and σ_{ws} are the expectation and variance of learnable w_s.*

Remark. Proposition 1 manifests that the error bounds of non-invariant learning are concerning with both observation expectation and relation variance. Since μ tends to be overwhelmingly larger than σ, the error upper bound will be approximately amplified by q times when inferring on e_q, leading to unacceptable generalization performances. Therefore, we can conclude the failure of non-invariance learning on OOD regressions.

Proposition 2. *Let the ST learner f^* be trained considering invariant relations, then the empirical risk under any environment $e_i (i = 1, ..., K)$ can asymptotically converge to 0 with $w_i^c = \frac{1}{p_i}$ where w_i^c is the weight for causal part neighbor aggregations.*

Remark. Proposition 2 delivers that capturing the invariant neighboring relations with spurious neighborhood eliminated can enable the independence between errors and the expectation of original observations, thus providing guarantee on performances of OOD scenarios. The complete proof of above two propositions can be found in Sec. A.1 and A.2.

3 METHODS

3.1 Framework Overview

Inspired by above analysis, CauSTG is proposed to model invariant relations in dynamic graphs across temporal environments. Specifically, CauSTG consists of three modules, a temporal environment partition that divides the periodicity into different sequence segments for hierarchical sample organization, the spatiotemporal consistency learner for easy relation extraction, and the hierarchical invariance explorer that filters and ensembles hierarchical local-global invariance. The overview of CauSTG is in Fig. 2.

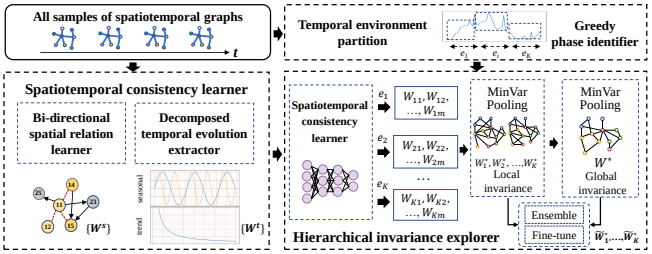


Figure 2: Framework overview of CauSTG. The W_{ij} refers to individual well-learned model weights within sub-environments, W^* indicates ensembled model weights.

3.2 Temporal environment partition

Given the inherent time-varying property in ST data, we naturally take time steps as the virtual environments, which essentially reflect various contexts like weather and daily tidal regularity. However, summarizing the step-level common invariance across all steps definitely leads to extreme sparsity of overall invariant relations. Therefore, achieving the tradeoff of temporal partition is the premise of spatiotemporal invariant learning.

In this work, we explore the invariance across time steps by further partitioning the periodicity into different segments where the time steps in each segment can be formed as samples within the same sub-environment. To facilitate data processing, we consider the same segment index in different periods as the same sub-environments so as to re-organize samples to separately train different models according to temporal partitions. Given spatiotemporal observations and the expected K sub-environments, we can formalize our temporal environment partition by maximizing the discrepancy distances between pair of segment partition,

$$\max \sum_{1 \leqslant p \neq q \leqslant K} d\left(\mathbf{D}_p, \mathbf{D}_q\right), s.t. \forall i, \Delta_1 < |\mathbf{D}_i| < \Delta_2; \sum_{i=1}^{K} |\mathbf{D}_i| = N \quad (2)$$

where $\mathbf{D}_p, \mathbf{D}_q$ are the p-th and q-th segments of partitioned observation, Δ_1 and Δ_2 are set as the minimal and maximal steps in each segment to avoid the trivial solution. However, directly optimizing Eq. 2 for K partitions with ST data is NP-hard. Inspired by the series-splitting solution in AdaRNN [9], we propose our spatiotemporal temporal environment partition. Following [9], we evenly construct D parts for each period and each part will be the minimal unit that cannot be further split. We then progressively search K in $\{4, 5, 6, 8\}$ to obtain the globally optimized partition points. The distinctions between our solution and [9] are two-fold. First, we consider the spatial nodes in each step and take the spatiotemporal observations as a whole to calculate the variations, where $\mathbf{D}_p \in \mathbb{R}^{N \times T_p}$ and T_p is the number of time steps in segment p. Second, to better identify invariant relations in each segment, we modify the discrepancy metric into the absolute value of cosine similarity, where higher absolute value of such similarity indicates more consistent evolution direction encapsulating both positive and negative correlations,

$$d\left(D_p, D_q\right) = \frac{1}{N} \sum_{i=1}^{N} |\cos(D_{p_i}, D_{q_i})| \quad (3)$$

We designate the Δ_1 as 2τ steps, covering two times of training sample periods, and Δ_2 is $\frac{\tau}{2}$ steps in a periodicity, for easy implementation. Based on the greedy strategy, we can obtain the globally optimized K temporal sub-environments. Formally, the p-th segment in periodicity naturally contributes to p-th sub-environment, which starts from step index p_s and ends at p_e, i.e., $e_i = \{p_s^i, ..., p_e^i\}$ and $\sum_{i=1}^{K} |p_e^i - p_s^i + 1| = T$. Therefore, we can hierarchically explore both segment-aware local invariance and global invariance under sub-environment partitions, and avoid the extreme sparse invariant connections by preserving such multi-level invariance.

3.3 Spatiotemporal consistency learner

Our spatiotemporal consistency learner is designed with two intuitions to adapt invariant relation extractions. (1) Multiple placeholders of learnable relations and representation decoupling [31] can encourage more opportunity to capture major common invariance across local steps. (2) Embedding relations into learnable parameters on both spatial and temporal domains allows relational quantification and enables invariance extraction via capturing parameter variations. We will elaborate our ST consistency learner on respective spatial and temporal aspects.

3.3.1 Bi-directional spatial relation learner. In spatial domain, we design a learning kernel that can better capture causal spatial correlations according to two observations. First, in ST elements, spatial transitions usually contribute to temporal evolution, thus the spatial relations can be measured by the consistency of temporal evolution directions. Second, spatial causal correlations can be attributed to physical structures, e.g., human daily routines and local functionality, and they can be either positive or negative. Therefore, we propose a spatial consistency measurement by considering temporal evolution and bi-directional relations, and further dissect how this measurement derives a scalable spatial learning kernel. Specifically, this measurement characterizes the consistency of variation directions in a node-wise manner, and normalizes the quantified difference via the smoothed average of previous κ steps for stability. Take steps from t to $t+1$ as an example, the spatial relation $r^t(v_i, v_j)$ between node i and j is formulated by,

$$r^t(v_i, v_j) = \frac{x_i^{t+1} - \overline{x}_i^t}{x_j^{t+1} - \overline{x}_j^t} \frac{\overline{x}_j^t}{\overline{x}_i^t} \quad (4)$$

where $\overline{x}_i^t$ is the averaged value of previous $(t : t-\kappa)$ steps, the sign of $r^t(v_i, v_j)$ indicates the correlated direction of two nodes while the ratio value quantifies the intensity of discrepancy. Then we can derive the predicted x_i^{t+1} with $r^t(v_i, v_j)$ (simplified as r_{ij}^t) by,

$$x_i^{t+1} = \overline{x}_i^t + \frac{r_{ij}^t (x_j^{t+1} - \overline{x}_j^t) \overline{x}_i^t}{\overline{x}_j^t} \quad (5)$$

As observed, Eq. (5) provides an opportunity to achieve x_i^{t+1} by regressing the relations. However, the second term in the numerator of Eq. (5) is over-complex and seems intractable due to unavailability of x_j^{t+1}. To enable scalable relation modeling and preserve such relational complexity, we impose a polynomial kernel function $g(v_i, v_j)$ on x_j^t and x_i^t to substitute the second term within numerator for describing their directional relationship. Concretely, we omit x_i for the existence of first term, and formulate the quadratic function as $g(v_i, v_j) = k_0 x_i \overline{x}_j + k_1 \overline{x}_j{}^2 + k_2$, where k_0, k_1 are learnable coefficients and k_2 is the bias for relational regression. Dividing $g(v_i, v_j)$ by $\overline{x}_j$ and merge with x_i, we can exploit the simplified but still scalable relation to obtain x_i^{t+1},

$$x_i^{t+1} = k_0 x_i^t + k_1 \overline{x}_j^t + \frac{k_2}{\overline{x}_j^t} \quad (6)$$

Actually, this aggregation function for v_i enjoys the nice property of capturing the bi-directions of both positive and negative correlations. When positive correlation dominates, we can learn a larger k_1 but smaller k_2 (near zero), and vice versa. From the invariance learning aspect, the stability of coefficient k_1 can be interpreted as a positively spatial consistency between x_i and x_j, i.e., they have the same variation direction at near steps, while a stable k_2 can interpret a negatively consistency where such consistency can be reliably transferred for extrapolation. This operation ensures the learning process to find the disentangled correlation patterns and increases the possibility of capturing causally invariant relations. Further, we take Eq. (6) into graph learning perspective, where the liner term of x_i can be viewed as the self loop while the latter two terms can be seen as the linear combinations of its neighborhood and corresponding inverses. Thus, the expectation of next step

prediction for x_i^{t+1} can be written as,

$$E(h_i^{t+1}) = k_0 x_i + \sum_{\substack{v_j \in N(v_i), \\ s \in [1, |N(v_i)|]}} k_{s1} x_j + \frac{k_{s2}}{x_j} \tag{7}$$

