Improving Generalization of Dynamic Graph Learning via Environment Prompt

Kuo Yang[1, 2], Zhengyang Zhou[1, 2, 3*], Qihe Huang[1, 2], Limin Li[1, 2],
Yuxuan Liang[4], Yang Wang[1, 2*]

[1] University of Science and Technology of China (USTC), Hefei, China
[2] Suzhou Institute for Advanced Research, USTC, Suzhou, China
[3] State Key Laboratory of Resources and Environmental Information System, Beijing, China
[4] The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
yangkuo@mail.ustc.edu.cn, zzy0929@ustc.edu.cn, {hqh, lilimin}@mail.ustc.edu.cn,
yuxliang@outlook.com, angyan@ustc.edu.cn

Abstract

Out-of-distribution (OOD) generalization issue is a well-known challenge within deep learning tasks. In dynamic graphs, the change of temporal environments is regarded as the main cause of data distribution shift. While numerous OOD studies focusing on environment factors have achieved remarkable performance, they still fail to systematically solve the two issue of environment inference and utilization. In this work, we propose a novel dynamic graph learning model named EpoD based on prompt learning and structural causal model to comprehensively enhance both environment inference and utilization. Inspired by the superior performance of prompt learning in understanding underlying semantic and causal associations, we first design a self-prompted learning mechanism to infer unseen environment factors. We then rethink the role of environment variable within spatio-temporal causal structure model, and introduce a novel causal pathway where dynamic subgraphs serve as mediating variables. The extracted dynamic subgraph can effectively capture the data distribution shift by incorporating the inferred environment variables into the node-wise dependencies. Theoretical discussions and intuitive analysis support the generalizability and interpretability of EpoD. Extensive experiments on seven real-world datasets across domains showcase the superiority of EpoD against baselines, and toy example experiments further verify the powerful interpretability and rationality of our EpoD.

1 Introduction

Dynamic graph learning aims to capture the evolution patterns of individual feature and global topology in spatio-temporal graphs over time. It has extensive applications in real-world scenarios, such as social relationship analysis [6, 61], traffic flow forecasting [3, 59, 64] and air quality prediction [16, 23]. The dynamic evolution is a prominent characteristic of spatio-temporal graphs [4, 16, 47], such as human interest and social development naturally undergo changes over time. This characteristic inevitably gives rise to a issue of data distribution shifts. Given this issue, enabling models to get temporal out-of-distribution (OOD) generalization ability poses a major challenge in dynamic graph learning [51, 55, 63].

Actually, recent studies have paid attention to tackling the issue of temporal OOD generalization [53, 58, 61]. They demonstrate that unseen temporal environments contribute to such distribution

*Yang Wang and Zhengyang Zhou are corresponding authors.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-2239

shift. For example, potential urban-hosted events can lead to out-of-distribution traffic state, and unrecorded academic communication between individuals in citation networks may hide new patterns of cooperation. Therefore, the research line relying on environment inference present a promising solution for addressing the temporal OOD issue. This paradigm focuses on inferring underlying environment factors, and utilizing the extracted environment information to enhance the robustness of dynamic graph learning against environment shifts. Although some works have achieved impressive performance [53, 58], there remain limitations in both environment inference and utilization.

The judgment of existing limitations stems from two crucial observations. **Firstly**, unseen environments invariably encompass a wide range of factors, posing challenges in accurately determining their quantities and scales. However, existing methods often rely on a predefined scale environment codebook for inferring unseen environments [53, 58], which may infer unrealistic environment results. **Secondly**, the shift of environments in dynamic graphs fundamentally reflects in the changes of structural associations [48, 65]. A real-world example is that the change of weather alters the future flow of the traffic network by changing human's trajectory. Nevertheless, existing methods often prioritize using the inferred environment as additional information to augment raw features, overlooking capturing the evolving structural associations [53, 58, 61]. Therefore, the existing dynamic graph OOD efforts face issues in both environment extraction and utilization, and the comprehensive solution to address both problems is currently lacking.

To tackle these limitations, we propose an **E**nvironment-**pro**mpted **D**ynamic Graph Learning (EpoD) architecture. **Firstly**, we propose a novel self-prompted learning mechanism to infer underlying environment representation. Given the lacking of environment labels and explicit scaling of environments, our aim is to guide the network generalizing environment factors from historical data in autonomous manner. The practices of language model inference on underlying semantics inspires us to utilize prompt learning to achieve this goal [14, 31]. We propose a self-prompted learning mechanism for spatio-temporal data to infer environment variables from historical data. By designing learnable prompt tokens and an interactive prompt-answer squeezing mechanism, we enable the model to effectively infer the compact and informative environment representations. **Secondly**, we propose a novel Structural Causal Model (SCM) with dynamic subgraph as mediating variable to enhance the adaptability of the network to environment shifts. Different from some approaches that obtain the causal subgraph by partitioning the original graph [61], we design a node-centered subgraph extractor specifically tailored for spatio-temporal data. This design is derived from a profound understanding of dynamic graph that the shift of environments within dynamic graphs invariably result in the changes of these asymmetric correlations between nodes. Our node-centered dynamic subgraphs extractor can capture node-wise asymmetry, where each node has its unique subgraph based on its environment states. **Lastly**, we conduct comprehensive experiments to evaluate the generalizability of EpoD. On the one hand, we perform experiments over multiple cross-domain datasets and introduce a more intricate long-series prediction task. On the other hand, we design an environment-shaded toy dataset, named *EnvST*, to verify the generalization ability of EpoD. The results show that EpoD can precisely perceive environment factors, and the generated dynamic subgraphs are equipped with both generalizability and interpretability. Our **contributions** are summarized as follows:

- We systematically investigate the environment-based efforts to tackle the temporal OOD issue, and observe the limitations of existing works in environment inference and utilization.

- To address existing challenges, we propose a novel **E**nvironment-**pro**mpted **D**ynamic Graph Learning (EpoD) architecture. Specifically, we introduce a self-prompted learning mechanism for spatio-temporal data to infer underlying environment variables without preset scale. For the exploitation of inferred environment factors, we propose a structural causal model with dynamic subgraphs for mediating variables to capture the effect of environment variable shifts on the data distribution. Our work presents a pioneering practice jointly focusing and solving the issue of environment inference and utilization.

- We conduct experiments over multiple cross-domain datasets, including traffics and social networks, to verify the effectiveness of our framework. And a toy dataset is designed to demonstrate the generalizability and interpretability of EpoD.

https://doi.org/10.52202/079017-2239

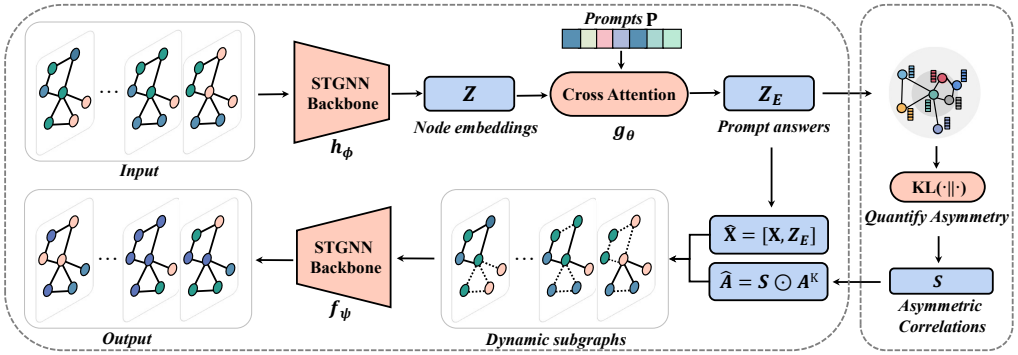


Figure 1: The architecture of EpoD. Left panel: the prediction of future evolution based on historical observations. Right panel: the extraction process of node-centered dynamic subgraph.

2 Background

Preliminaries. We denote $\mathcal{G} = \{\mathcal{G}^t\}_{t=1}^T$ as a dynamic graph across T steps, where $\mathcal{G}^t = (\mathcal{V}^t, \mathcal{X}^t, \mathcal{A}^t)$ represents a snapshot of the graph at step t. The tensor $\mathcal{X}^t \in \mathbb{R}^{N \times D}$ indicates observed features of $N = |\mathcal{V}^t|$ nodes at step t, where D denotes the feature dimension. $\mathcal{A}^t \in \{0,1\}^{N \times N}$ is the adjacency matrix describing the connectivity of graph $\mathcal{G}^t$. Given the historical data $\mathbf{X} = \{\mathcal{G}^t\}_{t=1}^T$, dynamic graph learning aims to predict the future evolution patterns $\mathbf{Y} = \{\mathcal{G}^t\}_{t=T+1}^{T+K}$, where K denotes the number of predicted future time steps. Historical observations $\mathbf{X}$ can be divided into the accessible environment features $\mathbf{X_X}$ and observed labels with evolution patterns $\mathbf{Y_X}$, e.g., volumes in traffic datasets, or links in social networks.

Problem Definition. There is a consensus that unseen environment factor, $\mathbf{E}$, considered as the primary reason for temporal OOD issue [58, 53]. In this work, we focus on capturing the invariant evolution pattern of dynamic graph to tackle the temporal OOD issue.

Spatio-Temporal Graph Forecasting for OOD issue. Our work aims to systematically tackle the challenges in existing environment-centered OOD approaches. Our specific practice in environment inference and utilization is significantly distinct from existing works and has substantial improvements. *On the one hand*, the self-prompted environment inference framework aims to guide the network to adaptively infer environment variables from historical data by using well-designed prompt tokens. Different from existing efforts relying on predefined environment scale [53, 58], our methods advantageously eliminates human biases and ensures accurate extraction of the environment from real historical data. *On the other hand*, we propose a novel SCM with dynamic subgraph as mediating variable. Compared to simply attaching environment embeddings to existing representations, our approach is more consistent with the understanding of dynamic graph data from a causal perspective. Furthermore, unlike some approaches that rely on partition strategies commonly used in static graph learning to extract subgraphs [61], we propose a node-centered dynamic subgraph extraction method that is better suited for spatio-temporal graph scenarios.

Prompt Learning. Prompt learning was initially introduced to address the challenge of data scarcity in language models [7, 12, 11]. By utilizing well-designed prompts, the model can effectively capture a broader space of data distributions and patterns during training, which is better at adapting to input samples from different distributions. For an extended period, the complexity of prompt design has been a hindrance preventing prompt learning from achieving broader applications. In contrast to language data, human beings lack the intuitive cognition of both image and graph data, making it challenging to design interpretable prompts templates. Recently, learnable prompts have also been proven to have superior performance [29, 20]. Therefore, prompt learning is thriving in computer vision research and graph learning [22, 30, 39]. Recently, the efforts of using prompt learning to enhance spatio-temporal prediction is beginning to emerge, including [62] and [60]. The former focuses on multi-attribute forecasting; the latter aims to enhance model generalization ability. However, there is lacking systematical research on prompt learning addressing temporal OOD problems.

https://doi.org/10.52202/079017-2239

3 Methodology

In this section, we introduce a novel **E**nvironment-**pro**mpted **D**ynamic Graph Learning architecture named EpoD. First, we propose a self-prompted learning mechanism to realize the awareness of unseen environment factors. Second, we revisit the winding causal path from environment to graph evolution, and a spatio-temporal learning framework utilizing dynamic subgraph as mediating variables is presented. Last, we provide theoretical analysis of EpoD from causal perspective to interpret its excellent generalization ability.

