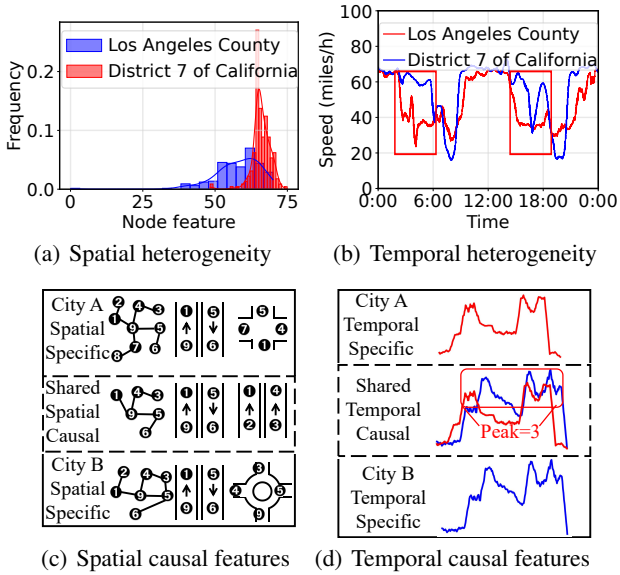
Causality-inspired Federated Learning for Dynamic Spatio-Temporal Graphs

Yuxuan Liu[1], Wenchao Xu[2], Haozhao Wang[3*], Zhiming He[1], Zhaofeng Shi[1], Chongyang Xu[1], Peichao Wang[1], Boyuan Zhang[1]

[1] School of Information and Communication Engineering, University of Electronic Science and Technology of China, China
[2] Division of Integrative Systems and Design, Hong Kong University of Science and Technology, China
[3] School of Computer Science and Technology, Huazhong University of Science and Technology, China
{eie.yuxuan.liu, zfshi, cyxu, peichaowang, boyuanzhang}@std.uestc.edu.cn, wenchao.xu@polyu.edu.hk, hz_wang@hust.edu.cn, zmhe@uestc.edu.cn

Abstract

Federated Graph Learning (FGL) has emerged as a powerful paradigm for decentralized training of graph neural networks while preserving data privacy. However, existing FGL methods are predominantly designed for static graphs and rely on parameter averaging or distribution alignment, which implicitly assume that all features are equally transferable across clients, overlooking both the spatial and temporal heterogeneity and the presence of client-specific knowledge in real-world graphs. In this work, we identify that such assumptions create a vicious cycle of spurious representation entanglement, client-specific interference, and negative transfer, degrading generalization performance in Federated Learning over Dynamic Spatio-Temporal Graphs (FSTG). To address this issue, we propose a novel causality-inspired framework named SC-FSGL, which explicitly decouples transferable causal knowledge from client-specific noise through representation-level interventions. Specifically, we introduce a Conditional Separation Module that simulates soft interventions through client conditioned masks, enabling the disentanglement of invariant spatio-temporal causal factors from spurious signals and mitigating representation entanglement caused by client heterogeneity. In addition, we propose a Causal Codebook that clusters causal prototypes and aligns local representations via contrastive learning, promoting cross-client consistency and facilitating knowledge sharing across diverse spatio-temporal patterns. Experiments on five diverse heterogeneity Spatio-Temporal Graph (STG) datasets show that SC-FSGL outperforms state-of-the-art methods.



(a) Spatial heterogeneity (b) Temporal heterogeneity

(c) Spatial causal features (d) Temporal causal features

Figure 1: Spatial and temporal heterogeneity across clients and shared causal patterns. **(a)** differences in graph structures (nodes and edges). **(b)** variation in traffic trends at the same time. **(c)** similar road layouts imply shared spatial causal structures. **(d)** recurrent temporal patterns suggest shared temporal causality.

Introduction

Spatio-Temporal Graphs (STGs) model dynamic systems by jointly capturing spatial dependencies and temporal patterns. They are widely used in traffic forecasting, sensor networks, and mobility analytics (Liang et al. 2023; Yao et al. 2019; Lyu et al. 2025). In practice, STG data are often distributed across regions or institutions, where data privacy concerns restrict centralized collection (Huang et al. 2024; Wang et al. 2024; Meng et al. 2024; Qi et al. 2025a). This motivates the need for Federated Learning over Dynamic Spatio-Temporal Graphs (FSTGs), which enables decentralized training across clients without exposing raw data (Liu et al. 2025d; Wang et al. 2024; Huang et al. 2023; Qi et al. 2023). To ensure privacy preservation, Federated Graph Learning (FGL) offers a decentralized framework that trains Graph Neural Networks across clients while keeping local data private (Wan et al. 2025a,b; Liu et al. 2025e).