In Eq. (7), $N(v_i)$ denotes the neighbor set of v_i, and k_0, k_{s1}, k_{s2} are corresponding learnable weights to model relations. Such scalable parameters extend binary adjacency to vector-level relation learning thus facilitating invariance extraction. For implementation, we formulate our learning scheme into a new graph learning framework, by transferring k_1 and k_2 into A_0 and A_{it},

$$\mathbf{h}^{t+1} = (\mathbf{A}_0\mathbf{X} + \mathbf{A}_{in}\mathbf{X}^{-1})\mathbf{U} \tag{8}$$

where $\mathbf{A}_0, \mathbf{A}_{in} \in \mathbb{R}^{N\times N}$ and $\mathbf{U} \in \mathbb{R}^{d_0\times d_m}$ are three learnable parameters for spatial learning, d_0 and d_m are input and output dimensions of node-level features. $\mathbf{A}_0$ naturally encapsulates the self loop. We denote the output feature maps and the learnable weight set as $\boldsymbol{X}^s \in \mathbb{R}^{N\times d_m}$ and $\boldsymbol{W}^s = \{\mathbf{A}_0, \mathbf{A}_{in}, \mathbf{U}\}$, respectively.

3.3.2 Decomposed temporal pattern extractor. Since causal learning calls for disentangled representation to counteract interventions, we design our temporal pattern extractor from two aspects, i.e., seasonal branch and trend branch [31] that respectively accommodate periodicity and evolution trends. To ensure independence and better extract factor-level invariance, these two branches will be learned independently by separated objectives and parameters.

Seasonal branch. We propose a multi-scale temporal convolution for pattern extractions where the sizes of convolution kernels are derived by Fast Fourier Transformer (FFT) [25]. In particular, we first exploit FFT to find top-l important frequency in the input sequence, i.e., $\{f_1, f_2, ..., f_l\} = \underset{max-l}{\arg\max} \text{FFT}(\boldsymbol{X})$ to identify the underlying seasonality. Then the inverses of frequency are considered as the seasonality and can be exploited as kernel sizes of temporal convolutions. Thus, the temporal convolution kernel falls into $w_{ts}^k \in \mathbb{R}^{d_{c_k}\times 1}$, where $d_{c_k} = 1/f_k$ $(k = 1, 2, ..., l)$. Then we can obtain the seasonal representation of $\widehat{\boldsymbol{Y}}_s \in \mathbb{R}^{N\times\tau}$ by,

$$\widehat{\boldsymbol{Y}}_s^{t+1:t+\tau} = \text{TCN}(\boldsymbol{X}^s; w_{ts}^1, ..., w_{ts}^l) \tag{9}$$

where w_{ts} can be viewed as periodicity-based amplitude for periodical pattern extraction. To capture the refined seasonal information, we exploit the extracted $\{d_{c_k}\}$ to recover the seasonal information from frequency domain to temporal domain $\widetilde{\boldsymbol{Y}}_s$, and minimize the difference between the predicted seasonal sequence $\widehat{\boldsymbol{Y}}_s$ and the reconstructed sequence $\widetilde{\boldsymbol{Y}}_s$ from the derived kernels. The seasonal learning objective can be denoted as,

$$Loss_{se} = \text{MAPE}(\widehat{\boldsymbol{Y}}_s, \widetilde{\boldsymbol{Y}}_s) \tag{10}$$

Noted that the sizes of kernels are calculated once and share across the same segment index in each period, and the reconstructed parameters will not count into our pattern extractor.

Trend branch. To encapsulate the trend learning capacity, we first impose learnable transformations $w_{tr} \in \mathbb{R}^{d_m\times\tau}$ on the input sequence and devise two trend-preserved objectives. Specifically, the result $\widehat{Y}_{tr} \in \mathbb{R}^{N\times\tau}$ after transformation on original sequence is,

$$\widehat{Y}_{tr} = \boldsymbol{X}^s * w_{tr} \tag{11}$$

After that, we calculate the first-order difference to characterize the evolving trend pattern, which is derived by $\Delta y^t(i) = y_i^{t+1} - y_i^t$. And then we devise two objectives to jointly preserve the sequence consistency via minimizing the values of cosine similarities respectively between targeted sequences, and their first-order differences,

$$Loss_{tr} = \min \cos(Y, \widehat{Y}_{tr}) + \cos(\Delta Y_{tr}^t, \Delta\widehat{Y}_{tr}^t) \tag{12}$$

Therefore, parameters w_{tr} account for capturing evolution trend transformation from inputs to outputs. To ensure the independence between these two representations, we impose an independence regularization through maximizing the cosine similarity between the two perspective representations, i.e., $\text{Reg}_{in} = -\min\cos(\widehat{Y}_s, \widehat{Y}_{tr})$.

Overall objective of temporal pattern extractor. We can obtain overall prediction representation by element-wisely fuse these two aspects, i.e., $\widehat{Y} = \widehat{Y}_{tr} \oplus \widehat{Y}_s$, and achieve the final objective,

$$Loss = \text{MAPE}(\widehat{Y}, Y) + \lambda_0 Loss_{se} + \lambda_1 Loss_{tr} + \lambda_2 \text{Reg}_{in} \tag{13}$$

where $\lambda_0, \lambda_1, \lambda_2$ are hyperparameters. Therefore, we can denote the set of temporal learning weight $\mathbf{W}^t = \{w_{ts}, w_{tr}\}$.

Finally, the combined parameters of our spatiotemporal consistency learner can be written as $\mathbf{W} = \text{Concat}\{\mathbf{W}^s, \mathbf{W}^t\} \in \mathbb{R}^{P\times Q}$ where these two weight sets capture the relations within input observations and relations between input and output. P and Q are two virtual dimensions of the learnable weights to facilitate the description of our model.

3.4 Hierarchical invariance explorer

Previous invariant learning works usually minimize the variance of risks derived from different environments [1, 33, 35], however, we argue that these solutions are not suitable for our regression tasks. The reason is that risks of classification are logistically homogeneous while the risks of ST regression are heterogeneous across topology and temporal environments. Thus, minimizing the variation of different risks must be biased. To this end, a more intuitive way of capturing invariance is to filter out stable trainable weights, which reflects invariant mapping relations, across environments.

Instead of imposing an IRM objective, we propose our hierarchical invariance explorer to explicitly identify stable spatiotemporal relations, which is elaborated in Fig. 3. Specifically, our invariance explorer is a training framemwork with a novel weight selection strategy. By cooperatively working with another two modules, our hierarchical invariance explorer generates diverse models within and across sub-environments, thus enabling parameter diversity and hierarchically capturing both local and global invariances. The hierarchical design of capturing two-level invariances is devised to avoid extreme sparsity of global invariant connections.

Concretely, with partitioned sub-environments, the generated samples and trained models are both hierarchically. We first organize different sample groups within each partitioned sub-environment and train a series of models for each sample group and environment. We denote weights in each well-trained models as $\Theta = \{\{\mathbf{W}_{11},...,\mathbf{W}_{1m}\}, ..., \{\mathbf{W}_{K1},...,\mathbf{W}_{Km}\}\}$, where $\mathbf{W}_{ij}$ represents the j-th grouped model within i-th sub-environment. Our invariance explorer receives well-learned model weights from differernt groups. To capture invariance with spatiotemporal observations, we propose a novel stable weight selection strategy named MinVar pooling.

Our MinVar pooling consists of two stages, a minimal variance-based stability filter to highlight relatively stable weights, and an average pooling to ensemble models for OOD inference.

Local invariance. Given the k-th sub-environment and the environment specific weight $\Theta(k)$, we first filter out the interpretable stable weights to achieve the locally invariant spatiotemporal correlations in this sub-environment. Denote $w_{kj}(p,q)$ as the element indexed by (p,q) in learned model weight $\mathbf{W}_{kj}$, we can measure the local stability by computing the element-wise variance, and then highlight the most $r\%$ stable weight indexes as invariant ST relations. Formally, the indexes of invariant weights (p,q) can be derived by,

$$(p,q) = \underset{\text{Min-r\%}}{\arg\min}(\underset{1\leqslant j\leqslant m}{\text{Var}}(\{w_{kj}(p,q)\})) \tag{14}$$

where $1 \leqslant p \leqslant P,\ 1 \leqslant q \leqslant Q$. Then only the selected entries will participate in following stage while other entries will be eliminated due to their unstability. After that, the second stage of average pooling calculates the average of weights at selected entries to obtain the local invariance $\mathbf{W}^*_k$,

$$w^*_k(p,q) = \underset{1\leqslant j\leqslant m}{\text{Avg}}(\{w_{kj}(p,q)\}) \tag{15}$$

$w^*_k(p,q)$ is the element in $\mathbf{W}^*_k$. Similarly, we achieve a series of local invariant models $\mathbf{W}^*_k (k = 1, 2, ..., K)$ for each sub-environment.