3.1 Self-prompted learning for environment awareness

Existing methods typically infer environments by predefining the scale of the environment codebook, which introduces human bias and has the potential to cause performance degradation. To tackle this issue, we introduce an environment inference principle that extract underlying environment representations from historical observations without predefining scales. Inspired by the remarkable success of prompt learning in inferring underlying semantics and the multimodal generalization of large language models (LLMs) [49, 35, 39], we propose a self-prompted learning mechanism (SPL) to realize this environment inference principle.

Spatial-specified prompt design. Our SPL focuses on guiding models to effectively extract environment variables from observed data by well-designed prompts. Therefore, the initial challenge we need to address is the design of prompt tokens. Within spatio-temporal graphs, it is common for different nodes to exhibit diverse evolution patterns. Thus, the design of prompts should be specified on spatial aspect to reflect node-wise distinctive evolution state. Given a dynamic graph $\boldsymbol{\mathcal{G}}$, we assign a prompt token $\mathbf{p}_i$ for each node, which is shared across temporal steps. However, the absence of prior knowledge about environment information hinders us from adopting the template-based approach to initialize $\mathbf{p}_i$. Fortunately, learnable prompts have revealed satisfactory performances on capturing hidden mappings [39, 30]. Therefore, we initialize a learnable prompt token $\mathbf{p}_i \in \mathbb{R}^d$ for each $i \in \mathcal{V}$, where d denotes the dimension of latent embedding space. The environment prompt tokens $\mathbf{P}$ for $\boldsymbol{\mathcal{G}}$ is then denoted as,

$$\mathbf{P} = \{\mathbf{p}_i\}_{i=1}^{N} \in \mathbb{R}^{N \times d}. \tag{1}$$

Since $\mathbf{P}$ is specified on the spatial perspective, the node across different temporal snapshots shares the same environment prompt. In the implementation, $\mathbf{P}$ is expanded as a new tensor of $\mathbb{R}^{T \times N \times d}$ by cloning T times.

Prompt-answer mechanism for environment squeezing. The next challenge to address is how to effectively utilize the well-designed prompts to guide the model in extracting underlying environment representations $\mathbf{Z}_E \in \mathbb{R}^{T \times N \times d}$. The premise of achieving this goal lies in profoundly understanding the relationships between the variables in dynamic graph. The ability of Structural Causal Model (SCM) to describe the relationship between variables offers a valuable framework for our analysis [26, 27]. The SCM of dynamic graph includes the temporal environment information $\mathbf{E}$, spatial context $\mathbf{C}$, historical observation $\mathbf{X}$, and future evolving signals $\mathbf{Y}$. Actually, $\mathbf{X}$ denotes historical observations, which is the combination of $\mathbf{X}_\mathbf{X}$ and $\mathbf{Y}_\mathbf{X}$ defined in Sec. 2. This causal model can be formalized as,

$$\mathbb{P}(\mathbf{X}, \mathbf{Y} | \mathbf{E}, \mathbf{C}) = \mathbb{P}(\mathbf{Y} | \mathbf{X}, \mathbf{E}, \mathbf{C}) \mathbb{P}(\mathbf{X} | \mathbf{E}, \mathbf{C}). \tag{2}$$

Therefore, we aim to squeeze environment variables $\mathbf{E}$ from the observable feature $\mathbf{X}_\mathbf{X}$ and the observed labels with evolution patterns $\mathbf{Y}_\mathbf{X}$, which is similar to solving cloze problems. To accomplish this, we design an interactive squeezing mechanism $g_\theta(\cdot)$ to guide the model to squeeze out the underlying environment variables through the interaction of learnable prompts $\mathbf{P}$ and the observable features $\mathbf{X}_\mathbf{X}$. Specifically, we first perform a spatio-temporal network backbone $h_\phi(\cdot)$ to get nodes embedding $\mathbf{Z} = \{\mathbf{z}_1^t, \cdots, \mathbf{z}_N^t\}_{t=1}^T \in \mathbb{R}^{T \times N \times d}$ by taking observable features $\mathbf{X}_\mathbf{X}$ as inputs. Then, $g_\theta(\cdot)$ decodes the learnable prompts $\mathbf{P}$ and encoded embedding $\mathbf{Z}$ to obtain prompt answers $\mathbf{Z}_E$. In implementation, $g_\theta(\cdot)$ consists of three families of learnable parameters, i.e., $W^Q, W^K, W^V \in \mathbb{R}^{T \times d \times d}$. Three hidden state matrices are calculated by,

$$\mathbf{P}^Q = \mathbf{P}W^Q,\ \mathbf{Z}^K = \mathbf{Z}W^K,\ \mathbf{Z}^V = \mathbf{Z}W^V. \tag{3}$$

The prompt answer $\mathbf{Z}_E$ is obtained by,

$$\mathbf{Z}_E = \mathrm{softmax}(\frac{\mathbf{P}^Q (\mathbf{Z}^K)^T}{\sqrt{d}})\mathbf{Z}^V + \boldsymbol{\epsilon}, \tag{4}$$

https://doi.org/10.52202/079017-2239

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Random noise is added to further enhance the robustness of inferred environment representation. Then, the ground-truth $\mathbf{Y_X}$ serves as the learning goal during this squeezing process. We design a tractable objective to squeeze unseen environment representation $\mathbf{Z}_E$,

$$\min_{\phi,\theta,\mathbf{P}} \mathcal{L}_P = -\mathbb{E}[\log \mathbb{P}(\mathbf{Y_X}|\mathbf{X_X}, \mathbf{E})] = \beta \mathbb{E}[\mathrm{KL}(\mathbb{P}_\theta(\mathbf{Z}_E)||\mathbb{P}_\phi(\mathbf{Z}))] - \mathbb{E}[\log \mathbb{P}_{\phi,\theta}(\mathbf{Y_X}|\mathbf{Z}_E)], \quad (5)$$

where $\beta \in [0, 1]$ is a preset hyperparameter, and its sensitivity analysis is provided in Appendix G.1.

- The first term captures the similarity between environment states $\mathbf{Z}_E$ and node embedding $\mathbf{Z}$.
- The second term predicts the evolution rule $\mathbf{Y_X}$ only using inferred unseen environments $\mathbf{Z}_E$.

The objective of Eq. 5 implies that our prompt answers not only reflect the evolution of dynamic graphs but also significantly differ from current available features.

Our design infers unseen environment factors into a continuous space, which remarkably distinctive from previous methods with predefined and discrete environment scale [53, 58]. Our extracted $\mathbf{Z}_E$ can eliminate the bias stemming from inadequate prior knowledge of environment information.

Provable Squeezed Environment Answer. SPL enables the awareness of environment factors via employing environment prompt framework. However, there is still indistinctness about the feasibility of our design. From the perspective of information theory [2, 5], the learning objective of SPL can be restated as,

$$\max_{\phi,\theta,\mathbf{P}} I(\mathbf{Z}_E; \mathbf{Y_X}) - \beta I(\mathbf{Z}_E; \mathbf{X_X}). \quad (6)$$

We will replace Eq. 5 with above Eq. 6 for later proof.

Theorem 3.1. *If there exists a causal relationship between unseen environment pattern $\mathbf{Z}_E^*$ and the label $\mathbf{Y_X}$, $\mathbf{Z}_E^*$ is the optimal result of SPL objective.*

Theorem 3.1 demonstrates the feasibility of SPL and we can always extract additional environment factor if it exist. Detailed proof is provided in Appendix C.

3.2 Spatial-temporal Learning with Dynamic Subgraph

With the well-learned environment answers, how to exploit such representations to achieve spatio-temporal prediction becomes a natural problem. Existing environment-centered approaches tend to directly treat the perceived environment embedding $\mathbf{Z}_E$ as the additional feature. However, from the perspective of data generation, the influence of environments on data distribution shifts is usually reflected in the changes of node-wise correlations. Current methods fail to capture the causal effects of environment variables in dynamic graphs. In this subsection, we introduce a spatio-temporal invariant learning approach using node-centered dynamic subgraph as the mediating variable.

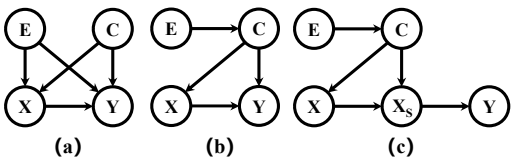


Figure 2: SCMs of dynamic graph. (a) Traditional generation understanding of dynamic graph; (b) Indirect effect of environment factors; (c) Dynamic subgraph as mediating variable.

A winding causal path in dynamic graph. We revisit the role of $\mathbf{E}$ in the spatio-temporal causal model, and propose a novel SCM with dynamic subgraph $\mathbf{X_S}$ as mediating variable as shown in Fig. 2(c). It can be formalized by,

$$\mathbb{P}(\mathbf{Y}, \mathbf{X}|\mathbf{E}, \mathbf{C}) = \mathbb{P}(\mathbf{Y}|\mathbf{X_S})\mathbb{P}(\mathbf{X_S}|\mathbf{X}, \mathbf{C})\mathbb{P}(\mathbf{X}|\mathbf{E}, \mathbf{C}). \quad (7)$$

In contrast to the symmetric correlations between nodes in static graphs, the dependencies between nodes in dynamic graphs are often directional and asymmetric. For example, there is a flow direction between nodes in traffic network and certain node pairs may have different influence or importance to each other in social network. As a result, the shift of environments within dynamic graphs always leads to the changes of these asymmetric correlations between nodes. A typical example is that the change of weather conditions always leads to a shift in the direction of traffic flow instead of bring any new paths. Different from partition-based subgraph learning strategies investigated in static

https://doi.org/10.52202/079017-2239

graphs, we design node-centered dynamic subgraph extractor tailored for dynamic graphs, where each node has its unique subgraph based on its environment states. Such strategy not only captures the impact of environments on internode dependencies, but also facilitates to mine the invariant pattern of spatio-temporal evolution more interpretably.