Existing FGL can be broadly categorized into two paradigms: Graph-FL, where each client holds multiple individual graphs, and Subgraph-FL, where clients own local subgraphs of a global graph (Li et al. 2025). Across both paradigms, various strategies have been proposed to address statistical heterogeneity, including structural alignment (Tan et al. 2023; Xie et al. 2021), pseudo graphs (Baek et al. 2023; Kim et al. 2025), or adaptive client modules (Li et al. 2024; Zhu et al. 2024). However, these techniques are primarily designed for static graphs and fall short when applied to dy-

*Corresponding author

namic environments. To support real-world systems where graphs evolve over time, recent works have extended FGL to FSTGs, incorporating temporal dynamics during federated training. For example, FUELS (Liu et al. 2025c) adopts contrastive learning with dual semantic alignment to address distribution shifts across FSTGs, while FedCroST (Zhang et al. 2024) introduces learnable spatio-temporal prompts to explicitly capture local STGs states. Despite such extensions, existing methods still rely on an assumption: *client knowledge is equally transferable and semantically aligned.*

However, such assumptions fail to hold in FSTGs, which are inherently characterized by complex and non-stationary heterogeneity across clients. As illustrated in Figures 1a–b, different clients possess varied spatial structures and temporal dynamics, violating the assumed cross-client homogeneous alignment in both feature and distributional space. These mismatches trigger a vicious cycle during federated optimization: (1) *Representation entanglement*, where generalizable patterns and specific variations are fused together in the learned features, making it hard to extract transferable knowledge; (2) *Specific interference*, where the global model overfits to non-transferable local signals, introducing noise into other clients' updates; and (3) *Negative transfer*, where improperly aggregated representations further degrade both global generalization and local adaptation. Such a self-reinforcing loop, where the inaccurate global model misguides local training and leads to even noisier representations and degraded aggregation quality, ultimately results in brittle personalization and poor performance. To address this fundamental challenge, we draw inspiration from recent advances in causal reasoning for graph learning (Lin et al. 2022; Lin, Lan, and Li 2021; Zhao and Zhang 2024), which emphasize disentangling stable causal mechanisms from spurious correlations. Building on these insights, we propose a novel categorization of causal knowledge in FSTGs: (1) *Shared causal knowledge*, which refers to invariant spatio-temporal patterns that are stable across clients, e.g., recurring traffic peaks or universal road layouts; (2) *Client-specific causal knowledge*, which captures factors unique to a particular region or system, e.g., city-specific road features (Figures 1c–d). Unfortunately, existing FGL methods ignore this distinction, treating all client knowledge as equally transferable, which leads to information leakage and false generalization. This raises a central question: *How can we extract shared causal knowledge while suppressing client-specific patterns in heterogeneous FSTGs, in order to enhance generalization and reduce negative transfer?*

Based on the above motivations, we propose **S**hared **C**ausality-inspired **F**ederated **S**patio-Temporal **G**raph **L**earning (**SC-FSGL**), a novel framework that disentangles transferable and client-specific knowledge in FSTG learning without explicitly constructing structural causal models. Specifically, a Conditional Separation Module with a learnable soft mask adaptively extracts invariant and client-specific causal features, mitigating entangled representations and spurious correlations due to heterogeneity. Invariant features are globally aggregated to enhance generalization, while client-specific knowledge remain local to prevent negative transfer. To further promote global consistency, we introduce a causal codebook that aligns shared spatio-temporal causal embeddings via a contrastive objective, reducing cross-client inconsistency and fostering a unified semantic space. Our key contributions are:

- We propose a causality-inspired representation learning framework for FSTGs, which adaptively separates shared and client-specific causal knowledge.
- We design a contrastive causal codebook to align spatial and temporal causal variables across clients, improving generalization and minimizing negative transfer under heterogeneous conditions.
- We conduct extensive experiments on five real-world spatio-temporal datasets under statistically heterogeneous settings. SC-FSGL consistently outperforms state-of-the-art baselines across multiple metrics.

Related Work

Causal Representation Learning in GNNs

Causal inference improves generalization by identifying invariant mechanisms across environments. In graph learning, methods like OrphicX (Lin et al. 2022) and Gem (Lin, Lan, and Li 2021) disentangle causal and spurious relations in static graphs to enhance robustness. Extensions to STGs, such as DyGNNExplainer (Zhao and Zhang 2024), decompose dependencies into static and dynamic components but assume centralized training, overlooking FSTGs heterogeneity. Moreover, existing approaches often ignore the distinction between shared and client-specific causal patterns, leading to causal interference when transferring non-generalizable knowledge. To address this, we separate spatio-temporal causal features via conditional masking and contrastive alignment, supporting robust and personalized FSTG learning.

Spatio-temporal Graph Learning

STG learning has been widely applied to traffic and mobility forecasting (Zhang et al. 2025a,b). Early models like DCRNN (Li et al. 2017) and STGCN (Yu, Yin, and Zhu 2017) integrate GCNs with recurrent or convolutional units to model dynamic dependencies. Later works such as GMAN (Zheng et al. 2020), MegaCRN (Jiang et al. 2023), STEAM (Gao et al. 2025) and TWIST (Wang et al. 2025c) employ attention or meta-graph learning to enhance performance. However, most assume centralized or homogeneous data, lacking robustness to distribution shifts. Unlike them, we introduce a causal representation framework for heterogeneous FSTGs.

Federated Graph Learning on STGs

FGL is challenged by heterogeneous graph data across clients. Prior works address this via parameter reweighting (Jiang et al. 2023; Huang et al. 2024) or subgraph generation (Zhang et al. 2021; Baek et al. 2023), but often overlook structural invariants. FedStar (Tan et al. 2023) extracts such invariants in static graphs, yet fails to handle dynamic STG settings. Recent approaches (Yuan et al. 2022; Zhang et al. 2024; Liu et al. 2025c) model local

spatio-temporal distributions via global networks, prompts or contrastive learning. However, they focus on distributional alignment without explicitly modeling transferable causal patterns. Our method instead disentangles shared and client-specific causal representations, enabling more robust generalization under spatio-temporal heterogeneity.