Global invariance. To achieve the global invariance, we impose our MinVar Pooling on all local invariant models and obtain the overall invariance model $\mathbf{W}^*$ by elimination and ensemble. With both local invariance $\mathbf{W}^*_k$ and overall invariance $\mathbf{W}^*$, the following question becomes into how to exploit the sub-environment models to achieve an ensembled global OOD model by fusing local and overall invariance. To resolve this challenge, we devise a stability-based ensemble to integrate local and global invariance via calculating the element-wise discrepancy. To be specific, the underlying discrepancy is traced back to its upper-level models. For instance, the potential stability of k-th local invariance $w^*_k(p,q)$ is referred to model variations within sub-environment k, i.e., $\underset{1\leqslant j\leqslant m}{\text{var}}(\{w_{kj}(p,q)\})$, while the stability of global invariance $w^*(p,q)$ is referred to the model variations across different sub-environment invariances, i.e., $\underset{1\leqslant k\leqslant K}{\text{var}}\ w^*_k(p,q)$. As the variance can be the inverse of stability, we take the inverse proportion of their stability (empirical variances) as ensemble coefficients. Formally, the ensembled global invariant ST learning weights for k-th sub-environment $\widetilde{\mathbf{W}}^*_k = \{\widetilde{w}^*(p,q)\}$ can be achieved by,

$$\begin{aligned}\widetilde{w}^*(p,q) &= \gamma_{(p,q)} w^*_k(p,q) + (1-\gamma_{(p,q)}) w^*(p,q)\\ \gamma_{(p,q)} &= \frac{\underset{1\leqslant k\leqslant K}{\text{var}}\ w^*_k(p,q)}{\underset{1\leqslant j\leqslant m}{\text{var}}(\{w_{kj}(p,q)\}) + \underset{1\leqslant k\leqslant K}{\text{var}}\ w^*_k(p,q)}\end{aligned} \tag{16}$$

Therefore, all positions of non-zero parameters in $\widetilde{\mathbf{W}}^*_k$ constitute the invariant relations in training set $\mathbb{G}_t$, and we impose a fine-tune strategy on it to update $\widetilde{\mathbf{W}}^*_k$ [26] and achieve the final f^*, where the new $\widetilde{\mathbf{W}}^*_k$ can better adapt to new pruned neural architectures.

Our hierarchical invariance explorer can be deemed as a novel training strategy with sample re-organization and model fusion. Given above, our global invariance model $\widetilde{\mathbf{W}}^*_k$ can preserve two-level parameter-level invariances and their fusion coefficients. When an OOD sample arrives, we first assign it to corresponding environment and exploit Eq. (16) to obtain the predictions.

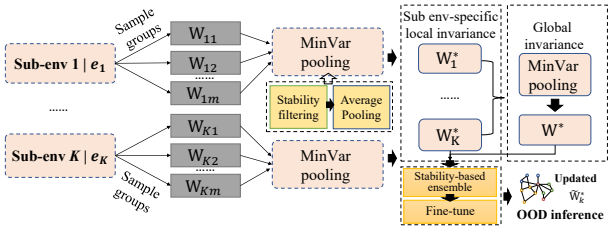


Figure 3: Overview of hierarchical invariance explorer

4 EXPERIMENT

4.1 Dataset

We employ four real-world spatiotemporal datasets across three domains. **Traffic.** (1) SIP: Camera surveillance capturing traffic volumes in Suzhou Industry Park (SIP), (2) Metr-LA: Traffic attributes detected by highway loop detectors of Los Angeles, USA. **Climate.** (3) KnowAir: $PM_{2.5}$ concentrations, covering 184 main cities of China [28]. **Smart grid.** (4) Electricity: Hourly urban electricity consumption of 321 clients from 2012 to 2014 [36]. The dataset descriptions are illustrated in the Appendix.

4.2 Baseline

We exploit eight prevalent baselines for spatiotemporal learning, and incorporate IRM with two best baselines to compare our solution against IRM. **Non-invariance learning: (1) STGCN:** A graph-based spatiotemporal framework with a sandwich structure by 1D temporal convolution [40]. **(2) MTGNN:** A graph-based multi-variate time series learning without defining explicit graph topology [36]. **(3) Graph WaveNet (GWN):** A graph-based traffic prediction model that integrates TCNs and GCNs [37]. **(4) DCRNN:** A diffusion convolutional recurrent neural network, which combines diffusion graph convolutions with RNNs [20]. **(5) ASTGNN:** An attention-based spatiotemporal network for capturing dynamic ST correlations [11]. **(6) AdaRNN:** An adaptive series learning with flexible distribution matching, which first tackles temporal covariate shifts [9] [1]. **(7) ST-SSL:** A self-supervised learning tailored for spatiotemporal framework [14]. **(8) STDEN:** A physics-guided SOTA solution to spatiotemporal learning [15]. **Invariant learning: (9) MTGNN+IRM:** We integrate MTGNN [36] with IRM [1] by minimizing variance of risks across sub-environments, inheriting the idea from [33, 35]. **(10) GWN-IRM:** Similar to (9), we integrate GWN [37] with IRM.

4.3 Implementation protocols

Following common settings in ST learnings [3, 11, 36], we implement 12-step ahead prediction by exploiting previous 12 steps. Our

[1] https://github.com/jindongwang/transferlearning/tree/master/code/deep/adarnn

Table 1: Periodicity and sub-environment partition

Dataset	SIP	Metr-LA	KnowAir	Electricity
Time steps	25,920	34,272	11,688	26,304
Interval	5 min	5 min	3 h	15 min
Periodicity	daily	daily	weekly	daily
Periodical steps	288	288	56	96
Env Partition K	6	6	6	6

temporal environment partition divides time steps within one periodicity into several segments where each sub-environment contains the even number of samples. The sub-environment partition K and intervals of periodicity are illustrated in Table 1. For training/validation/test set partition, we divide the whole sets into 2:1:1 for training, validation and testing. We implement our CauSTG based on GraphWavenet and do not pre-define any graph structure before model training. We adopt Adam [17] as the optimizer with learning rate of 1e-4. Mean Absolute Percentage Error (MAPE) is considered as the main evaluation metric, i.e., MAPE = $\frac{|\widehat{y}-y|}{y}$.

Given the problem of OOD inference, we consider **three OOD scenarios** as below to validate our CauSTG. The following settings are guaranteed across all baselines for fairness.

(1) Temporal covariate shift. To verify the performances on temporal covariate shifts, we construct training set by selecting $K/2$ sub-environments for training $\{e_{s_1}, ..., e_{s_{K/2}}\}$ while let half samples within sub-environments that have not appeared in training sets as validation and testing sets. With our partition principle on distinctions, this setting naturally forms different distributions between training and testing sets.

(2) Inductive setting. To explore the inductive extrapolation capacity on unseen samples, i.e., adaptation of new nodes in cities, we mask 5% of total nodes to imitate their non-involvements in training stage and involve them during testing. For implementation, we correspondingly find the most proximal node for each masked node, and copy their node-specific adjacency to new nodes, thus constructing an extended relational spatial adjacency for testing. We can leverage the power of our ST consistency learner to implement extrapolation. The prediction errors can be evaluated.

(3) Injective artificial noise. We explore the noise sensitivity via injecting Gaussian noise, which verifies whether our model can still capture the invariant relations under noisy data. The injected noise follows $N(\mu_0, \sigma_0)$, where μ_0 and σ_0 are the mean and variance of observations within corresponding sub-environment. We defer the detailed influences of noise intensity in Section 4.6.

4.4 Comparison results

Comparison results can be found in Table 2. Generally, regardless of IRM and CauSTG, learning with invariance can exactly improve OOD predictions. For IRM, we discover the less significance on temporal shifts for MTGNN+IRM, where we speculate the reason lies in MTGNN lacking the scalable relation designs for invariance extraction. In contrast, our CauSTG exactly competes IRM-based baselines and consistently achieves better results on almost all OOD scenarios, outperforming non-invariance baselines by approximately 1% to 10% while invariance ones by 0.5% to 7.35%. Specifically, by observing performances on different tasks across all baselines, we find that the task of new node involvement is much harder than temporally covariate shifts, correspondingly our CauSTG achieves most significant improvement. We especially attribute such superiority to two issues compared with other solutions, 1) the cooperative works between consistency learner and invariance explorer exactly identifies the causally spatial relations and temporal variation trends, which help better extrapolate on unseen data, 2) the copy of node neighborhood can empower the learned consistency to make sense on predicting unseen nodes. Moreover, our solution also preserves satisfactory performances as the disentangled and scalable relation learning can separate the factors thus be resilient to noise injections.