Dynamic subgraph extraction for environment enhancement. We argue that the asymmetry of environment factor between nodes often leads to clustering effects, such as the asymmetrical importance between the prominent individuals and their followers brings stable connectivity. We leverage the relative entropy (Kullback-Leibler Divergence) to measure such asymmetry. Relative entropy is a metric employed to quantify the disparity between two probability distributions, which exhibits a notable feature of asymmetry, i.e., $\mathrm{KL}(\mathbb{P}||\mathbb{Q}) \neq \mathrm{KL}(\mathbb{Q}||\mathbb{P})$. $\mathrm{KL}(\mathbb{P}||\mathbb{Q})$ quantifies the degree of match when using $\mathbb{P}$ as the reference distribution and approximating it with $\mathbb{Q}$, while $\mathrm{KL}(\mathbb{Q}||\mathbb{P})$ quantifies the degree of match when using $\mathbb{Q}$ as the reference distribution and approximating it with $\mathbb{P}$. Such property of relative entropy offers significant advantages in quantifying asymmetric environment distributions. This is because the asymmetric dependencies of environment distributions also considers the influence of one environment factor on another as the reference basis. However, computing the KL divergence between every pair of nodes $\mathrm{KL}(\mathbf{Z}^t_{E(i,:)}||\mathbf{Z}^t_{E(j,:)})$ is computationally intensive. To this end, we propose linear complexity quantification method using the mean distribution of node-level environment distributions as a proxy. $\mathrm{S}^t_{(i,j)}$ represents the dependence from i on j at time t,

$$\mathrm{S}^t_{(i,j)} = \mathrm{KL}(\bar{\mathbf{Z}}^t_E||\mathbf{Z}^t_{E(i,:)}) \times \mathrm{KL}(\mathbf{Z}^t_{E(j,:)}||\bar{\mathbf{Z}}^t_E), \tag{8}$$

where $\bar{\mathbf{Z}}^t_E = \mathrm{MEAN}(\mathbf{Z}^t_E)$ denotes the mean distribution of node-level environment embedding. This method utilizes $\bar{\mathbf{Z}}^t_E$ as an intermediary to measure the environment dependency from node i to j. The larger of $\mathrm{S}^t_{(i,j)}$ indicates a greater gap of environment difference from node i to j, which reflects the strong dependence. The asymmetric correlation matrix $\mathrm{S}^t \in \mathbb{R}^{N\times N}$ at time t can be calculated as,

$$\mathrm{S}^t = (\boldsymbol{M}^t)^{\mathrm{T}} \cdot \boldsymbol{N}^t. \tag{9}$$

$\boldsymbol{M}^t \in \mathbb{R}^{1\times N}$ and $\boldsymbol{N}^t \in \mathbb{R}^{1\times N}$ are respectively calculated from the two terms in Eq. 8,

$$\boldsymbol{M}^t = [\mathrm{KL}(\bar{\mathbf{Z}}^t_E||\mathbf{Z}^t_{E(1,:)}), \mathrm{KL}(\bar{\mathbf{Z}}^t_E||\mathbf{Z}^t_{E(2,:)}),..., \mathrm{KL}(\bar{\mathbf{Z}}^t_E||\mathbf{Z}^t_{E(N,:)})], \tag{10}$$

$$\boldsymbol{N}^t = [\mathrm{KL}(\mathbf{Z}^t_{E(1,:)}||\bar{\mathbf{Z}}^t_E), \mathrm{KL}(\mathbf{Z}^t_{E(2,:)}||\bar{\mathbf{Z}}^t_E),..., \mathrm{KL}(\mathbf{Z}^t_{E(N,:)}||\bar{\mathbf{Z}}^t_E)]. \tag{11}$$

$\mathrm{S}^t{}_{(:,i)}$ denotes potential nodes centered on node i and $\mathrm{KL}(\cdot||\cdot)$ indicates the Kullback-Leibler divergence. We can extract node-centered L-hop subgraph $\widehat{\mathcal{A}^t} \in \mathbb{R}^{N\times N}$ based on this environment-enhanced correlation matrix S^t,

$$\widehat{\mathcal{A}^t} = \mathrm{S}^t \odot \mathbb{I}((\mathcal{A}^t)^L), \tag{12}$$

where $\odot$ denotes element-wise multiplication of matrices and L is a hyperparameter. L is set to 5 and its sensitivity is discussed in Appendix G.1. $\mathbb{I}(\cdot)$ is an indicator function that assigns a value of 1 to elements in the matrix that are greater than 0, and assigns a value of 0 to the rest. We can get a series of dynamic subgraphs that evolves over time $\widehat{\mathcal{A}} \in \mathbb{R}^{T\times N\times N}$. Meanwhile, we combine the features of each node by concatenating environment answer $\mathbf{Z}^t_E$ to obtain enhanced $\widehat{\mathcal{X}}^t \in \mathbb{R}^{N\times(D+d)}$,

$$\widehat{\mathcal{X}}^t = \mathrm{CONCAT}([\mathcal{X}^t, \mathbf{Z}^t_E]). \tag{13}$$

The historical data enhanced by dynamic subgraph with environment representation is denoted as,

$$\mathbf{X_S} = \{\widehat{\mathcal{G}^t}\}^T_{t=1} = \{\mathcal{V}^t, \widehat{\mathcal{X}}^t, \widehat{\mathcal{A}}^t\}^T_{t=1}. \tag{14}$$

Dynamic graph prediction with generalizability. Finally, f_ψ encodes the dynamic subgraphs $\mathbf{X_S}$ via a spatio-temporal network backbone to predict future dynamic evolution. We obtain the learning objective of EpoD,

$$\min_{\phi,\theta,\mathbf{P},\psi} \mathcal{L} = -\mathbb{E}[\log \mathbb{P}_\psi(\mathbf{Y}|\mathbf{X_S}) - \log \mathbb{P}_{\phi,\theta}(\mathbf{Y_X}|\mathbf{X_X}, \mathbf{P})] - \beta\mathbb{E}[\mathrm{KL}(\mathbb{P}_\theta(\mathbf{Z}_E)||\mathbb{P}_\phi(\mathbf{Z}|\mathbf{X_X}))]. \tag{15}$$

We provide the training process of EpoD in Alg. 1. It is worth noting that our EpoD systematically resolves the limitations of environment inference and environment utilization faced by spatio-temporal invariant learning methods. The two aspects of the design are not independent, but rather tightly coupled. Moreover, our EpoD is pluggable that can be flexibly integrated with various backbones. In

Table 1: The performance of traffic prediction tasks $(12 \rightarrow 24)$ on four real-world datasets. The best results are shown in **bold** and the second best results are underlined.

Model	PEMS08		PEMS04		SD(2019-2020)		GBA(2019-2020)	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
GWNET	19.04±0.9	29.02±1.1	23.12±0.8	38.75±1.3	30.22±2.1	43.65±2.9	31.27±2.6	45.29±2.3
AGCRN	17.30±0.2	27.44±0.6	21.19±0.3	34.65±0.2	26.19±1.2	40.51±1.3	28.74±1.6	43.75±2.0
Z-GCNETs	19.24±0.3	28.40±0.2	22.55±0.5	36.27±0.7	28.21±1.7	41.32±1.8	29.87±1.2	43.11±2.2
DSTAGNN	17.56±0.3	26.29±0.2	21.22±0.7	36.65±0.2	26.34±1.4	41.31±1.6	30.11±2.0	42.99±2.7
STGNCDE	18.41±0.6	27.38±0.3	22.04±0.6	35.39±0.4	27.34±0.9	40.73±1.3	29.21±1.5	43.03±2.4
CaST	17.28±0.3	26.56±0.4	**20.79±0.4**	34.95±0.3	25.38±1.1	39.92±1.6	28.67±1.8	42.23±1.9
EopD(ours)	**16.92±0.2**	**25.66±0.6**	21.12±0.4	**34.02±0.3**	**23.58±1.2**	**38.25±1.4**	**27.26±1.5**	**40.14±1.8**

the experiments of traffic flow and social relationship prediction, we select Adaptive Graph Convolutional Recurrent Network (AGCRN) [3] and Disentangled Dynamic Graph Attention Networks (DDGAN) [61] as our STGNN backbone respectively.

Our approach does not require predictions of future unseen environments. Actually, we argue that the underlying unseen environment factors within historical observations harbors valuable information to guide evolution. Therefore, our focus lies in perceiving historical environment factors and exploiting them appropriately to capture evolution-invariant pattern for prediction. The subsequent experimental discussion can further validate such intuition.

3.3 Causal Interpretation of Dynamic Subgraphs

In this subsection, we provide a deeper understanding of EpoD in the causal theory perspective [28].

The mediating effect in the dynamic graph. The efforts on how temporal environment factors and spatial contexts influence the evolution of dynamics graph has been extensively made. However, even though some studies claim disentanglement of spatial-temporal dependencies, it is acknowledged that true separation may not be fully achieved. Most of them inherently concentrated on exploring the interplay between spatial and temporal dynamics. In fact, some pioneering researches have revealed the temporal evolution mostly stem from the changes over spatial dependencies. To this end, we summarize such indirect influence as mediating effect within dynamic graph, as shown in Fig. 2(b). But according to *the complete mediation effect theorem*, this SCM eliminates the direct effect of temporal variable $\mathbf{E}$ on future spatio-temporal evolution $\mathbf{Y}$. This requires us to contemplate whether $\mathbf{C}$ can serve as a mediation variable. Given the time-varying property of dynamic graph, the only spatial context cannot sufficiently interpret the graphs evolution. Therefore, a mediating variable simultaneously encapsulating spatial dependencies and temporal dynamics is required.

Dynamic subgraphs as mediation variable. The dynamic spatial variations induced by environment factors are essentially rooted in the changes of local dependencies. Thus, a novel SCM is introduced, which employ dynamic subgraph $\mathbf{X_S}$ as the mediation variable as illustrated in Fig. 2(c). Dynamic subgraphs exhibit both temporal and spatial characteristics, also serve as the mediation variable from $\mathbf{X}$ to $\mathbf{Y}$. This design offers us a chance to address distribution shift issue along the practices of causal adjustment [28]. We can observe a back-door path between causal path $\mathbf{X}$ and $\mathbf{Y}$, i.e., $\mathbf{X} \leftarrow \mathbf{C} \rightarrow \mathbf{X_S} \rightarrow \mathbf{Y}$. The backdoor adjustment pattern leveraging *do-calculus* on dynamic subgraph $\mathbf{X_S}$ is,

$$P(\mathbf{Y} = y|do(\mathbf{X} = x)) = \sum_{x_S} P(\mathbf{Y} = y|\mathbf{X} = x, \mathbf{X_S} = x_S)P(\mathbf{X_S} = x_S). \tag{16}$$

In essence, our EpoD can be viewed as employing backdoor adjustments to estimate $P(\mathbf{Y}|do(\mathbf{X}))$ by discovering dynamic subgraphs, where the prompted environment representations support the subgraph discovery process. More discussion is provided in Appendix D.

4 Experiments

4.1 Experiment Setup

We introduce datasets, baselines and experiment settings, and details are leaved in Appendix E.

https://doi.org/10.52202/079017-2239

Datasets. We employ seven cross-domain real-world dynamic graph datasets to evaluate our EpoD. PEMS08 and PEMS04 [34] are classic medium-scale traffic network datasets from California with 5-minute intervals; SD and GBA [24] are newly proposed large-scale traffic network datasets. COLLAB [40] is an academic collaboration dataset comprising papers published in 16 years; Yelp [33] is a business review dataset; ACT [18] shows students' actions on a MOOC platform over 30 days.

Baselines. We compare EpoD with two families of baselines, i.e., six traffic flow prediction models and six social link forecasting methods. Traffic flow prediction models: GWNET [52], AGCRN [3], Z-GCNETs[8], DSTAGNN [19], STGNCDE [9], CaST [53]. Social link forecasting models: DySAT[33], IRM[1], VREx[17], GroupDRO[32], DIDA[58], EAGLE [58].

Experiment settings. In the experiments of traffic flow prediction, our task is to predict the next 24 steps based on historical 12 steps observations ($12 \rightarrow 24$). Moreover, we choose traffic data from the SD and GBA datasets spanning from 2019 to 2020 in order to add the distribution shift scenarios arising from COVID-19, where the training set is composed of data from 2019, while the data from 2020 is divided into a validation set and a test set. The task of social relationship analysis is to exploit past graphs to make link prediction in the next time step. In the training stage, we selectively mask a shifted attribute link from COLLAB, Yelp and ACT to simulate the distribution shift scenario encountered in the real world [61].