Problem Description

Problem setting & notations

At time step t, the input STG $\mathcal{G}_t^k = \{X_t^k, A_t^k\}$ of client k ($k \in K$) comprises a historical node feature matrix $\mathbf{X}_t^k \in \mathbb{R}^{|V^k| \times d}$ and an adjacency matrix $\mathbf{A}_t^k \in \mathbb{R}^{|V^k| \times |V^k|}$, where K represents the total number of clients, V^k represents the sets of nodes feature and d represents the dimension of node features. The prediction results at time t, denoted by $\mathcal{Y}_t^k = \{\hat{X}_t^k, A_t^k\}$, includes predicted future node features $\hat{\mathbf{X}}_t^k \in \mathbb{R}^{|V^k| \times d}$. At this point, our local task is to train a model f_{θ^k} with model parameters θ^k. This model aims to establish causal relationships between variables based on the historical γ steps to predict the future β steps of spatio-temporal states:

$$\{X_{t-\gamma}^k, X_{t-\gamma+1}^k, ..., X_t^k\} \overset{f_{\theta^k}}{\underset{\theta^k}{\rightarrow}} \{\hat{X}_{t+1}^k, \hat{X}_{t+2}^k, ..., \hat{X}_{t+\beta}^k\}, \quad (1)$$

Causal View of FSTGs

In FSTGs, each client k ($k \in K$) is treated as an environment with distinct data-generating processes (See Figure 2), and K is the set of participating clients. We decompose the spatial and temporal variables $\mathcal{S}_t^k, \mathcal{T}_t^k$ into shared and client-specific components:

$$\mathcal{S}_t^k = \mathcal{S}_{t,c}^k \cup \mathcal{S}_{t,o}^k, \quad \mathcal{T}_t^k = \mathcal{T}_{t,c}^k \cup \mathcal{T}_{t,o}^k. \quad (2)$$

To analyze the effect of shared causal variables, we adopt the Structural Causal Model (SCM) framework (Pearl 2009). In this framework, the **do**-operator $\mathbf{do}(\cdot)$ denotes an intervention that sets a variable to a fixed value and removes all incoming causal influences. Using this formalism, the interventional distribution $P(\mathcal{Y}_t^k \mid \mathbf{do}(\mathcal{T}_{t,c}^k))$ quantifies the causal effect of the shared temporal variables on the outcome. Under the assumption that $\mathcal{T}_{t,o}^k$ and $\mathcal{S}_t^k$ are conditionally independent given $\mathcal{G}_t^k$, we approximate the interventional distribution via observational distributions as:

$$\begin{aligned} P(\mathcal{Y}_t^k|\mathbf{do}(\mathcal{T}_{t,c}^k)) &= \sum P(\mathcal{Y}_t^k|\mathbf{do}(\mathcal{T}_t^k))P(\mathcal{T}_{t,c}^k) \\ &= \sum P(\mathcal{T}_{t,o}^k) \sum P(\mathcal{T}_t^k|\mathcal{G}_t^k)P(\mathcal{S}_t^k), \end{aligned} \quad (3)$$

where $\mathcal{G}_t^k$ denotes the observed graph. Similarly, the interventional distribution with respect to the spatial is given by $P(\mathcal{Y}_t^k|\mathbf{do}(\mathcal{S}_{t,c}^k)) = \sum P(\mathcal{T}_t^k) \sum P(\mathcal{Y}_t^k|\mathcal{G}_t^k)P(\mathcal{T}_t^k)$. See Appendix for details.

Since direct interventions are impractical in the federated scenarios, we introduce a learnable soft mask $M_t^k \in [0,1]^d$ to approximate intervention effects in latent space:

$$\Phi(X_t^k) = (1 + \mathrm{LN}(M_t^k)) \odot X_t^k, \quad (4)$$

where $\mathrm{LN}(\cdot)$ denotes layer normalization and $\odot$ is the Hadamard product. This soft mask simulates intervention by attenuating the influence of client-specific variables, thereby approximating $\mathbf{do}(X_t^k = x_t^k)$ in representation space.

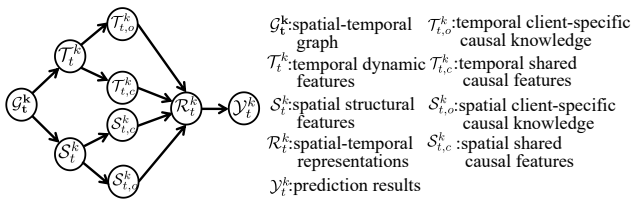


Figure 2: A conceptual illustration inspired by the SCM on client k, showing the relationships among observed variables and latent components.

To promote the invariance of retained features across heterogeneous clients, we adopt the principle of *Invariant Risk Minimization (IRM)* (Arjovsky et al. 2020; Chang et al. 2020), which encourages a predictor ω to remain optimal across all environments. Formally, this principle seeks a feature representation Φ such that a single predictor achieves optimality under all environment-specific risks:

$$\min_{\Phi} \sum_{k \in K} \mathcal{R}_k(\omega, \Phi) \quad \text{s.t.} \quad \omega \in \arg\min_{\bar{\omega}} \mathcal{R}_k(\bar{\omega}, \Phi), \quad (5)$$

where $R_k(\omega, \Phi) = \mathbb{E}_{(x,y)\sim\mathcal{D}_k}[\ell(\omega(\Phi(x)), y)]$ denotes the prediction risk of client k, with ω being a linear classifier and Φ the shared encoder. This constraint enforces that Φ should extract invariant features that enable ω to generalize across all client distributions.