In addition, it is worth noting that our model reveals better performance in Electricity. The underlying reason is that our scalable bi-directional spatial learning is more superior on node-level datasets without explicit graph structures. For Electricity, the users do not have explicit neighborhood but tend to reveal similar consumption patterns with consistent variation trends. Thus, it is opportunely suitable to model such user-level similarity (i.e., spatial consistency) and evolutional trend consistency by our ST consistency learner.

In summary, our CauSTG is with superiority in three aspects, 1) more adaptation capacity on inductive settings (new nodes) due to extracted invariant relation, 2) more friendly to datasets without pre-defined adjacencies due to scalable relation learning, 3) resilient to injective noise with its hierarchical invariance explorer.

4.5 Ablation study

To uncover the significance of each module to the success of CauSTG, we perform an ablation study on **temporal covariate shifts** via removing each module or replacing it with a vanilla one. The ablated variants are as follows. **(1) CauSTG-Adj:** Replace bi-directional spatial relation learning with binary adjacent matrix constructed by geographical distances or node behavior similarity. [2] **(2) CauSTG-GRU:** Replace the disentangled temporal learning with GRU. **(3) CauSTG-NoHier:** Skip local invariance and directly obtain global invariance across all steps. The detailed results are in Table 6. As observed, the hierarchical invariance explorer plays the most significant role in OOD learning as the sparse connections of invariance across all steps are unacceptable. Concretely, removing the inverse of $\boldsymbol{X}$ leads to the loss of modeling negatively causal correlations thus deteriorating performances. Replacing temporal learning with only sequential perspective GRU becomes less scalable for invariance learning, thus inferior to seasonal-trend ones. Remarkably, these results can match with baseline performances on similar settings, which further verifies better designs of our modules.

4.6 Model hyperparameter analysis

We take temporal covariate shift to evaluate the impacts of hyperparameter on model performances. We report averaged MAPE of SIP and KnowAir on testing set in Fig. 4 while results of another two sets are in the Sec. B.4.

(1) Number of sub-environment partition K. More sub environments can lead to finer granularity of temporal partitions thus more stable relations can be captured, but more environments

[2] Adjacencies of SIP, Metr-LA and KnowAir are established by geographical distance while graph of Electricity is constructed by cosine similarity between users.

Table 2: Performance comparisons on OOD scenarios against baselines. The overall best results are bold, results of best non-invariance and invariant learning are respectively marked with underline and *.

	SIP			Metr-LA			KnowAir			Electricity		
	Temporal shift	New nodes	Artifical noise	Temporal shift	New nodes	Artifical noise	Temporal shift	New nodes	Artifical noise	Temporal shift	New nodes	Artifical noise
STGCN	22.75%	26.7key 2%	23.36%	12.62%	15.13%	13.53%	31.71%	42.87%	33.94%	2.65%	4.93%	3.93%
MTGNN	20.09%	23.74%	20.70%	10.05%	12.56%	11.25%	24.06%	36.22%	29.59%	2.12%	4.41%	3.65%
GWN	20.13%	23.65%	20.84%	10.01%	12.52%	11.17%	24.13%	36.21%	29.53%	2.08%	4.34%	3.51%
DCRNN	21.17%	24.64%	21.88%	10.50%	13.01%	11.41%	25.17%	36.23%	30.43%	2.31%	4.68%	3.83%
ASTGNN	22.31%	25.87%	22.92%	10.04%	12.55%	10.99%	26.31%	37.43%	31.27%	2.33%	4.71%	3.79%
AdaRNN	21.22%	24.78%	22.79%	10.14%	13.08%	11.58%	24.60%	36.47%	30.76%	2.10%	4.45%	3.88%
ST-SSL	21.75%	23.43%	22.44%	10.61%	12.42%	12.86%	24.01%	36.07%	30.03%	2.15%	4.04%	3.56%
STDEN	21.88%	24.67%	21.70%	11.03%	12.83%	12.36%	25.13%	36.97%	30.65%	2.23%	4.88%	3.97%
MTGNN+IRM	20.21%	23.86%	20.52%*	10.02%	12.53%	11.03%	24.03%	35.14%	29.46%	2.05%	4.21%	3.28%*
GWN+IRM	20.01%*	23.56%*	20.64%	9.94%*	12.45%*	10.95%*	24.01%*	35.12%*	29.34%*	2.04%*	4.13%*	3.33%
CauSTG	**19.91%**	**23.03%**	**20.35%**	**9.75%**	**12.34%**	**10.64%**	**23.63%**	**34.32%**	**28.95%**	**1.89%**	**3.89%**	**3.15%**
Beyond non-inv	0.90%	1.71%	1.69%	2.60%	0.64%	3.18%	1.58%	4.85%	1.96%	9.13%	3.71%	10.26%
Beyond inv	0.50%	2.25%	0.83%	1.91%	0.88%	2.83%	1.58%	2.28%	1.33%	7.35%	5.81%	3.96%

Table 3: Ablation study

Variants	SIP	Metr-LA	KnowAir	Electricity
CauSTG-Adj	21.10%	11.60%	26.14%	2.24%
CauSTG-GRU	21.62%	10.55%	25.17%	2.30%
CauSTG-NoHier	23.26%	13.42%	26.68%	2.84%
CauSTG	19.91%	9.75%	23.63%	1.89%

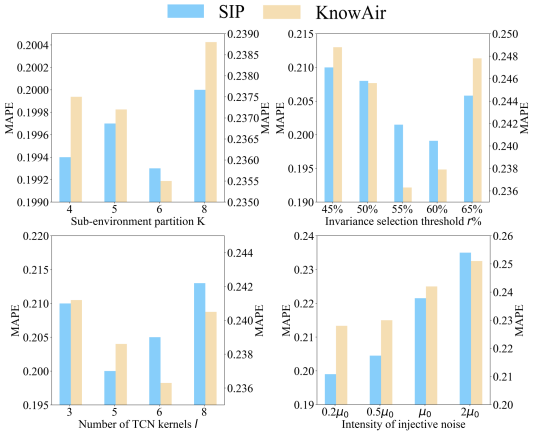


Figure 4: Hyperparameter analysis on SIP and KnowAir

will also introduce higher spatial and temporal complexity in the training process. We thus adjust $K \in \{4, 5, 6, 8\}$ to find out the eclectic partition number. We find the performances are stable across partition numbers and choose $K = 6$ as an eclectic one for all sets.

(2) Invariance filtering threshold $r\%$. We let $r\%$ vary from $\{45\%, 50\%, 55\%\ 60\%\}$ to obtain a well-fit threshold. There is the trade-off between reliable but sparse stability with smaller r and the less reliable but richer relation connections with larger r. We find $r\% = 60\%, 55\%, 55\%, 50\%$ for SIP, Metr-LA, KnowAir and Electricity, which means that unstable relations accounting for 40%, 45%, 45%, 50% of total parameters are removed for invariant learning.

(3) Number of TCN kernels l. We enable l varying from $\{3, 5, 6, 8\}$ for each dataset to get suitable ones for temporal pattern extractions. In fact, we arrive $l = 5$ for SIP and Metr-LA while $l = 6$ for KnowAir and Electricity.

(4) Intensity of injective noise δ. We inject artificial noise into training samples of SIP and KnowAir, where the noise is set as $\delta = \{0.2\mu_0, 0.5\mu_0, \mu_0, 2\mu_0\}$, μ_0 is the average of observations of corresponding environment. As observed, our model can be resilient to noise even the intensity rising to μ_0 and $2\mu_0$, where the reason lies in the superiority of disentangled relation and trend extraction.

4.7 Case study

In this subsection, we visualize model weights within three temporal environments on Metr-LA in Fig. 5, including both spatial relations and temporal patterns.

The non-zero entries in $\mathbf{A}_0$ describe stable positive correlations between node pairs. In particular, we observe that Nodes 51 and 52 (highlighted with red circles) reveal relatively consistent relations across 3 environments, especially in the morning and evening. We look up to real-world maps and find that these two nodes lie in highway around Universal Studios Hollywood, which inherently share an up-downstream relation with physically positive correlation and further confirm the rationale of our CauSTG.

In temporal patterns, we clip the matrix w_{tr} into a sub weight sized by 15×12, and discover that trend patterns w_{tr} are more easier to reach consistency across sub-environments while the intensity of seasonal patterns w_{ts} usually cannot reach a consensus. This reason can be that the convolution kernels in w_{ts}, which determine the intensity values, are distinctive across environments, but trend patterns usually share across environments. This observation can also be verified in Fig. 1 (a). Therefore, in predictions, the value intensity (i.e., w_{ts}) is more dependent on local invariance while the trend (i.e., w_{tr}) can be attributed to invariance across environments.