4.2 Performance Analysis on Real-world Datasets

Traffic flow prediction. We evaluate our EpoD with baselines based on Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), where lower values of them represent better performance. Tab. 1 shows the results of EpoD on traffic flow data. We have two observations: 1) our EpoD outperforms all baselines on three datasets. The powerful long-sequence prediction ability of EpoD demonstrates that our design is robust to environment perturbations and excels in capturing the evolution patterns of dynamic graph. We also note that CaST [53] obtains suboptimal results on most datasets and even optimal results on PEMS04. This indicates that it is effective to tackle the temporal distribution shift issue by studying environment factors under the guidance of causal theory. 2) EpoD exhibits a more pronounced capability for prediction improvement on two large-scale traffic datasets. It indicates that EpoD is better suited for addressing the distribution shift issue caused by extremely intricate environment perturbations, which is the main challenge posed by large-scale traffic data. We also discuss the interpretability of dynamic subgraphs in Appendix G.3.

Table 2: AUC score (%) of future link prediction task on real-world social relationship datasets. The best results are shown in **bold** and the second best results are underlined.

Model	Collab	Yelp	ACT
DySAT	76.59±0.20	66.09±1.42	66.55±1.21
IRM	75.42±0.87	56.02±16.08	69.19±1.35
VREx	76.24±0.77	66.41±1.87	70.15±1.09
GroupDRO	76.33±0.29	66.97±0.61	74.35±1.62
DIDA	81.87±0.40	75.92±0.90	78.64±0.97
EAGLE	**84.41±0.87**	77.26±0.74	82.70±0.72
EopD(ours)	83.21±0.35	**80.85±0.81**	**83.85±0.52**

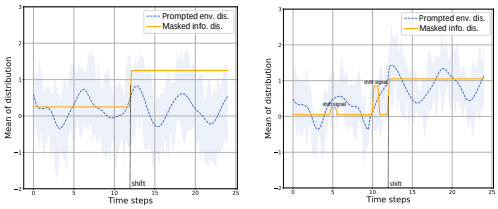
Social link forecasting. Tab. 2 presents the performance of EpoD on social link prediction tasks. Our model outperforms all baselines on two datasets under distribution shifts. We also observe that EAGLE [58] achieves one best performance and other sub-optimal results, which is comparable to our method. It proves that the approaches of perceiving environments can tackle the distribution shifts issue in dynamic graph. Besides, our extracted continuous environment representations are more expressive than the environment factors with pre-defined scales.

4.3 Toy Dataset

We manually design a toy dataset *EnvST* with temporal distribution shift to explore the generalizability of EpoD. The feature of each node in *EnvST* encompasses three components, i.e., [x_A, x_B, x_C], where x_A and x_B represent evolution-causal features but x_B is masked after the data is generated, x_C indicates available evolution-spurious feature. To simulate the temporal distribution shift in the dynamic graph, the training and test dataset of x_A, x_B and x_C are sampled from distributions with significant differences separately. The label of each node on the *EnvST* at t step is activated by updated feature $y_i^t \sim \mathrm{Bern}(\sigma(z_i^t))$, where $\sigma(\cdot)$ is the sigmoid func-

tion. We conduct experiments from the following two aspects: 1) we investigate whether EpoD has the capability to perceive masked environment feature x_B, 2) we study whether EpoD can identify and remove the spurious correlation x_C. More analysis can be found in Appendix E.4.

Powerful perception for unseen environment. Fig. 3(a) and 3(b) show the distribution difference between masked feature x_B and prompted environment feature $\mathbf{Z}_E$, where experiments are conducted on *EnvST* under the scenario of distribution shift. Fig. 3(a) shows our prompted environment feature $\mathbf{Z}_E$ can cover more than half of the shifted features in the future steps. As shown in Fig. 3(b), we observe that our prompted environment variables can effectively cover slight early signal and utilize it to tackle OOD issue.



(a) Sharp temporal distribution shift. (b) Temporal distribution shift with slight signals.

Figure 3: Analysis on the toy dataset.

Robust spurious information identification ability. We then explore whether our EpoD can filter out the disturbance of x_C. Specifically, we have the following experiment design. Consider x_C is sampled from $\mathcal{N}(\mu, I)$, we set $\mu \in [0, 10]$ and record the performance of EpoD under the influence of different spurious information as shown in Fig. 7. We can observe that the fluctuations in prediction performance consistently fall within the acceptable error bounds. Therefore, we can conclude that EpoD have the ability to identify spurious information x_C.

4.4 Ablation Study

We conduct ablation studies from the following two aspects.

Temporal shared learnable prompts. In spatio-temporal graphs, we can observe different nodes always reveal heterogeneous evolution patterns. Thus, we design temporal shared node-wise learnable prompts. In essence, this design is driven by both resource consumption and real-world scenarios. There are still two potentially effective design approaches: a single learnable prompt shared globally (SingleP) and node-private learnable prompts (PrivateP). The former only initializes a globally shared prompt for the dynamic graph $\mathbf{P} \in \mathbb{R}^d$; the latter assigns learnable prompts to each node of each snapshot $\mathbf{P} \in \mathbb{R}^{T \times N \times d}$. To this end, we compare three design methods in terms of accuracy and efficiency. Fig. 4 shows the results of ablation, where the bars represent the time consuming and the lines depict accuracy. We have the following two observations. 1) From the aspects of accuracy, the design of EpoD and PrivateP have similar performance, and SingleP is significantly inferior to them. 2) From the aspects of efficiency, it is understandable that SingleP has the highest training efficiency and PrivateP is the least efficient. The efficiency of EpoD falls between them, yet it remains competitive with SingleP. Therefore, we can conclude that our design stands out as the optimal choice considering both efficiency and accuracy.

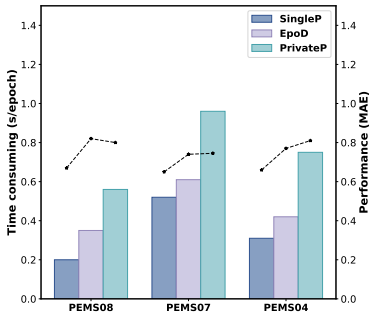


Figure 4: A comparison of learnable prompts design approaches.

The necessity of dynamic subgraph design. We aim to investigate the importance of dynamic subgraphs and explore which subgraph extraction strategy is more suitable for dynamic graph learning. Specifically, we first construct two variants of EpoD, i.e., EpoD-NoSub and EpoD-PartitionSub. EpoD-NoSub is a variant of EpoD that does not utilize dynamic subgraphs for spatiotemporal prediction, which just uses the perceived environment only by incorporating features.

Table 3: Ablation results on dynamic subgraph in EpoD. MAE performance on PEMS08 and PEMS04, AUC(%) score on Yelp.

Model	PEMS08	PEMS04	Yelp
EpoD-NoSub	17.45	21.93	76.34
EpoD-PartitionSub	18.09	22.04	77.35
EopD(ours)	**16.92**	**21.12**	**80.85**

https://doi.org/10.52202/079017-2239

EpoD-PartitionSub means an EpoD variant with partition-based subgraph extraction strategy like static graph learning. We perform ablation experiments on PEMS08, PEMS04 and Yelp, as shown in Tab. 3. First, we observe that the performance difference between EpoD and EpoD-NoSub is substantial, with a maximum gap exceeding 2. It means the utilization of dynamic subgraphs not only enhances interpretability but also is crucial for improving generalization performance. Second, we EpoD had a more pronounced effect than EpoD-PartitionSub. In addition, more experiments show that EpoD is more stable than EpoD-PartitionSub.

4.5 Efficiency Analysis

We analyze the efficiency of EpoD theoretically and practically. We utilize $|V|$ and $|E|$ to denote the number of nodes and edges in the graph, d to represent the dimension of implicit representation, and T to represent the time step of historical observations. The time consumption mainly comprises three components: the spatio-temporal graph aggregation process, the prompt answer process, and the dynamic subgraphs sampling process. The time complexity of the spatio-temporal aggregation is $O(T\cdot(|E|\cdot d+|V|\cdot d^2))$. The prompt answer process primarily involves a cross-attention operation, with a time complexity of $O(T\cdot|V|\cdot d)$. The dynamic subgraphs sampling module implements node-centered sampling, with a time complexity of $O(T\cdot|V|)$. Therefore, the time complexity of EpoD is $O(T\cdot d\cdot|E|+T\cdot(1+d+d^2)\cdot|V|)$. In conclusion , EpoD exhibits linear time complexity concerning the number of nodes and edges, which is competitive with existing dynamic GNNs such as DIDA, EAGLE, and CaST.

Table 4: Efficiency analysis of EpoD (s/epoch).

Dataset	DIDA	EAGLE	EpoD
COLLAB	11.21	12.05	11.84
Yelp	6.89	7.38	7.34
ACT	9.27	9.76	9.59

We also conduct efficiency comparisons of EpoD, DIDA, and EAGLE in COLLAB, Yelp, and ACT datasets, measuring the time taken per epoch (s/epoch). All experiments are run on an NVIDIA A100-PCIE-40GB. Empirically, we observe that the operational efficiency of our method is competitive with existing approaches.

5 Conclusion and Future Work

In this paper, we propose a novel dynamic graph learning framework EpoD to tackle the temporal distribution shift issue by exploiting prompt learning. Inspired by the powerful ability of prompt learning in perceiving underlying semantic and causal associations, we first introduce a self-prompted environment inference mechanism. This approach aims to extract underlying environment variables that potentially influence temporal distribution shift. Subsequently, we propose a novel causal pathway that leverages dynamic subgraphs as mediating variables to effectively utilize the inferred environment embedding. Experiments on real-world datasets and toy examples show that our EpoD effectively improve the dynamic graph learning under temporal shifts, especially boosting the interpretability via dynamic subgraphs.

Limitations. One of the limitation of our work is its strong dependence on graph topology. Specifically, our subgraph discovery strategy essentially is the node filtering based on K-hop neighbors, which can be regarded as subtracting elements from the graph topology. However, the graph topology constructed by distance-based adjacency matrix always lacks adaptability to dynamic changes in the relationships between nodes [53]. In the future, we aim to improve the extraction process of node-centered dynamic subgraphs. We intend to form subgraphs from an empty topology by taking into account both the perceived environment embeddings and the initial distance information.

Acknowledgements

This paper is partially supported by the National Natural Science Foundation of China (No.12227901, No.62072427, No.62402414), Natural Science Foundation of Jiangsu Province (BK.20240460), the Project of Stable Support for Youth Team in Basic Research Field, CAS (No.YSBR-005), Academic Leaders Cultivation Program, USTC, and the grant from State Key Laboratory of Resources and Environmental Information System. We sincerely thank all reviewers for their insightful and constructive comments in improving this paper.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[2] Robert B Ash. *Information theory*. Courier Corporation, 2012.

[3] Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. Adaptive graph convolutional recurrent network for traffic forecasting. *Advances in neural information processing systems*, 33:17804–17815, 2020.

[4] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[5] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR, 2018.