Methodology

In this section, we introduces the SC-FSGL in the Figure 3, a prediction model for FSTG that incorporates shared causal relationships to enhance accuracy and mitigate the impact of client-specific causal knowledge on the global model. Next, we will delineate the various modules of the SC-FSGL.

Spatio-Temporal Embedding (STE)

The spatial network structure and historical observations play a crucial role in FSTG prediction. To encode spatial information, we employ node2vec (Grover and Leskovec 2016) as the **Structure Extractor**, generating spatial embeddings (SE) that preserve the graph topology. For temporal representation, we encode historical timestamps into $\mathcal{T}_{his}^k \in \mathbb{R}^{\gamma \times |V^k| \times D}$ using a week-day-hour format, and similarly encode future time steps into $\mathcal{T}_{pred}^k \in \mathbb{R}^{\beta \times |V^k| \times D}$, where $|V^k|$ is the number of nodes and D the embedding dimension. We then concatenate temporal and spatial embeddings to obtain time-varying vertex representations following GMAN (Zheng et al. 2020): $STE_{his}^k = \mathbf{concat}[\mathcal{T}_{his}^k, SE^k]$ and $STE_{pred}^k = \mathbf{concat}[\mathcal{T}_{pred}^k, SE^k]$, which are input into the Temporal Feature Extractor Φ_t^k. Meanwhile, the $\mathcal{G}_t^k$ is processed by the Spatial Feature Extractor Φ_s^k to generate the spatial representation $\mathcal{S}_t^k$.

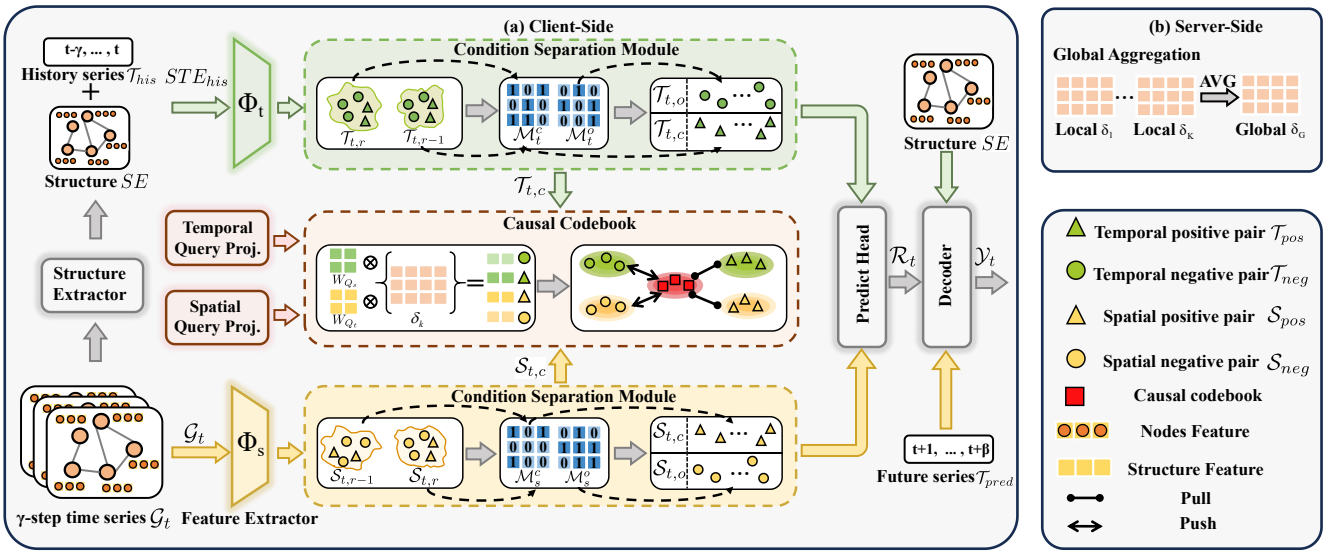


Figure 3: The figure illustrates the overall architecture of SC-FSGL, including the client-side feature extraction, causal codebook construction, and global aggregation process. The model extracts spatio-temporal features, separates shared and client-specific causal variables, and leverages the causal codebook to enhance the prediction accuracy through soft intervention and contrastive representation alignment.

Conditional Separation Module

To effectively disentangle transferable causal patterns from client-specific spurious factors, we propose a Conditional Separation Module, based on the assumption that shared causal factors remain stable across communication rounds, while client-specific patterns tend to fluctuate due to local heterogeneity. However, in practice, these stable causal patterns are often entangled with noisy client-specific signals due to variations in data distributions and latent confounders. To address this, we introduce a *soft intervention mechanism* that simulates causal interventions in the latent space, selectively attenuating spurious influences and facilitating the extraction of invariant causal representations.