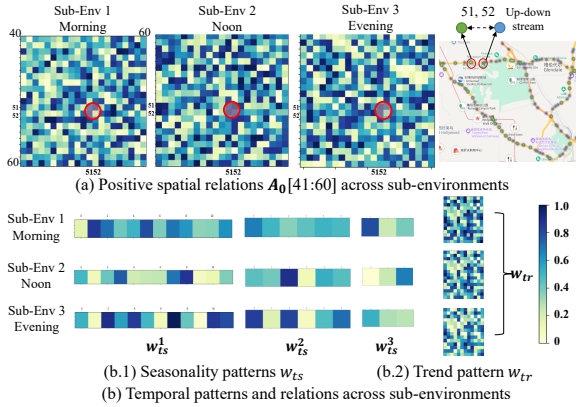


Figure 5: Visualization of invariant relations

5 RELATED WORK

Spatiotemporal learning addresses challenging tasks such as traffic prediction [11, 24], precipitation forecasting [13, 16] and real estate value estimation [23, 27]. Recent spatiotemporal learning usually falls into deep learning-based solutions. At an early stage, the urban areas are partitioned into grids and CNN architectures are exploited to capture spatial dependencies [43]. Further, spatiotemporal data is organized into graph structure for non-Euclidean modeling where each node possesses location-specific observation and edges carry node-wise relations. Followed up, multi-view graph convolution [6, 10, 12], inner product-based adaptive graph learning [36] are proposed to boost the performances. Besides, temporal learning is also introduced with GRU [3, 21] or TCN [2, 40]. Despite prosperity, the city expansions and urban constructions raise covariate shift concern on spatiotemporal observation. To this end, TrafficStream devises to periodically update the training models by re-sampling and re-weighting importance to realize continuous learning and confront distribution shifts [5]. However, this work still suffers the challenge of how to tradeoff historical and new knowledge. Technically, continuous learning models require periodically rolling training thus failing to capture the invariance within changing environments. Thus, exploring OOD generalization of ST models is of great need but hardly explored.

Time-series learning with covariate shift. Traditional sequential regressions usually adopt auto-regressive based solution such as ARIMA [44]. With the success of deep learning, variants of RNNs, such as LSTM [41] and GRU [7] with gated modules are proposed to avoid the gradient vanishment. Also, the powerful language model, Transformer, is introduced to enable longer horizon predictions of structured time series [32, 34]. More recently, Du, et, al [9] point out the covariate shift issue in time series, and correspondingly propose the AdaRNN to enable adaptive aligned regression with modules of distribution characterization and matching. Follow-up works transfer the series from temporal domain into frequency to realize multi-periodicity detection [30], and empower more robust [45] and longer-term forecasting [32]. Specifically, CoST decomposes seasonality-trend information to obtain the disentangled representations, which theoretically justifies its robustness via causal perspective [31]. Even though, these works all focus on series-level without introducing spatial dependencies. In fact, spatial correlations in ST learning can bring in further challenges due to the non-independence identical distribution (non-i.i.d.) issue, which is inherently not considered in the traditional series learning.

OOD learning is usually tackled by causal theory, and these solutions can be classified into counterfactual-based and invariant learning-based. Counterfactual prediction incorporates virtual but rational samples derived by causal inference, which expects maximal generalization on unseen domains by increasing sample diversity [42]. On graphs, these solutions take neighbor pairs as context and global graph structural properties as treatment, to realize rationale discovery [46]. However, such solution is inherently a kind of data augmentation and fails to refine intrinsic data regularity. For invariant learning, it usually assumes the existence of virtual environments and the distributions of covariates (i.e., input x) can be varying across different environments. Specifically, it leverages IRM [1] to capture invariance across environments. This idea has been adapted to graph-level classification tasks by proposing DIR [35], OOD-GNN [18], and MoleOOD [38] via finding an invariant but sufficient substructure on graphs. However, these studies are built on static graphs that cannot well fit in node-level dynamic regressions. Very recently, EERM [33] overcomes the non-i.d.d. issue on node-level learning and learns the invariance via maximizing the risk variances from multiple environments. Even though, all these works cannot simultaneously accommodate spatial correlation and temporal dependence from causal perspective, which respectively faced with non-i.d.d. issue and heterogeneous temporal environments.

6 CONCLUSION

In this paper, we propose a causal spatiotemporal learning framework, CauSTG, to tackle the covariate shift in ST learning. CauSTG converts invariant representation learning into capturing stable trainable weights. Specifically, we take temporal steps as environments and partition temporal environments by identifying distinctive segments, enabling hierarchical invariance extraction. To facilitate quantified invariance extraction, we modify traditional ST model into a spatiotemporal consistency learner, which empowers bi-directional spatial relation and decomposed seasonality-trend knowledge to be well-captured. Finally, a novel training scheme, the hierarchical invariance explorer, is devised to filter and ensemble the stable weights via measuring weight discrepancy across sub-environments, capturing both segment-level local invariance and global invariance. Experimental results on three OOD scenarios validate the performance superiority and exhibit intrinsic interpretation of our OOD solutions. For future work, we will work on scalable training strategies for iteratively refining the stable models and explore model adaptations on different OOD scenarios.

ACKNOWLEDGMENTS

This paper is partially supported by the National Natural Science Foundation of China (No.62072427, No.12227901), the Project of Stable Support for Youth Team in Basic Research Field, CAS (No.YSBR-005), Academic Leaders Cultivation Program, USTC.

REFERENCES

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2020. Invariant Risk Minimization. *stat* 1050 (2020), 27.

[2] Lei Bai, Lina Yao, Salil S Kanhere, Xianzhi Wang, and Quan Z Sheng. 2019. STG2seq: spatial-temporal graph to sequence model for multi-step passenger demand forecasting. In *IJCAI*. International Joint Conferences on Artificial Intelligence, 1981–1987.

[3] Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. 2020. Adaptive Graph Convolutional Recurrent Network for Traffic Forecasting. *NIPS* 33 (2020).

[4] Beatrice Bevilacqua, Yangze Zhou, and Bruno Ribeiro. 2021. Size-invariant graph representations for graph classification extrapolations. In *International Conference on Machine Learning*. PMLR, 837–851.

[5] Xu Chen, Junshan Wang, and Kunqing Xie. 2021. TrafficStream: A Streaming Traffic Flow Forecasting Framework Based on Graph Neural Networks and Continual Learning. *arXiv preprint arXiv:2106.06273* (2021).

[6] Shaojie Dai, Jinshuai Wang, Chao Huang, Yanwei Yu, and Junyu Dong. 2022. Dynamic Multi-View Graph Neural Networks for Citywide Traffic Inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)* (2022).

[7] Rahul Dey and Fathi M Salem. 2017. Gate-variants of gated recurrent unit (GRU) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*. IEEE, 1597–1600.

[8] Wenjie Du, Lianliang Chen, Haoran Wang, Ziyang Shan, Zhengyang Zhou, Wenwei Li, and Yang Wang. 2023. Deciphering urban traffic impacts on air quality by deep learning and emission inventory. *Journal of Environmental Sciences* 124 (2023), 745–757.

[9] Yuntao Du, Jindong Wang, Wenjie Feng, Sinno Pan, Tao Qin, Renjun Xu, and Chongjun Wang. 2021. Adarnn: Adaptive learning and forecasting of time series. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 402–411.

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

[11] Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. 2019. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *AAAI*, Vol. 33. 922–929.

[12] Liangzhe Han, Bowen Du, Leilei Sun, Yanjie Fu, Yisheng Lv, and Hui Xiong. 2021. Dynamic and multi-faceted spatio-temporal deep learning for traffic speed forecasting. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 547–555.

[13] Pradeep Hewage, Marcello Trovati, Ella Pereira, and Ardhendu Behera. 2021. Deep learning-based effective fine-grained weather forecasting model. *Pattern Analysis and Applications* 24, 1 (2021), 343–366.

[14] Jiahao Ji, Jingyuan Wang, Chao Huang, Junjie Wu, Boren Xu, Zhenhe Wu, Junbo Zhang, and Yu Zheng. 2023. Spatio-Temporal Self-Supervised Learning for Traffic Flow Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[15] Jiahao Ji, Jingyuan Wang, Zhe Jiang, Jiawei Jiang, and Hu Zhang. 2022. STDEN: Towards physics-guided neural networks for traffic flow prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 4048–4056.

[16] Seongchan Kim, Seungkyun Hong, Minsu Joh, and Sa-kwang Song. 2017. Deeprain: Convlstm network for precipitation prediction using multichannel radar data. *arXiv preprint arXiv:1711.02316* (2017).