[6] Tanya Y Berger-Wolf and Jared Saia. A framework for analysis of dynamic social networks. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 523–528, 2006.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[8] Yuzhou Chen, Ignacio Segovia, and Yulia R Gel. Z-gcnets: Time zigzags at graph convolutional networks for time series forecasting. In *International Conference on Machine Learning*, pages 1684–1694. PMLR, 2021.

[9] Jeongwhan Choi, Hwangyong Choi, Jeehyun Hwang, and Noseong Park. Graph neural controlled differential equations for traffic forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6367–6374, 2022.

[10] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29, 2016.

[11] Chen Gao, Si Liu, Jinyu Chen, Luting Wang, Qi Wu, Bo Li, and Qi Tian. Room-object entity prompting and reasoning for embodied referring expression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[12] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020.

[13] Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 922–929, 2019.

[14] Yutai Hou, Hongyuan Dong, Xinghao Wang, Bohan Li, and Wanxiang Che. Metaprompting: Learning to learn better prompts. *arXiv preprint arXiv:2209.11486*, 2022.

[15] Qihe Huang, Lei Shen, Ruixin Zhang, Shouhong Ding, Binwu Wang, Zhengyang Zhou, and Yang Wang. Crossgnn: Confronting noisy multivariate time series via cross interaction refinement. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[16] G Jin, Y Liang, Y Fang, J Huang, J Zhang, and Y Zheng. Spatio-temporal graph neural networks for predictive learning in urban computing: A survey. arxiv 2023. *arXiv preprint arXiv:2303.14483*.

[17] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.

 https://doi.org/10.52202/079017-2239

[18] Srijan Kumar, Xikun Zhang, and Jure Leskovec. Predicting dynamic embedding trajectory in temporal interaction networks. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1269–1278, 2019.

[19] Shiyong Lan, Yitong Ma, Weikang Huang, Wenwu Wang, Hongyu Yang, and Pyang Li. Dstagnn: Dynamic spatial-temporal aware graph neural network for traffic flow forecasting. In *International conference on machine learning*, pages 11906–11917. PMLR, 2022.

[20] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

[21] Jianxin Li, Qingyun Sun, Hao Peng, Beining Yang, Jia Wu, and S Yu Philip. Adaptive subgraph neural network with reinforced critical structure mining. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8063–8080, 2023.

[22] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

[23] Yuxuan Liang, Songyu Ke, Junbo Zhang, Xiuwen Yi, and Yu Zheng. Geoman: Multi-level attention networks for geo-sensory time series prediction. In *IJCAI*, volume 2018, pages 3428–3434, 2018.

[24] Xu Liu, Yutong Xia, Yuxuan Liang, Junfeng Hu, Yiwei Wang, Lei Bai, Chao Huang, Zhenguang Liu, Bryan Hooi, and Roger Zimmermann. Largest: A benchmark dataset for large-scale traffic forecasting. *arXiv preprint arXiv:2306.08259*, 2023.

[25] Siqi Miao, Mia Liu, and Pan Li. Interpretable and generalizable graph learning via stochastic attention mechanism. In *International Conference on Machine Learning*, pages 15524–15543. PMLR, 2022.

[26] Judea Pearl. Causal inference in statistics: An overview. 2009.

[27] Judea Pearl. Causal inference. *Causality: objectives and assessment*, pages 39–58, 2010.

[28] Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversity-Press*, 19(2):3, 2000.

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[30] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022.

[31] Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2021.

[32] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[33] Aravind Sankar, Yanhong Wu, Liang Gou, Wei Zhang, and Hao Yang. Dysat: Deep neural representation learning on dynamic graphs via self-attention networks. In *Proceedings of the 13th international conference on web search and data mining*, pages 519–527, 2020.

[34] Chao Song, Youfang Lin, Shengnan Guo, and Huaiyu Wan. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 914–921, 2020.

[35] Alessandro Sordoni, Eric Yuan, Marc-Alexandre Côté, Matheus Pereira, Adam Trischler, Ziang Xiao, Arian Hosseini, Friederike Niedtner, and Nicolas Le Roux. Joint prompt optimization of stacked llms using variational inference. *Advances in Neural Information Processing Systems*, 36, 2024.

[36] Qingyun Sun, Ziying Chen, Beining Yang, Cheng Ji, Xingcheng Fu, Sheng Zhou, Hao Peng, Jianxin Li, and Philip S Yu. Gc-bench: An open and unified benchmark for graph condensation. In *Advances in Neural Information Processing Systems*, 2024.

[37] Qingyun Sun, Jianxin Li, Hao Peng, Jia Wu, Xingcheng Fu, Cheng Ji, and S Yu Philip. Graph structure learning with variational information bottleneck. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4165–4174, 2022.

[38] Qingyun Sun, Jianxin Li, Hao Peng, Jia Wu, Yuanxing Ning, Philip S Yu, and Lifang He. Sugar: Subgraph neural network with reinforcement pooling and self-supervised mutual information mechanism. In *Proceedings of the Web Conference 2021*, pages 2081–2091, 2021.

[39] Xiangguo Sun, Hong Cheng, Jia Li, Bo Liu, and Jihong Guan. All in one: Multi-task prompting for graph neural networks. 2023.

[40] Jie Tang, Sen Wu, Jimeng Sun, and Hang Su. Cross-domain collaboration recommendation. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1285–1293, 2012.

[41] Binwu Wang, Jiaming Ma, Pengkun Wang, Xu Wang, Yudong Zhang, Zhengyang Zhou, and Yang Wang. Stone: A spatio-temporal ood learning framework kills both spatial and temporal shifts. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2948–2959, 2024.

[42] Binwu Wang, Pengkun Wang, Yudong Zhang, Xu Wang, Zhengyang Zhou, Lei Bai, and Yang Wang. Towards dynamic spatial-temporal graph learning: A decoupled perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 9089–9097, 2024.

[43] Binwu Wang, Yudong Zhang, Xu Wang, Pengkun Wang, Zhengyang Zhou, Lei Bai, and Yang Wang. Pattern expansion and consolidation on evolving graphs for continual traffic prediction. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2223–2232, 2023.

[44] Kun Wang, Yuxuan Liang, Xinglin Li, Guohao Li, Bernard Ghanem, Roger Zimmermann, Huahui Yi, Yudong Zhang, Yang Wang, et al. Brave the wind and the waves: Discovering robust and generalizable graph lottery tickets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[45] Kun Wang, Hao Wu, Yifan Duan, Guibin Zhang, Kai Wang, Xiaojiang Peng, Yu Zheng, Yuxuan Liang, and Yang Wang. Nuwadynamics: Discovering and updating in causal spatio-temporal modeling. In *The Twelfth International Conference on Learning Representations*, 2024.

[46] Kun Wang, Hao Wu, Guibin Zhang, Junfeng Fang, Yuxuan Liang, Yuankai Wu, Roger Zimmermann, and Yang Wang. Modeling spatio-temporal dynamical systems with neural discrete learning and levels-of-experts. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[47] Senzhang Wang, Jiannong Cao, and S Yu Philip. Deep learning for spatio-temporal data mining: A survey. *IEEE transactions on knowledge and data engineering*, 34(8):3681–3700, 2020.

[48] Xu Wang, Lianliang Chen, Hongbo Zhang, Pengkun Wang, Zhengyang Zhou, and Yang Wang. A multi-graph fusion based spatiotemporal dynamic learning framework. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 294–302, 2023.

[49] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023.

 https://doi.org/10.52202/079017-2239

[50] Hao Wu, Yuxuan Liang, Wei Xiong, Zhengyang Zhou, Wei Huang, Shilong Wang, and Kun Wang. Earthfarsser: Versatile spatio-temporal dynamical systems modeling in one model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15906–15914, 2024.

[51] Qitian Wu, Hengrui Zhang, Junchi Yan, and David Wipf. Handling distribution shifts on graphs: An invariance perspective. *arXiv preprint arXiv:2202.02466*, 2022.

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

[53] Yutong Xia, Yuxuan Liang, Haomin Wen, Xu Liu, Kun Wang, Zhengyang Zhou, and Roger Zimmermann. Deciphering spatio-temporal graph forecasting: A causal lens and treatment. *arXiv preprint arXiv:2309.13378*, 2023.

[54] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

[55] Chenxiao Yang, Qitian Wu, Qingsong Wen, Zhiqiang Zhou, Liang Sun, and Junchi Yan. Towards out-of-distribution sequential event prediction: A causal treatment. *Advances in neural information processing systems*, 35:22656–22670, 2022.

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

[57] Hao Yuan, Jiliang Tang, Xia Hu, and Shuiwang Ji. Xgnn: Towards model-level explanations of graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 430–438, 2020.

[58] Haonan Yuan, Qingyun Sun, Xingcheng Fu, Ziwei Zhang, Cheng Ji, Hao Peng, and Jianxin Li. Environment-aware dynamic graph learning for out-of-distribution generalization. *arXiv preprint arXiv:2311.11114*, 2023.

[59] Qianru Zhang, Chao Huang, Lianghao Xia, Zheng Wang, Siu Ming Yiu, and Ruihua Han. Spatial-temporal graph learning with adversarial contrastive adaptation. In *International Conference on Machine Learning*, pages 41151–41163. PMLR, 2023.

[60] Qianru Zhang, Lianghao Xia, Zhonghang Li, Siu Ming Yiu, and Chao Huang. Simple yet effective spatio-temporal prompt learning.

[61] Zeyang Zhang, Xin Wang, Ziwei Zhang, Haoyang Li, Zhou Qin, and Wenwu Zhu. Dynamic graph neural networks under spatio-temporal distribution shift. *Advances in Neural Information Processing Systems*, 35:6074–6089, 2022.

[62] Zijian Zhang, Xiangyu Zhao, Qidong Liu, Chunxu Zhang, Qian Ma, Wanyu Wang, Hongwei Zhao, Yiqi Wang, and Zitao Liu. Promptst: Prompt-enhanced spatio-temporal multi-attribute prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3195–3205, 2023.

[63] Zhengyang Zhou, Qihe Huang, Kuo Yang, Kun Wang, Xu Wang, Yudong Zhang, Yuxuan Liang, and Yang Wang. Maintaining the status quo: Capturing invariant relations for ood spatiotemporal learning. 2023.

[64] Zhengyang Zhou, Yang Wang, Xike Xie, Lianliang Chen, and Hengchang Liu. Riskoracle: A minute-level citywide traffic accident forecasting framework. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 1258–1265, 2020.

[65] Zhengyang Zhou, Kuo Yang, Wei Sun, Binwu Wang, Min Zhou, Yunan Zong, and Yang Wang. Towards learning in grey spatiotemporal systems: A prophet to non-consecutive spatiotemporal dynamics. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, pages 190–198. SIAM, 2023.

A Broader impacts

Dynamic graph learning models are widely used to support social development, such as recommendation systems and smart cities. However, with the increasing complexity of data scale and application scenarios, the distribution shifts between training and test data have become a significant obstacle in the development of dynamic graph learning. In light of this, our work aims to address the issue of data distribution shifts in the model and promote the broader application of graph learning in various fields. Therefore, our work aims to develop a model with the out-of-distribution generalization ability and thereby promote the widespread application of dynamic graph learning in various fields.