Specifically, the Conditional Separation Module comprises two symmetric submodules, applied respectively to temporal and spatial representations. Given an input feature $\mathcal{F}_t^k \in \{\mathcal{T}_t^k, \mathcal{S}_t^k\}$, we denote $\mathcal{F}_{t,r-1}^k$ as the representation obtained at the previous round $r-1$, serving as a reference to extract causal patterns that are invariant across rounds, where r represents the current communication round. Then, we learn a soft mask $\mathcal{M}_t^{k,c} = \sigma(\mathcal{F}_{t,r-1}\omega_{\mathcal{F}} + b_{\mathcal{F}})$ that highlights shared causal dimensions, and a complementary mask $\mathcal{M}_t^{k,o} = 1 - \mathcal{M}_t^{k,c}$ to capture client-specific components. The $\omega_{\mathcal{F}}$ denotes the weight matrix and $b_{\mathcal{F}}$ the bias term of a MLP, $\sigma(\cdot)$ is a nonlinear activation function. These masks are applied to the input representations to disentangle causal and spurious components via:

$$\begin{cases} \mathcal{F}_{t,c}^k = \sigma\left[(1 + \mathrm{LN}(\mathcal{M}_t^{k,c})) \odot \mathcal{F}_{t,r}^k\right] \\ \mathcal{F}_{t,o}^k = \sigma\left[(1 + \mathrm{LN}(\mathcal{M}_t^{k,o})) \odot \mathcal{F}_{t,r}^k\right], \end{cases} \tag{6}$$

where $\mathrm{LN}(\cdot)$ is layer normalization. This decomposition enables the model to approximate causal interventions by suppressing client-specific noise, thus promoting the extraction of stable, transferable features across clients.

To guide the mask learning, we jointly optimize two objectives: (1) a *contrastive alignment loss* that encourages the shared representations $\mathcal{F}_{t,c}^k$ to be semantically consistent across clients, using the global causal codebook δ as a reference (detailed in the next section); (2) an *IRM-inspired regularization* that promotes representation invariance across heterogeneous client environments by penalizing the gradient of the prediction loss with respect to the predictor ω. Following (Arjovsky et al. 2020), we relax the original IRM constraint into a differentiable surrogate objective using a gradient penalty, which encourages the predictor to remain optimal across different environments. Formally, the IRM loss is defined as:

$$\mathcal{L}_{\text{irm}} = \sum_{\Phi_\delta \in \{\Phi_s, \Phi_t\}} \left\| \nabla_\omega \mathcal{L}_{\text{pred}}(\omega \circ \Phi_\delta) \right\|_2^2, \tag{7}$$

where $\mathcal{L}_{\text{pred}}$ is the prediction loss, measured by MAE between predicted values and ground truth. Φ_δ is the encoder aligned with the codebook δ, and ω is the prediction layer. By minimizing this gradient norm, the model enforces stability in predictive performance across clients, thereby reinforcing the extraction of invariant causal features.

Causal Codebook

Existing methods often increase the amount of information fed into the model, leading the model to overly focus on client-specific knowledge that is not beneficial to the current client during the aggregation phase (Qi et al. 2025b; Wang et al. 2025a; Liu et al. 2024). To reduce the interference of client-specific causal parts and enhance the contribution

of shared causal knowledge, we propose a causal codebook $\delta \in \mathbb{R}^{\phi \times d}$, where ϕ is the number of causal prototypes and d is the feature dimension of items. This is a shared learnable prototype matrix that acts as a *semantic anchor*, i.e., a set of latent reference vectors that define a consistent representation space to align spatial and temporal causal features across clients. To align causal variables with δ, we compute similarity scores between projected features and codebook:

$$\begin{cases} \alpha_j^s = \frac{exp((W_{Q_s} * \mathcal{S}_{t,c}^k + b) * \delta_j^\top)}{\sum_{j=1}^{\phi} exp(((W_{Q_s} * \mathcal{S}_{t,c}^k + b) * \delta_j^\top)} \\ \alpha_j^t = \frac{exp((W_{Q_t} * \mathcal{T}_{t,c}^k + b) * \delta_j^\top)}{\sum_{j=1}^{\phi} exp(((W_{Q_t} * \mathcal{T}_{t,c}^k + b) * \delta_j^\top)}, \end{cases} \tag{8}$$

where j denote the row indices of δ, W_Q represents the learnable parameters of the local Query model, which performs a projection transformation of $\mathcal{T}_{t,c}^k$ and $\mathcal{S}_{t,c}^k$. Considering the causal relationship between the spatial causal variables and the temporal causal variables and their impact on the final result $\mathcal{Y}_t^k$, we further strengthen the shared spatial and temporal causal relationships in local model. The top-1 (max) and second-best ($\hat{max}$) matching entries in the codebook are used as positive and negative pairs:

$$\begin{cases} \mathcal{S}_{pos}^k = max_j(\alpha_j^s * \delta_j) \\ \mathcal{S}_{neg}^k = \hat{max}_j(\alpha_j^s * \delta_j) \end{cases} \begin{cases} \mathcal{T}_{pos}^k = max_j(\alpha_j^t * \delta_j) \\ \mathcal{T}_{neg}^k = \hat{max}_j(\alpha_j^t * \delta_j). \end{cases} \tag{9}$$

Based on the selected pairs, we define a contrastive loss to align $\mathcal{S}_{t,c}^k$ and $\mathcal{T}_{t,c}^k$ with their positive codebook entries while separating them from negatives:

$$\begin{aligned} \mathcal{L}_{\text{com}} = &\log \frac{\exp(\text{sim}(\delta, \mathcal{S}_{\text{pos}}^k)/\tau)}{\exp(\text{sim}(\delta, \mathcal{S}_{\text{pos}}^k)/\tau) + \sum \exp(\text{sim}(\delta, \mathcal{S}_{\text{neg}}^k)/\tau)} \\ &+ \log \frac{\exp(\text{sim}(\delta, \mathcal{T}_{\text{pos}}^k)/\tau)}{\exp(\text{sim}(\delta, \mathcal{T}_{\text{pos}}^k)/\tau) + \sum \exp(\text{sim}(\delta, \mathcal{T}_{\text{neg}}^k)/\tau)}, \end{aligned} \tag{10}$$

where $\text{sim}(\cdot)$ denotes cosine similarity and τ is the temperature parameter. For each client, the top-1 matched prototype serves as the positive anchor, while the second-best is used as a hard negative, following hard negative mining to enhance discriminability. This encourages shared causal features to align with semantically consistent prototypes while being separated from close but suboptimal ones.