[17] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

[18] Haoyang Li, Xin Wang, Ziwei Zhang, and Wenwu Zhu. 2022. Ood-gnn: Out-of-distribution generalized graph neural network. *IEEE Transactions on Knowledge and Data Engineering* (2022).

[19] Sihang Li, Xiang Wang, An Zhang, Yingxin Wu, Xiangnan He, and Tat-Seng Chua. 2022. Let invariant rationale discovery inspire graph contrastive learning. In *International Conference on Machine Learning*. PMLR, 13052–13065.

[20] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. 2018. Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. In *International Conference on Learning Representations*.

[21] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. 2018. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *ICLR*.

[22] Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. 2021. Heterogeneous risk minimization. In *International Conference on Machine Learning*. PMLR, 6804–6814.

[23] Xiaolong Liu. 2013. Spatial and temporal dependence in house price prediction. *The Journal of Real Estate Finance and Economics* 47, 2 (2013), 341–369.

[24] Hao Miao, Jiaxing Shen, Jiannong Cao, Jiangnan Xia, and Senzhang Wang. 2022. MBA-STNet: Bayes-enhanced Discriminative Multi-task Learning for Flow Prediction. *IEEE Transactions on Knowledge and Data Engineering* (2022).

[25] Vincent Morello, ED Barr, BW Stappers, EF Keane, and AG Lyne. 2020. Optimal periodicity searching: revisiting the fast folding algorithm for large-scale pulsar surveys. *Monthly Notices of the Royal Astronomical Society* 497, 4 (2020), 4654–4671.

[26] Manali Shaha and Meenakshi Pawar. 2018. Transfer learning for image classification. In *2018 second international conference on electronics, communication and aerospace technology (ICECA)*. IEEE, 656–660.

[27] Pengkun Wang, Chuancai Ge, Zhengyang Zhou, Xu Wang, Yuantao Li, and Yang Wang. 2021. Joint Gated Co-attention Based Multi-modal Networks for Subregion House Price Prediction. *IEEE Transactions on Knowledge and Data Engineering* (2021).

[28] Shuo Wang, Yanran Li, Jiang Zhang, Qingye Meng, Lingwei Meng, and Fei Gao. 2020. Pm2. 5-gnn: A domain knowledge enhanced graph neural network for pm2. 5 forecasting. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*. 163–166.

[29] Senzhang Wang, Hao Miao, Hao Chen, and Zhiqiu Huang. 2020. Multi-task adversarial spatial-temporal networks for crowd flow prediction. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 1555–1564.

[30] Qingsong Wen, Kai He, Liang Sun, Yingying Zhang, Min Ke, and Huan Xu. 2021. RobustPeriod: Robust time-frequency mining for multiple periodicity detection. In *Proceedings of the 2021 International Conference on Management of Data*. 2328–2337.

[31] Gerald Woo, Chenghao Liu, Doyen Sahoo, Akshat Kumar, and Steven Hoi. 2022. CoST: Contrastive Learning of Disentangled Seasonal-Trend Representations for Time Series Forecasting. In *International Conference on Learning Representations*.

[32] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. 2021. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems* 34 (2021), 22419–22430.

[33] Qitian Wu, Hengrui Zhang, Junchi Yan, and David Wipf. 2022. Handling Distribution Shifts on Graphs: An Invariance Perspective. In *International Conference on Learning Representations*.

[34] Sifan Wu, Xi Xiao, Qianggang Ding, Peilin Zhao, Ying Wei, and Junzhou Huang. 2020. Adversarial Sparse Transformer for Time Series Forecasting. In *NIPS*.

[35] Yingxin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. 2021. Discovering Invariant Rationales for Graph Neural Networks. In *International Conference on Learning Representations*.

[36] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. 2020. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *KDD*. 753–763.

[37] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. 2019. Graph wavenet for deep spatial-temporal graph modeling. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 1907–1913.

[38] Nianzu Yang, Kaipeng Zeng, Qitian Wu, Xiaosong Jia, and Junchi Yan. 2022. Learning substructure invariance for out-of-distribution molecular representations. In *Advances in Neural Information Processing Systems*.

[39] Xiuwen Yi, Junbo Zhang, Zhaoyuan Wang, Tianrui Li, and Yu Zheng. 2018. Deep distributed fusion network for air quality prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 965–973.

[40] Bing Yu, Haoteng Yin, and Zhanxing Zhu. 2018. Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. In *IJCAI*. 3634–3640.

[41] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. 2019. A review of recurrent neural networks: LSTM cells and network architectures. *Neural computation* 31, 7 (2019), 1235–1270.

[42] Guozhen Zhang, Jinwei Zeng, Zhengyue Zhao, Depeng Jin, and Yong Li. 2022. A Counterfactual Modeling Framework for Churn Prediction. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 1424–1432.

[43] Junbo Zhang, Yu Zheng, and Dekang Qi. 2017. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Thirty-first AAAI conference on artificial intelligence*.

[44] Mingda Zhang. 2018. Time Series: Autoregressive models AR, MA, ARMA, ARIMA. *University of Pittsburgh* (2018).

[45] Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. [n. d.]. Self-Supervised Contrastive Pre-Training For Time Series via Time-Frequency Consistency. In *Advances in Neural Information Processing Systems*.

[46] Tong Zhao, Gang Liu, Daheng Wang, Wenhao Yu, and Meng Jiang. 2021. Counterfactual graph learning for link prediction. *arXiv preprint arXiv:2106.02172* (2021).

[47] Zhengyang Zhou, Yang Wang, Xike Xie, Lianliang Chen, and Chaochao Zhu. 2020. Foresee Urban Sparse Traffic Accidents: A Spatiotemporal Multi-Granularity Perspective. *IEEE TKDE* (2020).

A PROOF FOR THEORETICAL ANALYSIS

In this section, we inherit the assumptions and notations from Sec. 2.2.

A.1 Proof of Proposition 1

Prop 1. Failure of non-invariance learning on OOD regression. We consider one-time GNN aggregation of its neighbors, from T-step to achieve the expected regression prediction of $T+1$-step. Based on above assumptions, for node v_i, we take $\mathcal{N}_c(v_i)$ as the causally correlated neighbor set of v_i while $\mathcal{N}_s(v_i)$ denotes the set of non-causally correlated neighbors. Given degree d_i, the traditional one-time aggregation for v_i with all nodes can be formulated by decomposing the causal and non-causal parts,

$$E(h_i^T) = \frac{x_i^T + \sum\limits_{c_j \in \mathcal{N}_c(v_i)} w_{ij}^c x_{c_j}^T + \sum\limits_{s_j \in \mathcal{N}_s(v_i)} w_{ij}^s x_{s_j}^T}{1+d_i} \tag{17}$$

where c_j and s_j are the subscripts of two neighborhood sets, and w_{ij}^c and w_{ij}^s are learnable weights for causal parts and non-causal parts (spurious correlations). By calculating the difference between the aggregated expectation $E(h_i^T)$ and groundtruth x_i^{T+1}, we can derive the prediction error ε_0 after one-time aggregation by neglecting the non-linear activations,

$$\begin{aligned} \varepsilon_0 &= ||E(h_i^T) - x_i^{T+1}|| \\ &= || \frac{x_i^T + \sum\limits_{c_j \in \mathcal{N}_c(v_i)} w_{ij}^c x_{c_j}^T + \sum\limits_{s_j \in \mathcal{N}_s(v_i)} w_{ij}^s x_{s_j}^T - (1+d_i) x_i^{T+1}}{1+d_i} || \end{aligned} \tag{18}$$

Assume that observations on both current step x_i^T and next step x_i^{T+1} follow the same Gaussian distribution $N(\mu_0, \sigma_0)$, and $p_i = \frac{||\mathcal{N}_c(v_i)||}{||\mathcal{N}(v_i)||}$ accounts for the proportion of causal neighbors. To facilitate the expression, we let μ_0^t, μ_0^{t+1} denote the expectation of observation x_i at t and $t+1$, and μ_0^c, μ_0^s represent the expectation of the expected observation of its causal neighborhood and non-causal (spurious) neighborhood. The initial error of Eq. 18 can be modified by,

$$\varepsilon_0 = \frac{\mu_0^t + p_i d_i \mu_0^c w_i^c + (1-p_i) d_i \mu_0^s w_i^s - (1+d_i)\mu_0^{t+1}}{1+d_i} \tag{19}$$

where we ignore the sign for absolute value, and assume that the expectation and learnable weights all preserve positive.