We ensure the full ethical compliance of our work, and all the datasets we utilize are publicly available. Our work does not involve human subjects and does not introduce any potential negative social impacts or issues related to privacy and fairness.

B Notations

Table 5: Classification accuracies for naive Bayes and flexible Bayes on various data sets.

NOTATIONS	DESCRIPTIONS
$\boldsymbol{\mathcal{G}} = \{\mathcal{G}^t\}_{t=1}^T$	A DYNAMIC GRAPH ACROSS T STEPS
$\mathcal{G}^t = (\mathcal{V}^t, \mathcal{X}^t, \mathcal{A}^t)$	A t-STEP GRAPH WITH THE NODES $\mathcal{V}^t$, FEATURES $\mathcal{X}^t$ AND EDGES $\mathcal{A}^t$
$\mathcal{X}^t \in \mathbb{R}^{N \times D}$	THE FEATURE MATRIX $\mathcal{X}^t$ OF t-STEP GRAPH SNAPSHOT
$\mathbf{P} \in \mathbb{R}^{T \times N \times d}$	LEARNABLE PROMPTS MATRIX OF $\mathbb{R}^{T \times N \times d}$
$\mathbf{X} = (\mathbf{X_X}, \mathbf{Y_X})$	HISTORICAL OBSERVATIONS IN PREDICTION TASKS
$\mathbf{X_S} = \{\mathcal{V}^t, \widehat{\mathcal{X}}^t, \widehat{\mathcal{A}}^t\}_{t=1}^T$	HISTORICAL OBSERVATIONS REPRESENTED BY DYNAMIC SUBGRAPHS
$\mathbf{E}$	TEMPORAL ENVIRONMENT VARIABLE IN PREDICTION TASKS
$\mathbf{Y}$	THE FUTURE EVOLUTION TREND IN PREDICTION TASKS
$\mathbf{X_X}$	THE OBSERVABLE FEATURE OF HISTORICAL OBSERVATIONS
$\mathbf{Y_X}$	THE OBSERVED LABELS WITH EVOLUTION PATTERNS
$\mathbf{Z}_E$	CONTINUOUS FEATURES OF ENVIRONMENT VARIABLES
$\mathbf{Z}$	NODE EMBEDDING BY ENCODING ORIGINAL FEATURES
$\mathbf{C}$	SPATIAL VARIABLE IN STRUCTURAL CAUSAL MODEL
$g_\theta(\cdot)$	A CROSS-ATTENTION NETWORK
$h_\phi(\cdot)$	A SPATIO-TEMPORAL NETWORK BACKBONE
$f_\psi(\cdot)$	A SPATIO-TEMPORAL NETWORK BACKBONE IN FINAL PREDICTION STAGE

C Detailed Proof of Theorem 3.1

Lemma 3.1. (*The Chain Rule for Mutual Information* [2]) For any set of random variables X, Y, and Z, the chain rule is expressed as

$$I(X; Y, Z) = I(X; Y) + I(X; Z|Y), \tag{17}$$

where $I(X; Y, Z)$ denotes the mutual information between X, Y, and Z, $I(X; Y)$ represents the mutual information between X and Y, and $I(X; Z|Y)$ represents the conditional mutual information between X and Z given Y. The rule signifies that the overall mutual information in the system can be decomposed into two components: the mutual information between X and Y, and the conditional mutual information between X and Z given Y. It reflects how information is transmitted and shared within complex systems.

Theorem 3.1. If there exists a causal relationship between unseen environment pattern $\mathbf{Z}_E^*$ and the label $\mathbf{X_X}$, $\mathbf{Z}_E^*$ is the optimal solution of SPL objective.

Proof. Consider *The Chain Rule for Mutual Information*, we derive the following equation:

$$I(\mathbf{Z}_E; \mathbf{Y_X}) = I(\mathbf{Y_X}; \mathbf{X_X}, \mathbf{Z}_E) - I(\mathbf{X_X}; \mathbf{Y_X}|\mathbf{Z}_E) \tag{18}$$

https://doi.org/10.52202/079017-2239

$$I(\mathbf{Z}_E; \mathbf{X}_\mathbf{X}) = I(\mathbf{X}_\mathbf{X}; \mathbf{Z}_E, \mathbf{Y}_\mathbf{X}) - I(\mathbf{X}_\mathbf{X}; \mathbf{Y}_\mathbf{X} | \mathbf{Z}_E) \tag{19}$$

We then get $I(\mathbf{Z}_E; \mathbf{Y}_\mathbf{X}) - \beta I(\mathbf{Z}_E; \mathbf{X}_\mathbf{X}) = I(\mathbf{Y}_\mathbf{X}; \mathbf{X}_\mathbf{X}, \mathbf{Z}_E) - (1-\beta) I(\mathbf{X}_\mathbf{X}; \mathbf{Y}_\mathbf{X} | \mathbf{Z}_E) - \beta I(\mathbf{X}_\mathbf{X}; \mathbf{Z}_E, \mathbf{Y}_\mathbf{X})$. Thus, optimizing Eq. 6 is to maximize $I(\mathbf{Y}_\mathbf{X}; \mathbf{X}_\mathbf{X}, \mathbf{Z}_E)$ and minimize $(1-\beta) I(\mathbf{X}_\mathbf{X}; \mathbf{Y}_\mathbf{X} | \mathbf{Z}_E) + \beta I(\mathbf{X}_\mathbf{X}; \mathbf{Z}_E, \mathbf{Y}_\mathbf{X})$. Next, we investigate whether $\mathbf{Z}_E^*$ is the result of optimization.

(i) Maximizing $I(\mathbf{Y}_\mathbf{X}; \mathbf{X}_\mathbf{X}, \mathbf{Z}_E)$ reflects the combination of $\mathbf{X}_\mathbf{X}$ and $\mathbf{Z}_E$ can make optimal prediction. Since $\mathbf{X}_\mathbf{X}$ is constant, $\mathbf{Z}_E = \mathbf{Z}_E^*$ ensures that the features used for prediction have the greatest overlap with the ground-truth.

(ii) Since $I(\mathbf{X}_\mathbf{X}; \mathbf{Y}_\mathbf{X} | \mathbf{Z}_E) \geq 0$ and $I(\mathbf{X}_\mathbf{X}; \mathbf{Z}_E, \mathbf{Y}_\mathbf{X}) \geq 0$. If $\beta \in [0, 1]$, the lower bound of $(1 - \beta) I(\mathbf{X}_\mathbf{X}; \mathbf{Y}_\mathbf{X} | \mathbf{Z}_E) + \beta I(\mathbf{X}_\mathbf{X}; \mathbf{Z}_E, \mathbf{Y}_\mathbf{X})$ is 0. Next, we prove our SPL can achieve $I(\mathbf{X}_\mathbf{X}; \mathbf{Y}_\mathbf{X} | \mathbf{Z}_E^*) = 0$ and $I(\mathbf{X}_\mathbf{X}; \mathbf{Z}_E^*, \mathbf{Y}_\mathbf{X})) = 0$ in detail. $I(\mathbf{X}_\mathbf{X}; \mathbf{Y}_\mathbf{X} | \mathbf{Z}_E^*)$ represents the conditional mutual information between $\mathbf{X}_\mathbf{X}$ and $\mathbf{Y}_\mathbf{X}$ given $\mathbf{Z}_E^*$. Because we manually add random noise $\boldsymbol{\epsilon}$ at Eq. 4, our model can be viewed as a map $\mathbf{Y}_\mathbf{X} = f(\mathbf{Z}_E^*) + \boldsymbol{\epsilon}$. Obviously, $\boldsymbol{\epsilon}$ is independent of $\mathbf{Y}_\mathbf{X}$. We can get $I(\mathbf{X}_\mathbf{X}; \mathbf{Y}_\mathbf{X} | \mathbf{Z}_E^*) = 0$. $I(\mathbf{X}_\mathbf{X}; \mathbf{Z}_E^*, \mathbf{Y}_\mathbf{X})) = 0$ is also proven the independence between $\mathbf{X}_\mathbf{X}$ and $\mathbf{Z}_E^*$.

We have proved it.

D Causal Interpretation of Dynamic Subgraphs

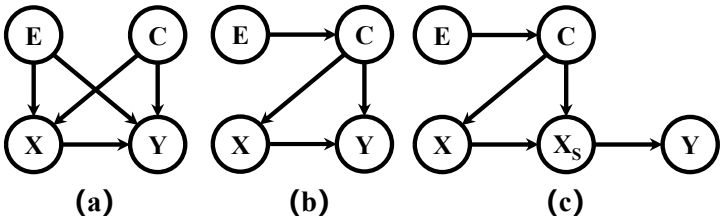


Figure 5: SCMs of dynamic graph. (a) Traditional generation understanding of dynamic graph; (b) Indirect effect of environment factors; (c) Dynamic subgraph as mediating variable.

With the causal theory [28], we can build a coherent progression for the proposal of EpoD.

The Structural Causal Model in the dynamic graph. The Structural Causal Model (SCM) in the dynamic graph fosters a deeper understanding of the generation process of spatio-temporal graphs. Many methodologies have achieved impressive performance in addressing the issue of spatial-temporal distribution shifts. The widely utilized SCM is shown in Fig. 5(a).

- $\mathbf{E} \rightarrow \mathbf{X} \leftarrow \mathbf{C}$. The historical observation $\mathbf{X}$ consists of two parts: temporal environment variable $\mathbf{E}$ and spatial context $\mathbf{C}$. In many works, these two aspects are often treated as disjoint factors and discussed decouplingly, i.e, $\mathbf{E} \perp \mathbf{C}$.
- $\mathbf{E} \rightarrow \mathbf{Y} \leftarrow \mathbf{C}$. Temporal environment variable $\mathbf{E}$ and spatial context $\mathbf{C}$ also determine the future evolution trend $\mathbf{Y}$. Most studies keep this assumption that historical observations $\mathbf{X}$ and future evolution $\mathbf{Y}$ are influenced by the same $\mathbf{E}$ and $\mathbf{C}$. However, this is not the case in the real-world dynamic graphs, giving rise to the distribution shift issue.
- $\mathbf{X} \rightarrow \mathbf{Y}$. Historical observations are useful for predicting future evolution. A thorough understanding of historical observations serves as a crucial foundation for exploring the invariant evolution model of dynamic graphs.

The mediating effect in the dynamic graph. The efforts on how temporal environment factors and spatial contexts influence the evolution of dynamics graph has been extensively made. However, even though some studies claim disentanglement of spatial-temporal dependencies, it is acknowledged that true separation may not be fully achieved. Most of them inherently concentrated on exploring the interplay between spatial and temporal dynamics. In fact, some pioneering researches have revealed the temporal evolution mostly stem from the changes over spatial dependencies. To this end, we further summarize such indirect influence as mediating effect within dynamic graph, as shown in Fig. 5(b).

- $\mathbf{E} \rightarrow \mathbf{C} \rightarrow \mathbf{X}$. The historical observation $\mathbf{X}$ is seen as the evolution process of spatial context $\mathbf{C}$, where $\mathbf{C}$ covers the temporal environment information $\mathbf{E}$.
- $\mathbf{X} \leftarrow \mathbf{C} \rightarrow \mathbf{Y}$. Spatial context $\mathbf{C}$ directly determine current observations $\mathbf{X}$ and future evolution trend $\mathbf{Y}$.