Local Training and Inference

During the training phase, the client's encoder represents $\mathcal{R}_t^k$ as input to the decoder for predicting future value $\mathcal{Y}_t^k$. To minimize the discrepancy between the predicted output and the actual ground truth, we define a predictor loss function as $L_{mse} = ||\mathcal{R}_t^k - \mathcal{Y}_t^k||^2$. Finally, we formulate the complete local loss function, which integrates additional components for comprehensive error minimization:

$$\mathcal{L}_{\text{local}} = \mathcal{L}_{\text{mse}} + \alpha \cdot \mathcal{L}_{\text{com}} + \beta \mathcal{L}_{irm}, \tag{11}$$

where $\mathcal{L}_{\text{com}}$ aligns shared causal representations via contrastive learning, and the IRM penalty enforces predictor invariance across clients. The α and β are hyperparameters that balance the influence of the contrastive loss and the IRM regularization term, respectively. During local training, clients extract spatio-temporal features, apply conditional separation to obtain disentangled representations, and perform prediction. The codebook δ is updated via federated averaging and used to align causal features during local training.

Convergence Analysis

In this section, we analyze and prove the convergence of the model. We have the following assumptions:

Assumption 1 *The objective function L_k is convex,*

$$L_k(a) \geq L_k(b) + \langle \nabla L_k(b), a - b \rangle. \tag{12}$$

Assumption 2 *For L_k, all gradients of the model parameter θ associated with it are constrained by a constant M,*

$$\mathbb{E}(||\nabla L_k(\theta)||^2) \leq M^2. \tag{13}$$

Suppose there are K participating clients, $\mathcal{G}_{k,t}$ representing the STG input of client k($k \in K$) at the edge of time t, and the local model parameter is defined as $\theta_r^k = \{\theta_{r,a}^k, \theta_{r,b}\}$, where $r \in R$ represents the aggregation Round. $\theta_{r,a}$ represents a local model parameter that does not participate in aggregation, $\theta_{r,b}$ is a model parameter of causal codebook, and $\theta_{r,b}$ participates in global aggregation. If L_k is a convex function, then θ is a convex set, and we assume a bound between the model parameters and the optimal model parameters, where $||\theta_r^k - \theta^{k,*}|| \leq I$. Partial loss function is $L(\theta_r^k) = L(\theta_{r,a}^k) + L(\theta_{r,b}^k)$.

Theorem 1 *causal codebook converges to the following bound when the learning rate is η,*

$$\frac{1}{R}\sum_{r=1}^{R}[L_k(\theta_{r,b}^k) - L_k(\theta_b^{k,*})] \leq I^2 \frac{1}{2R\eta} + \frac{M^2}{2}\eta. \tag{14}$$

Detailed proofs are provided in the Appendix.

Experiments

Experimental settings

Datasets. We introduce the datasets used in our experimentation, which comprise real traffic data from five distinct cities: METRLA (Li et al. 2017), PEMSD4 (Guo et al. 2019), PEMSD7(M) (Yu, Yin, and Zhu 2017), PEMSD8 (Guo et al. 2019), and PEMSBAY (Li et al. 2017). Each dataset corresponds to one client to preserve strong spatio-temporal heterogeneity. For details on dataset statistics, preprocessing, and partitioning protocols, please refer to the Appendix.

Baseline. We compared our approach with state-of-the-art methods, including the baseline FedAvg (McMahan et al. 2017), methods addressing federated data heterogeneity such as FedProx (Li et al. 2020), FedRep (Collins et al. 2021), Moon (Li, He, and Song 2021), and FedStar (Tan et al. 2023), as well as models for spatio-temporal prediction tasks like GMAN (Zheng et al. 2020), MegaCRN (Jiang et al. 2023) and FUELS (Liu et al. 2025c).

Metrics. We evaluate SC-FSGL using three standard metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE).

Client	METRLA		PEMSD4		PEMSD7(M)		PEMSD8		PEMSBAY		Avg.	
Metric	**MAE**	**MAPE**	**MAE**	**MAPE**	**MAE**	**MAPE**	**MAE**	**MAPE**	**MAE**	**MAPE**	**MAE**	**MAPE**
FedAvg	12.21	23.58%	3.41	7.13%	5.02	14.25%	3.72	7.22%	3.44	7.69%	5.56	11.97%
FedProx	7.80	19.37%	3.20	6.83%	7.15	16.22%	3.09	6.70%	4.35	10.15%	5.12	11.85%
Moon	7.54	18.00%	2.65	5.57%	6.21	15.66%	3.11	5.84%	4.35	8.11%	4.77	10.64%
FedRep	6.78	17.71%	8.88	14.95%	4.37	12.15%	2.46	5.79%	7.89	13.67%	6.07	12.85%
FedStar	28.71	56.69%	4.35	8.17%	16.41	34.24%	3.25	7.16%	33.71	53.31%	17.29	31.91%
GMAN	9.51	22.26%	3.62	6.60%	3.64	10.53%	3.12	5.84%	3.29	6.45%	4.64	10.34%
MegaCRN	10.39	35.54%	4.39	10.75%	8.84	29.52%	3.78	8.54%	5.10	14.05%	6.50	19.68%
FUELS	7.30	18.84%	4.87	8.44%	4.70	12.12%	3.36	6.56%	2.34	4.92%	4.51	10.18%
Local	7.08	**17.46%**	4.17	9.75%	7.11	14.81%	2.54	6.06%	**2.39**	**6.13%**	4.66	10.84%
SC-FSGL	**6.38**	17.57%	**2.46**	**4.65%**	**3.50**	**9.94%**	**2.29**	**5.72%**	6.44	12.53%	**4.21**	**10.08%**

Table 1: Performance comparison of different approaches for FSTG on various datasets (Clients).