Since the non-causal based learning is formulated by regression function of $\widehat{y}_i = w_i^c x_c + w_i^s x_s$, the prediction residual res_i will be derived by $res_i = \widehat{y}_i - w_i^c x_c = w_i^s x_s$. Therefore, we can substitute the difference between aggregated causal parts and groundtruth with aggregated non-causal part, and obtain the following equations,

$$\begin{aligned} \varepsilon_0 &= \frac{\mu_0^t + p_i d_i \mu_0^c w_i^c - (1+d_i)\mu_0^{t+1} + (1-p_i) d_i \mu_0^s w_i^s}{1+d_i} \\ &= \frac{2(1-p_i) d_i \mu_0^s w_i^s}{1+d_i} \end{aligned} \tag{20}$$

With Eq. 20, we can arrive that the derived error is not reducible as $w_i^c \neq 0$. Then, we can further disentangle the influence factors of this error. As causal parts are defined on the stable relations while spurious parts are defined on highly variant correlations across distribution (environments), we can make the assumption of the distributions of corresponding learnable weights by,

$$\begin{gathered} w_i^c \sim N(\mu_w, \sigma_{wc}), \quad w_i^s \sim N(\mu_w, \sigma_{ws}) \\ s.t.\ \sigma_{ws} \gg \sigma_{wc} \end{gathered} \tag{21}$$

Given that if one random variable follows Gaussian distribution, then 99.73% of the samples fall into the ranges between $[\mu - 3\sigma, \mu + 3\sigma]$, according to the 'Three Sigma Principe'. This principle tells us almost all samples must fall into above ranges excluding very few extreme values. Then we can approximate error ε_0 with restoring μ_0^s to μ_0,

$$\varepsilon_0 \sim \frac{2(1-p_i) d_i \mu_0 (\mu_w \pm 3\sigma_{ws})}{1+d_i} \tag{22}$$

Since two variables of p_i satisfying $0 < p_i < 1$ and $\sigma_{ws} \gg \sigma_{wc}$ are constants that cannot be ignorable, the errors for non-invariance learning will be positively proportional to both $\mu_0 \mu_w$ and σ_{ws} while negatively correlated with p_i. That's to say, the less causal parts within observations, i.e., smaller p_i, and the larger variations of relations across environments, i.e., larger σ_{ws}, the performance deterioration will be more serious. Moreover, when f is transferred to OOD testing set $N(\mu_q, \sigma_q | e_q)$ satisfying $\mu_q = q\mu_0$ where $q \in \mathbb{N}^+$. The approximated error of OOD testing risks is amplified to,

$$\varepsilon_q \sim \frac{2(1-p_i) d_i q \mu_0 (\mu_w \pm 3\sigma_{ws})}{1+d_i} \tag{23}$$

To this end, such result manifests an unacceptable amplification of error bounds from in-distribution to out-of-distribution samples. Thus, we can arrive the conclusion that non-invariant relation learning is inclined to fail on OOD regressions.

As a result, such representative example can be further generalized to all regression tasks based on aggregations. □

A.2 Proof of Proposition 2

Prop 2. Potential to eliminate the error amplification by only capturing invariant relations.

If we only capture the neighbors that are causally correlated, the errors ε_c can be derived by following equation,

$$\begin{aligned} \varepsilon_c &= \frac{\mu_0^t + p_i d_i \mu_0^c w_i^c - (1+d_i)\mu_0^{t+1}}{1+d_i} \\ &= \frac{(p_i w_i^c - 1) d_i \mu_0^c}{1+d_i} \end{aligned} \tag{24}$$

Since w_i^c are learnable parameters, there must exist an optimal point $w_i^c = \frac{1}{p_i}$ satisfying $\varepsilon_c \sim 0 \ll \varepsilon_0$. Thus, solution with invariant relation preserved can enable the error converge to 0 by appropriate optimization.

Finally, we can conclude the proof of this proposition and verify the correctness of our motivation, i.e., capturing invariant relation invariance for OOD regression. □

B DETAILS ON EXPERIMENTS

In this section, we will present some more details on experiments, regarding experimental configurations, dataset statistics, as well as parameter influences on performance.

B.1 Experimental configurations

To facilitate the reproducibility, we list the detailed configuration of our CauSTG in Table 4 for reference.

Table 4: Configuration of CauSTG. The configurations are displayed in the order of SIP, Metr-LA, KnowAir and Electricity if different values are specified by datasets.

Parameter	Concrete values
Backbone of CauSTG	GraphWaveNet (GWN)
Sample split	Train/Validation/Test: 2/1/1
Sub-environment partition K	6
Model number of sub-environment m	4
Balance coefficient $(\lambda_0, \lambda_1, \lambda_2)$	(0.5, 0.5, 0.2)
Learning rate	1e-3
Invariance filtering threshold $r\%$	(60%,55%,55%,50%)
Number of TCN kernels l	(5,5,6,6)
TCN kernel dimension	(12,6,3)
Proportion of masked nodes	5%
Hidden dimension of GNN d_m	64
Optimizer	Adam

Table 5: Dataset statistics

Dataset	Node #	Time step #	Time span	Interval length	Mean/Std var in periodicity
SIP	108	25,920	01/01/2017-03/31/2012	5min	72.17/55.45
Metr-LA	207	34,272	03/01/2012-06/30/2012	5min	27.45/30.25
KnowAir	184	11,688	01/01/2015-12/31/2018	3h	52.69/61.60
Electricity	321	26,304	01/01/2012-12/31/2014	15min	2538.79/820.92

B.2 Statistics of datasets

We figure out the statistics of four cross-domain datasets in Table 5, to illustrate their properties. Noted that the mean within periodicity represents the statistical average of each dataset while the standard variance within periodicity describes the averaged standard variances of interval-level observations within each periodicity on corresponding datasets. **The statistics exactly reveal the strong variation within each periodicity that verifies the segment-level heterogeneity, and environment partition can exactly contribute to (imitate) the distribution shifts.**

B.3 Results on CauSTG+X

To support the generalization of our CauSTG, we combine our proposed the hierarchical invariance, i.e., the strategies of hierarchical training and stable weight selection, with two best ST learning baselines, GWN, and MTGNN. The experiments are conducted on temporal covariate shift scenario. By comparison, we can see that the invariance explorer can indeed improve the performances on OOD scenarios, verifying the model-agnostic property of our solution. In addition, our integrated CauSTG still outperforms others due to the improved design of our scalable relation-enhanced learning.

Table 6: Ablation study

Variants	SIP	Metr-LA	KnowAir	Electricity
MTGNN	20.09%	10.05%	24.06%	2.12%
CauSTG-MTGNN	19.96%	9.98%	23.85%	2.10%
GWN	20.13%	10.01%	24.13%	2.08%
CauSTG-GWN	20.01%	9.96%	24.02%	1.98%
CauSTG	19.91%	9.75%	23.63%	1.89%

B.4 Hyperparameter analysis on performance

Result analysis and further insights. In this subsection, we further demonstrate the hyperparameter analysis on other two datasets, Metr-LA and Electricity, to support the completeness of our experiments. The detailed results are shown in Fig. 6. As observed, different parameters have various impacts on final predictions and we can exploit these results to review and consolidate the designs of our model and promote its adaptation to different datasets. For instance, larger-scale datasets, such as Metr-LA and Electricity, may require more learnable parameters to fit therefore the element-wise variations of weights become larger. To this end, a smaller r should be imposed to filter the stable relations for avoiding the sparse common invariance. In addition, we set $m = 4$ as the number of sub-models within each sub-environment. In an intuitive observation, the increasing number of models can first increase the performance and then decrease as more models will lead to more diversity thus challenging for capturing invariance. We can also modify the fine-tune epochs for each dataset to realize the optimized results which will be left as our future work. In summary, these results of more detailed analysis can provide insights into our further research on the wider universality of our CauSTG.

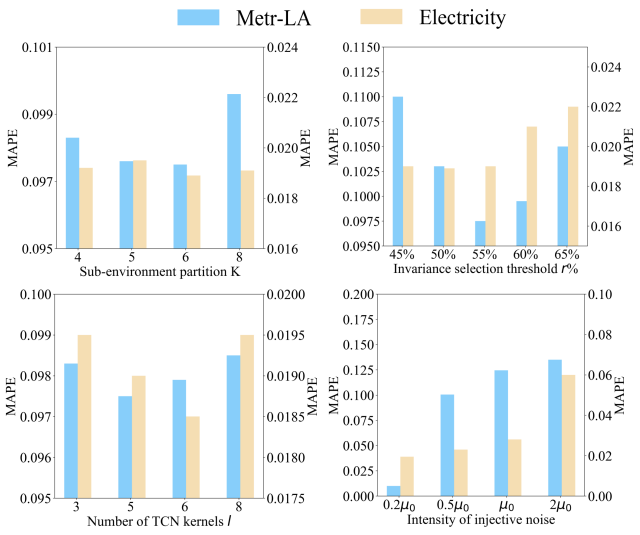


Figure 6: Hyperparameter analysis on MetrLA and Electricity

Table 7: Efficiency study

Dataset	Metr-LA		KnowAir	
Stage	train	inference	train	inference
MTGNN	0.41M	0.41M	0.38M	0.38M
GWN	0.28M	0.28M	0.28M	0.28M
DCRNN	0.37M	0.37M	0.37M	0.37M
ASTGNN	0.57M	0.57M	0.55M	0.55M
CauSTG	1.83M	0.43M	1.77M	0.41M

B.5 Efficiency study across different models

We evaluate the efficiency of our proposed CauSTG from three perspectives, time complexity, space complexity and the empirical parameter numbers.