But according to *the complete mediation effect theorem*, this SCM eliminates the direct effect of temporal variable $\mathbf{E}$ on future spatio-temporal evolution $\mathbf{Y}$. This necessitates us to contemplate whether $\mathbf{C}$ has the capacity to serve as a mediation variable. Given the time-varying property in dynamic graph, the only spatial context cannot sufficiently interpret the temporal evolution of graphs. Therefore, a mediating variable simultaneously encapsulating spatial dependencies and temporal dynamics is required.

Dynamic subgraphs as mediating variables. The dynamic spatial variations induced by environment factors are essentially rooted in the changes of local dependencies. Thus, a novel SCM is introduced, which employ dynamic subgraph $\mathbf{X_S}$ as the mediation variable as illustrated in Fig. 5(c).

- $\mathbf{X} \leftarrow \mathbf{C} \rightarrow \mathbf{X_S} \rightarrow \mathbf{Y}$. Dynamic subgraph $\mathbf{X_S}$ is more abundant than the spatial context $\mathbf{C}$. In other words, dynamic subgraph $\mathbf{X_S}$ utilizes substructures to encompass spatial information $\mathbf{C}$, capturing temporal environment factors $\mathbf{E}$ through dynamic evolutive subgraphs.

Dynamic subgraphs exhibit both temporal and spatial characteristics, also serve as the mediation variable from $\mathbf{X}$ to $\mathbf{Y}$. This design offers us a chance to address distribution shift issue along the practices of causal adjustment [28]. We can observe a back-door path between causal path $\mathbf{X}$ and $\mathbf{Y}$, i.e., $\mathbf{X} \leftarrow \mathbf{C} \rightarrow \mathbf{X_S} \rightarrow \mathbf{Y}$. The backdoor adjustment pattern leveraging *do-calculus* on dynamic subgraph $\mathbf{X_S}$ is,

$$P(\mathbf{Y} = y|do(\mathbf{X} = x)) = \sum_{x_S} P(\mathbf{Y} = y|\mathbf{X} = x, \mathbf{X_S} = x_S)P(\mathbf{X_S} = x_S) \tag{20}$$

In essence, our EpoD can be viewed as employing backdoor adjustments to estimate $P(\mathbf{Y}|do(\mathbf{X}))$ by discovering dynamic subgraphs, where the prompted environment supports the subgraph discovery process.

E Experiment Details

E.1 Datasets

Our experimental design included the selection of seven real-world dynamic graph datasets from two distinct domains. The detailed statistics of the datasets are as shown in Tab. 6. We select a shifted link attribute from COLLAB, Yelp and ACT datasets respectively to simulate the distribution shift scenario in the real world. The shifted attribute links become accessible only during the out-of-distribution (OOD) testing stage. This scenario is more practical and challenging in real-world situations, as the model cannot capture any information about the filtered links during training and validation.

PEMS08 [34] is collected from the Caltrans Performance Measurement System (PeMS), which records the real traffic network flow data from 07/01/2016 to 08/31/2016. It delineates a dynamic graph data of a traffic network with 170 sensors across 17,856 steps. Among the known traffic datasets, it falls into the category of small-scale dataset.

PEMS04 [34] records the real traffic network flow data from 01/01/2018 to 02/28/2018. It describes a dynamic graph data of a traffic network with 307 sensors across 16,992 steps. It belongs to a medium-scale traffic dataset.

SD [24] is a sub-dataset of the large-scale dataset CA proposed by [24]. It comprises traffic flow data recorded by 716 sensors in San Diego county from 01/01/2017 to 12/31/2021.

GBA [24] is a larger traffic dataset than SD, which is also a sub-dataset of the large-scale dataset CA. It contains traffic flow data provided by 2,352 sensors in 11 counties situated in the Greater Bay Area from 01/01/2017 to 12/31/2021.

 https://doi.org/10.52202/079017-2239

COLLAB [40] is an academic collaboration dataset comprising papers published between 1990 and 2006, spanning 16 graph snapshots. In this dataset, nodes represent authors, and edges represent co-authorship relationships. The edges include five attributes based on co-authored publications: "Data Mining", "Database", "Medical Informatics", "Theory" and "Visualization".

Yelp [33] is a dataset containing customer reviews on businesses. In this dataset, nodes represent customers and businesses, while edges capture review behaviors. The edges are associated with five attributes based on business categories: "Pizza", "American (New) Food", "Coffee & Tea", "Sushi Bars" and "Fast Food".

ACT [18] characterizes student interactions on a MOOC platform over a span of one month, consisting of 30 graph snapshots. In this dataset, nodes represent students or the targets of actions, while edges signify various student actions.

Table 6: Statistics of the real-world dynamic graph datasets.

DATASET	# NODES	# EDGES	# GRAPH SNAPSHOTS	TEMPORAL INTERVAL
PEMS08	170	276	17,856	5 MINUTES
PEMS04	307	338	16,992	5 MINUTES
SD	716	17,319	525,888	5 MINUTES
GBA	2,352	61,246	525,888	5 MINUTES
COLLAB	23,035	151,790	16	1 YEAR
YELP	13,095	65,375	24	1 MONTH
ACT	20,408	202,339	30	1 DAY

E.2 Detailed Implementation

Algorithm 1: The training process of EpoD

Input: historical dynamic graph data $X = \{\mathcal{G}^t\}_{t=1}^T$
Initial: dynamic graph encoders h_ϕ and f_ψ, cross-attention mechanism decoder g_θ, learnable prompt $\mathbf{P}$, the number of epochs K
for $i = 1$ **to** K **do**
  Environment prompt stage:
  $\mathbf{Z}_E = g_\theta(\mathbf{P}, \mathbf{Z}) + \boldsymbol{\epsilon},\ \mathbf{Z} = h_\phi(\mathbf{X_X})\ \ (\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}))$
  $\hat{\mathbf{Y}}_\mathbf{X} = \mathrm{Linear}(\mathbf{Z}_E)$
  Environment utilization stage:
  $\mathrm{S}^t = (\boldsymbol{M}^t)^{\mathrm{T}} \cdot \boldsymbol{N}^t$ as shown in Eq. 9
  $\widehat{\mathcal{A}^t} = \mathrm{S}^t \odot (\mathcal{A}^t)^K, \quad \widehat{\mathcal{X}} = \mathrm{CONCAT}([\mathcal{X}, \mathbf{Z}_E])$
  $\mathbf{X_S} = \{\widehat{\mathcal{G}^t}\}_{t=1}^T = \{\mathcal{V}^t, \widehat{\mathcal{X}}^t, \widehat{\mathcal{A}}^t\}_{t=1}^T$
  $\widehat{\mathbf{Y}} = f_\psi(\mathbf{X_S})$
  Optimizing:
  $\min_{\phi,\theta,\mathbf{P},\psi} \mathcal{L} = -\mathbb{E}[\log \mathbb{P}_\psi(\mathbf{Y}|\mathbf{X_S}) + \log \mathbb{P}_{\phi,\theta}(\mathbf{Y_X}|\mathbf{X_X}, \mathbf{P})] + \beta\mathbb{E}[\mathrm{KL}(\mathbb{P}_\theta(\mathbf{Z}_E)||\mathbb{P}_\phi(\mathbf{Z}|\mathbf{X_X}))]$
end for
Return h_ϕ, f_ψ, g_θ and $\mathbf{P}$

We implement our EpoD with PyTorch 1.11.0 on a server with NVIDIA A100-PCIE-40GB. The detailed training process is shown in Alg. 1. All experiments are repeated with 10 different random seeds of [1,2,3,4,5,6,7,8,9,10]. The reported results include the mean and standard deviation obtained from these 10 runs.

Traffic flow prediction. In the experiments of traffic flow prediction, our task is to predict the next 24 steps based on the historical 12 steps observations ($12 \rightarrow 24$). Besides, we choose traffic data from the SD and GBA datasets from 2019 to 2020 in order to add the distribution shift scenarios arising from COVID-19. The selected spatio-temporal graph neural network backbone is Adaptive Graph Convolutional Recurrent Network (AGCRN) [3].

Social link prediction. The task of social relationship analysis is to exploit past graphs to make link prediction in the next time step. Following the measures of [61], we introduce perturbations to test data to simulate the scenario of distribution shift in those datasets. Specifically, we select Data Mining and Pizza as the shifted attributes in COLLAB and Yelp. For dataset ACT, we employ K-Means to cluster the action features into five categories and randomly select a certain category (the 5th cluster) of edges as the shifted attribute. Besides, Disentangled Dynamic Graph Attention Networks (DDGAN) [61] is chosen as our spatio-temporal graph neural network backbone.

E.3 Metrics

We utilize Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to assess the performance of our EpoD and baselines. Both metrics quantify the error between model predictions and actual observations in regression tasks. A smaller value for these metrics indicates better model performance. MAE is less sensitive to outliers due to its use of absolute differences, while RMSE is more sensitive to large errors due to its use of squared differences. Given the actual observation Y_i and the corresponding predicted value $\hat{Y}_i$ for n samples, two metrics are calculated as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i| \tag{21}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}. \tag{22}$$

E.4 Toy dataset

We manually design a toy dataset *EnvST* with temporal distribution shift to explore the generalizability of EpoD. Tab. 7 describes the statistical information of *EnvST*. Specifically, *EnvST* illustrates the evolution sequence of a graph with 100 nodes across 1000 steps, where the topology of each snapshot does not change over time. The feature of each node in *EnvST* encompasses three components, i.e., [x_A, x_B, x_C], where x_A and x_B represent evolution-causal features but x_B is masked after the data is generated, x_C indicates available evolution-spurious feature. To simulate the temporal distribution shift in the dynamic graph, the training and test dataset of x_A, x_B and x_C are sampled from distributions with significant differences separately. The label of each node on the *EnvST* at t step is activated by updated feature $y_i^t \sim \mathrm{Bern}(\sigma(z_i^t))$, where $\sigma(\cdot)$ is the sigmoid function.

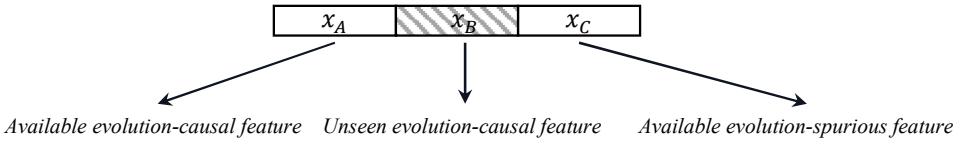


Figure 6: Feature description of the toy dataset.

Table 7: Statistics of the toy dataset.