Time	60 min		30 min		15 min	
Metric	**MAE**	**MAPE**	**MAE**	**MAPE**	**MAE**	**MAPE**
FedAvg	5.56	11.97%	4.80	10.99%	5.32	10.83%
FedProx	5.12	11.85%	4.68	9.85%	4.15	9.04%
Moon	4.77	10.64%	5.12	10.95%	5.31	13.30%
FedRep	6.07	12.85%	3.93	8.48%	4.55	10.83%
FedStar	17.29	31.91%	23.32	42.59%	19.93	35.00%
GMAN	4.64	10.34%	3.92	8.80%	3.64	8.34%
MegaCRN	6.50	19.68%	6.38	19.55%	6.33	19.46%
FUELS	4.51	10.18%	4.24	10.31%	3.87	8.80%
Local	4.66	10.84%	5.00	10.92%	4.64	13.09%
SC-FSGL	**4.21**	**10.08%**	**3.67**	**8.19%**	**3.29**	**7.76%**

Table 2: Performance across prediction horizons.

Implementation Details. For all federated learning methods, communication rounds were set to early stop, and local training consisted of one epoch. Our implementation, developed using PyTorch, was executed on two NVIDIA 3090 GPUs for all experiments. Each dataset has unique traffic network structures and timestamps.

The Results of SC-FSGL

Performance Across Prediction Times. To comprehensively evaluate SC-FSGL, we compare it with eight representative baselines across three prediction horizons 60, 30, and 15 minutes, on MAE and MAPE metrics (see Table 2). SC-FSGL consistently achieves the best performance across all metrics and time spans. For 60 minute forecasts, it has the lowest MAE (4.21), outperforming FUELS (4.51) and GMAN (4.64), as well as the MAPE (10.08%). Similar trends are observed at 30 and 15 minutes, where SC-FSGL achieves MAE of 3.67 and 3.29, respectively, surpassing both spatio-temporal (e.g., GMAN) and federated (e.g., FedProx, Moon) baselines. Notably, methods like FedRep and Moon, though partially addressing heterogeneity, suffer from unstable performance due to lacking causal disentanglement. Models such as FedStar and MegaCRN underperform in all settings, indicating poor robustness under spatiotemporal heterogeneity. In contrast, SC-FSGL's consistent superiority confirms the effectiveness of its causal disentanglement and alignment mechanisms in heterogeneous FSTG learning.

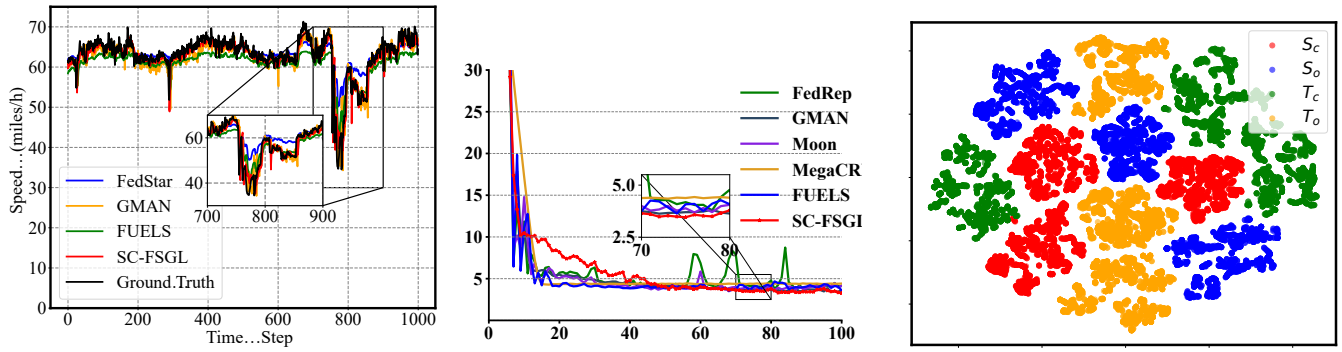
The prediction performance on different clients. To evaluate SC-FSGL's robustness under spatio-temporal heterogeneity, we compare its predictive accuracy across five real-world traffic datasets, each representing a distinct client. As shown in Table 1, SC-FSGL achieves the best or second-best results. For instance, on PEMSD4 and PEMSD8, SC-FSGL achieves the lowest MAEs of 2.46 and 2.29, outperforming both federated (e.g., FedAvg, FedProx) and spatio-temporal models (e.g., GMAN, FUELS). On PEMSD7(M), it achieves a MAE of 3.50, significantly better than Moon (6.21) and MegaCRN (8.84), demonstrating resilience to temporal drift. SC-FSGL effectively extracts shared causal features, mitigating negative transfer and enhancing generalization. When averaged across all clients, it attains the best overall score.