Time complexity. Assuming K sub-environment and m models in each sub-environment, the total training workloads can be derived by the $K \times m$ times of traditional ST learning. Actually, our ST learning enables the model to automatically learn node-wise relations and temporally seasonal-trend pattern, the parameters becomes two-fold. Parameters of spatial learning is linearly with node number N while parameters of temporal pattern extraction are the summation of sizes of convolution kernels and trend transformation weights, which is also limited to linearity of node number. Therefore, the time complexity of training process is approximately $O(N * K * m)$. And for inference stage, our model combines the environment-specific local invariance weight $\mathbf{W}_k^*$ with global invariance $\mathbf{W}^*$ without additional computations. Thus, our model is efficient for one-training and permanent OOD inference. In addition, our solution does not require additional training time as the training set are sampled separated and our models including different sub-models can be trained parallelly on GPUs.

Space complexity. Although we train multiple (i.e., K) sub-models for each task, we subsequently integrate them into one model via a stability-based ensemble strategy. Such strategy is performed by filtering non-stable relation placeholders and therefore, our solution is actually more parameter-efficient with no increases of parameter numbers in inference stage. In the training stage, the introduced additional cost is much more deserved than making efforts on re-designs and modifications of models when new OOD data arriving.

Empirical parameter numbers. We derive the number of learnable parameters for several selected baselines as well as our CauSTG. As the parameter numbers are diverse across different training sets, we take the datasets Metr-LA and KnowAir as an example and show parameter numbers of models respectively at the training stage and inference stage in Table 7. The results illustrate that in inference stages, our CauSTG is with similar and comparable parameter numbers when compared with other baselines. Also, our CauSTG is with a medium scale among all these baselines and even has fewer parameters at inference stage than training due to the stable weight selection (filtering) process.

C TECHNICAL DETAILS OF CAUSTG

In this section, we provide some more technical details of our CauSTG to supplement the main texts. First, we provide the pseudo code of our CauSTG in Algorithm 1.

Algorithm 1 Causally spatiotemporal graph learning (CauSTG)

Input: Dataset $\mathbb{G}_s$, ST observations $\boldsymbol{X}$, number of samples m
Output: Well-learned invariant relations $\widetilde{\mathbf{W}}_k^*$ and predictor $f^*(\cdot)$

Initialization: ST consistency learner $f(\cdot)$ and its parameter set $\mathbf{W}$.
Get environment partition $\{e_1, e_2, ..., e_K\} \leftarrow$ Eq. 2 and Eq. 3 for $\boldsymbol{X}$
$\{\{\mathcal{G}^t\}_{t=1}^B\}_{k=1}^K \leftarrow$ Re-organize B samples for each sub-environment
Divide B samples into different m sub-set for sub-environment
for n = 0 to K **do**
  for t = 0 to m **do**
    Learn the weight $\mathbf{W}_{ij}$ based on ST consistency learner $f(\cdot)$
    Total loss: $\mathcal{L}$ = MAPE$(\widehat{Y}, Y) + \lambda_0 Loss_{se} + \lambda_1 Loss_{tr} + \lambda_2 \text{Reg}_{in}$
  end for
  Get local invariance $\mathbf{W}_k^* \leftarrow$ Eq. 15 on $\mathbf{W}_{kj}$
end for
Get global invariance $\mathbf{W}^*$ by filtering $\mathbf{W}_k^*$
Get the fused invariance $\widetilde{\mathbf{W}}_k^*$ and $f^*(\cdot) \leftarrow$ Eq. 16 on $\mathbf{W}^*$ and $\mathbf{W}_k^*$
Fine tune and update $\widetilde{\mathbf{W}}_k^* \leftarrow$ Shuffled $(\{\mathcal{G}^t\}_{t=1}^B)$
return $\widetilde{\mathbf{W}}_k^*, f^*(\cdot)$

Also, we present the detailed architecture of decomposed temporal pattern extractor in Fig. 7, which consists of a seasonal branch for multi-scale seasonality pattern extraction and a trend learning branch with multi-view trend constraints. Such disentangled learning is inspired by that the independence of learned representation can benefit causal learning and capture diverse independent relations. The learnable weights of seasonal learning $\boldsymbol{w}_{ts}$ and trend extraction $\boldsymbol{w}_{tr}$ encapsulate the relations between inputs and predicted observations including underlying periodicity-based amplitude and evolution trend transformation.

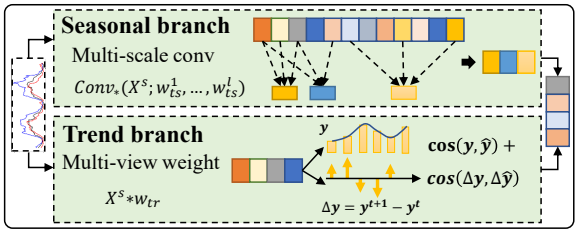


Figure 7: Detailed architecture of decomposed temporal pattern extractor

D FURTHER DISCUSSIONS OF CAUSTG

Model summary. Our CauSTG is a novel deep learning framework that introduces causal theory, i.e., invariant learning across environments, into spatiotemporal forecasting. The core idea is hierarchically re-organizing training samples guided by temporal environment partitions and maximally exploit available data to capture invariant relations from environment-specific model parameters. In addition, with the stability-based filtering, our CauSTG can be seen as interpreting uncertainty-based weight filtering to realize the invariant learning on OOD regression.

Besides, our CauSTG, with serveral emerging techniques, has a nice scalability on various series-based learning and regression tasks, which promotes the OOD studies from classification towards

regression and spatiotemporal learning. On OOD time-series learning, we can adopt our adaptive temporal partition strategy to progressively identify the distinguished sequence segment to construct diverse temporal environments. Then the consistent local patterns within sequence segments and global patterns across sequence segments can be well captured for OOD sample prediction. For regression tasks, it usually aims to achieve a series of learnable weights to fuse feature-level observations for obtaining targets. At this time, we can borrow the idea of environment partition in CauSTG and enable learnable weights to reflect the relations between features and targets. Then the relatively stable relations (i.e., weights) across different virtual environments can be further extracted for OOD inference.

Relation to existing causal inference literature. In this work, we accommodate spatiotemporal learning into a new invariant learning framework in causal perspective, even there is no explicit Structural Causal Model that is commonly appear in causality-based literature. Actually, we aim to learn the invariant spatiotemporal relations and disentangled seasonal-trend patterns across environments thus eliminating the non-stable relations or patterns that account for the shortcut features in structural causal model. Also, alternately training samples across different sub-environments can be viewed as the do-calculus on spatiotemporal variables and this mechanism can be the backdoor adjustment in causal theory, where the environments can be considered as a vertical variable to main observations. Thus, our hierarchical invariance explorer can eliminate the backdoors through the invariance filtering. Therefore, our model can be well adapted in the causal theory with satisfactory interpretations.

Relation to uncertainty theory. In the theory of uncertainty quantification, the epistemic uncertainty models learnable model weights as a distribution and captures such uncertainty by estimating the distribution parameters, such as the expectation and variance. In our solution, the weight variance within or across environments, indicating the stability of learnable relations, can naturally be viewed as the epistemic uncertainty of corresponding weights. Therefore, our model can also be seen as filtering the weights with less epistemic uncertainty during training process from uncertainty perspective.

Limitations and future works. In our work, we hierarchically partition samples into different environments and sample groups for invariance extraction, and exploit the rolling back strategy to fine-tune the neural networks after stability-based pruning. Therefore, there still remain two limitations, 1) how to systematize the hierarchical division process with the dataset property and 2) how to explore a better strategy to iteratively fine-tune the new pruned network structure? The potential solution to the first question is to conduct extensive experiments to empirically observe the performance variations and another solution becomes deriving the optimized partition from a well-designed spatial-temporal consistency metric. For the second one, the learning strategy can be further promoted by uncertainty quantification, parameter sensitivity analysis and theory of lottery tickets. The rolling-back and fine-tuning process can work alternately and iteratively. Both of them will be left as our future works. For generality, we will also explore how to adapt the learning strategies on different OOD scenarios such as new node involvement and covariate shifts, with slight modification.

Fairness and ethic issues. Our work performs extensive analysis and experiments on datasets including traffics, climate datasets concerning air quality, and consumptions of electricity, without any personal identity and privacy issues. Therefore, our work is with no ethics and privacy issues. In addition, all baselines and methods are compared with fairness.