NOTATION	DESCRIPTION	DIMENSION
N	THE NUMBER OF NODES IN THE GRAPH	100
T	THE NUMBER OF EVOLUTION STEPS OF DYNAMIC GRAPHS	1000
x_A	AVAILABLE EVOLUTION-CAUSAL FEATURE	3
x_B	UNSEEN EVOLUTION-CAUSAL FEATURE	3
x_C	AVAILABLE EVOLUTION-SPURIOUS FEATURE	3
y	EVOLUTION LABEL	1

As shown in Fig. 6, the feature of each node in *EnvST* encompasses three components, i.e., [x_A, x_B, x_C]. x_A and x_B represent evolution-causal features but x_B is masked after the data is generated,

https://doi.org/10.52202/079017-2239

x_C indicates available evolution-spurious feature. To simulate the temporal distribution shift in the dynamic graph, the training and test dataset of x_A, x_B and x_C are sampled from distributions with significant differences separately. The label of each node on the *EnvST* at t step is activated by updated feature $y_i^t \sim \text{Bern}(\sigma(z_i^t))$, where $\sigma(\cdot)$ is the sigmoid function. The feature is updated by aggregating the neighbor information of the last k time steps,

$$z_i^t = \sum_{l=t,\ldots,t-k} \sum_{j \in \mathcal{N}_i} \text{Aggregate}(x_A^l(j), x_B^l(j)). \tag{23}$$

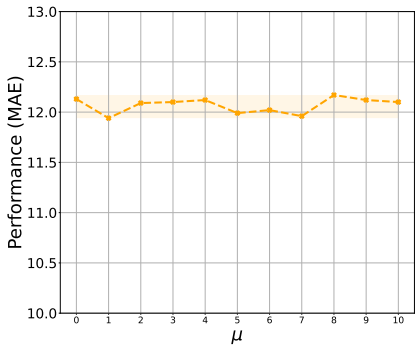


Figure 7: Performance is influenced by spurious information.

We conduct experiments from the following two aspects: 1) we investigate whether EpoD has the capability to perceive masked environment feature x_B, 2) we study whether EpoD can identify and remove the spurious correlation x_C.

Powerful perception for unseen environment. Fig. 3(a) and 3(b) show the distribution difference between masked feature x_B and prompted environment feature $\mathbf{Z}_E$, where experiments are conducted on *EnvST* under the scenario of distribution shift. There is no doubt that $\mathbf{Z}_E$ can match the masked x_B in the steady historical observation sequences, before *shift point* in these figures. More importantly, we focus on the perception ability of prompt learning after the distribution shift occurs. First, we categorize the scenarios of temporal distribution shift into two types: sharp shifts and shifts with signals. The former indicates that any distribution shift signal can not be obtained from historical sequences. Fig. 3(a) shows our prompted environment feature $\mathbf{Z}_E$ can still cover more than half of the shifted features in the future steps. The latter illustrates the signals of distribution shift has begun to emerge slightly in historical observations, which is a commonplace scenario in the real world. For example, those individuals who were used to planning purchased warm clothes in advance before the temperatures plummeted. As shown in Fig. 3(b), we observe that our prompted environment variables can effectively cover slight early signal and utilize it to tackle OOD issue.

Robust spurious information identification ability. We then explore whether our EpoD can filter out the disturbance of x_C. Specifically, we have the following experiment design. Consider x_C is sampled from $\mathcal{N}(\mu, I)$, we set $\mu \in [0, 10]$ and record the performance of EpoD under the influence of different spurious information as shown in Fig. 7. We can observe that the fluctuations in prediction performance consistently fall within the acceptable error bounds. Therefore, we can conclude that EpoD have the ability to identify spurious information x_C.

F More Realted Works

Dynamic Graph Learning. Graph Neural Networks (GNNs) [10, 37, 54, 44, 36] and Sequence Neural Networks (SNNs) [56, 15] have been extensively studied and have achieved great success in real-world tasks. The GNN models we often use are GCN, GIN, PNA, etc; SNNs include LSTM, RNN, TCN, etc. Therefore, spatio-temporal graph learning models have been widely studied in recent years [50, 46, 45, 41, 43, 42]. Based on varying interpretations of the correlation between temporal and spatial information, the current works are undertaken along two research lines. One claim is to study temporal and spatial information in a decoupled manner, which is potentially present in most current works [3, 19, 61]. Due to the intricate temporal and spatial relationships in reality, they

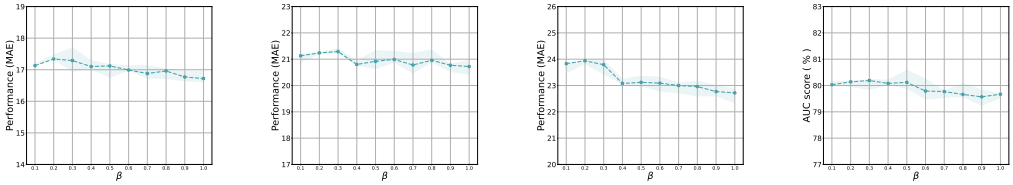
often fail to offer sufficient interpretability and generalizability. The other one argues that spatial context is influenced by the temporal information [65, 53]. These methods depict spatiotemporal scenes that align more closely with the complexities of the real world. Moreover, this also presents a feasible approach to tackle the distribution shift issue arising from the changes of temporal environment. Our work falls into advancing the latter research line by exploiting causal structure model.

Subgraph Learning. Subgraph learning, with its robust causal interpretability, has achieved remarkable success in static graph applications, such as molecular property prediction and social network analysis [25, 57, 38, 21]. In essence, it captures the inductive bias inherent in graph data that local dependencies are invariant patterns predicting ground-truth. We argue that dynamic subgraphs may exhibit a similar inductive bias spatio-temporal evolution. But dynamic subgraphs remain an unexplored area with no existing studies. To this end, this paper introduces a dynamic subgraph learning mechanism to address the issue of temporal distribution shift resulting from the changes of environment factors.

G Additional Results

G.1 Hyperparameter Sensitivity Analysis

We analyze the sensitivity of the hyperparameter β in Eq. 5, which functions as the trade-off for the loss in Eq. 5. The value range of β is $[0, 1]$. We perform experiments on four real-world datasets, i.e, PEMS08, PEMS04, SD, Yelp, and present the results in Fig 8. The results show that the sensitivity of the prediction results to β is not very drastic. But the performances of EpoD on four datasets are the best when $\beta \in [0.2, 0.5]$. Therefore, we set $\beta = 0.2$ is in our implementation.



(a) PEMS08. (b) PEMS04. (c) SD. (d) Yelp.

Figure 8: Sensitivity analysis of the hyperparameter β on four real-world datasets.

We then analyze the sensitivity of hyperparameter L in Eq. 12. As shown in Tab. 8, we study the performance of EpoD when L is set to $[1, 10]$. We observe that increasing the value of L tends to improve its performance. However, it is important to note that this improvement comes at the cost of increased time consumption. Therefore, we set $L = 5$ as a trade-off between performance and time consumption.

Table 8: The performance (MAE) of different L in Eq. 12 on PEMS08.

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
EopD	18.87	18.21	17.76	17.25	17.43	17.13	16.92	16.65	16.58	16.96

G.2 Backbone Sensitivity Analysis

Our EpoD essentially provides a solution for temporal distribution shits issue in dynamic graphs. The most prominent characteristic of EpoD is the pluggability, which denotes that we can be applied to numerous existing backbones. In the task of traffic flow prediction, we explore the performance associated with the selection of different models as backbones as shown in Tab. 9. We get the following two **obs**ervations.

Obs 1. The EpoD-enhanced version consistently shows a significant improvement over the raw backbone. On large-scale datasets with distributional shits, EpoD tends to exhibit more substan-

https://doi.org/10.52202/079017-2239

Table 9: The MAE performance of EpoD with three different backbones ($12 \rightarrow 24$).

MODEL	PEMS08	PEMS04	SD(2019-2020)	GBA(2019-2020)
ASTGCN [13]	19.34	22.89	28.36	32.58
EOPD+ASTGCN	17.75	21.78	26.72	28.76
DSTAGNN [19]	17.56	21.22	26.34	30.11
EPOD+DSTAGNN	<u>17.21</u>	**20.76**	<u>24.89</u>	<u>27.89</u>
AGCRN [3]	17.30	21.19	26.19	28.74
EOPD+AGCRN	**16.92**	<u>21.12</u>	**23.58**	**27.26**

tial performance improvements. This highlights the effectiveness of our EpoD in addressing the distribution shift issue.

Obs 2. The expressive capacity of the backbone directly influences the predictive ability of the model enhanced by EpoD. Since the performance of AGCRN is already excellent, the enhancements introduced by EpoD often result in optimal results. This underscores the importance of selecting a proficient backbone model for forecasting.

G.3 Interpretable Dynamic Subgraphs

In this subsection, we explore the interpretability of dynamic subgraphs under real-world scenarios. In recent years, the most notable temporal-distribution shift phenomenon is the outbreak of COVID-19. The introduction of LargeST [24] provides a chance for us to study this distribution shift of traffic flow under COVID-19. We choose a local sensors network in GBA dataset and reconstruct its evolution data at monthly intervals from 2019 ($t = 1, ..., T$) to 2020 ($t = T + 1, ..., T + k$). Then, we apply our EpoD trained on 2019 data to partition each snapshot into a bag of subgraphs, as shown in the top panel of Fig. 9. It is evident that the nodes tend to suppress communication among themselves when the distribution is shifting, and this phenomenon is particularly pronounced around the three blue nodes. The bottom panel of Fig. 9 visualizes the ground-truth of this sequence, which aligns with the information reflected in our dynamic subgraphs. This suggests that our EpoD exhibit sensitivity to environment changes.

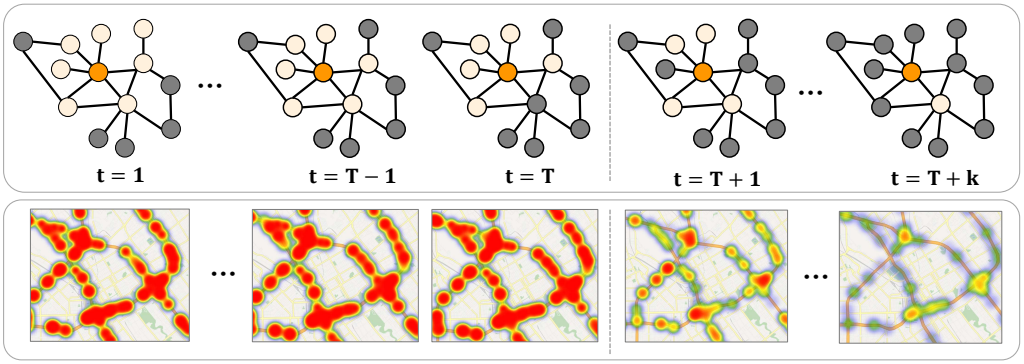


Figure 9: The interpretability of dynamic subgraphs within real-world scenarios.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction of this paper clearly outline our contributions and important assumptions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Our manuscript contains "Limitations and Future Works" section in Sec. 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

https://doi.org/10.52202/079017-2239

Justification: We provide a comprehensive proof of our proposed theory.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We release the code using anonymous links.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We release the code using anonymous links.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so No is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We describe our experiment setup in detail.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The quantified results we provide are statistical results (mean and standard deviation), obtained from conducting experiments with 10 different random seeds.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

https://doi.org/10.52202/079017-2239

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the resources (type and number of GPUs) that we conduct our experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have already read the NeurIPS Code of Ethics

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss both potential positive societal impacts and negative societal impacts of our work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our effort aim to address unresolved questions in the field, and all datasets we used are public datasets.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We thoroughly introduce all works that are related to our research, and carefully check the original license.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

https://doi.org/10.52202/079017-2239

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We have released the license of our code.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.