Ablation Study

In this section, we conducted ablation experiments on SC-FSGL. Using MAE as the evaluation metric, we performed ablation experiments on five datasets. We then reported the average results. In the Figure 5, "w/o Book" indicates the removal of the causal codebook, "w/o CS" indicates the removal of the Conditional Separation Module, and "w/o Book and CS" indicates the simultaneous removal of both the causal codebook and Conditional Separation Modules. We observed that the SC-FSGL with the complete modules showed the lowest MAE in all variants, and this demonstrated the effectiveness of our approach. In Figure 5(f), the SC-FSGL showed the lowest MAE. In contrast, the "w/o Book" showed a significant increase in MAE, indicating the importance of the causal codebook for predictive performance. The "w/o CS" variant mainly investigated whether to conduct conditional separation or not, and the results indicated that the CS contributes to the overall predictive performance of the model.

Predictive Performance Analysis

Figure 4(a) shows the visual comparison of the predicted results with the ground truth on the PEMS04 dataset. Compared to other baseline methods like FedStar, FUELS, and GMAN, SC-FSGL fits the ground truth much more closely.



(a) (b) (c)

Figure 4: Performance Analysis of SC-FSGL on the PEMSD4 Dataset: (a) Prediction Results, (b) MAE Curves, and (c) t-SNE Visualization of Disentangled Causal Representations.

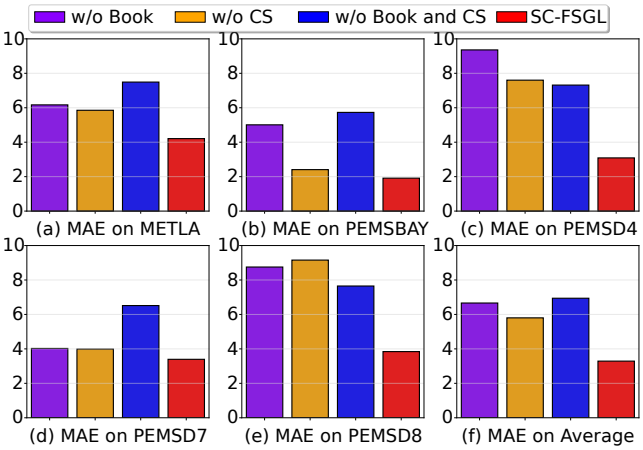


Figure 5: Ablation Study of SC-FSGL.

The predicted result by SC-FSGL not only aligns well with the magnitude and trend of the actual data but also captures the temporal fluctuations and peak variations more accurately. In contrast, competing methods tend to either smooth out the signal, missing key variations, or exhibit phase shifts in temporal alignment. This demonstrates SC-FSGL's ability to model fine-grained temporal dynamics in a federated setting. The predictive accuracy is due to the disentanglement of shared representations and client-specific representations, which can better generalize and preserve client-specific knowledge.

MAE Trends over Communication Rounds

Figure 4(b) depicts the trend of MAE over the training rounds on the PEMSD4 dataset. Compared to the representative baseline methods like FedRep, Moon, GMAN, FUELS, and MegaCRN, SC-FSGL shows lower MAE throughout the training process. While some baselines suffer from a drop of MAE followed by plateaus or fluctuation, SC-FSGL shows a smooth and steady decrease, indicating the stable learning dynamics and enhanced robustness under the heterogeneous conditions. This performance can be attributed to SC-FSGL's causal disentanglement mechanism, which can separate transferable knowledge and suppress client-specific knowledge. The observed trend show cases SC-FSGL's capability to utilize global causal structures for enhanced federated prediction.

Causal Feature Visualization via t-SNE

To evaluate the effectiveness of SC-FSGL in disentangling causal representations, we visualize the learned features using t-SNE, as shown in Figure 4(c). The embeddings of $\mathcal{S}_c$, $\mathcal{S}_o$, $\mathcal{T}_c$, and $\mathcal{T}_o$ are clearly separable by semantic type and show multiple compact clusters inside each category. This indicates the heterogeneity and non-stationary characteristics of the data. The shared representations ($\mathcal{S}_c$, $\mathcal{T}_c$) form tight sub-clusters, which means that the model can capture diverse and transferable patterns across clients, rather than following a single unified mode. This is further strengthened by the prototype-based alignment mechanism of the causal codebook, which promotes the representations to gather around semantic anchors. In contrast, the private representations ($\mathcal{S}_o$, $\mathcal{T}_o$) seem more scattered, allowing local heterogeneity to be more evident. The SC-FSGL can effectively discriminate transferable knowledge from client-specific factors while maintaining structural diversity, which is conducive to robust learning in heterogeneous FSTGs.

Conclusion

In this study, we propose SC-FSGL, a methodology for predicting heterogeneous FSTG data. The SC-FSGL methodology is designed to enhance the influence of shared causal relationships on model predictions, reduce the effects of client-specific causal knowledge on the global model, and decrease the granularity distance between temporal and spatial causal variables. We conduct experiments using five real-world datasets to evaluate the performance of our approach. Results demonstrate superior performance compared to prior methods on multiple baselines.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under grants 62376103, 62302184, 62436003 and 62206102; Major Science and Technology Project of Hubei Province under grant 2024BAA008.

References

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2020. Invariant Risk Minimization. arXiv:1907.02893.

Baek, J.; Jeong, W.; Jin, J.; Yoon, J.; and Hwang, S. J. 2023. Personalized subgraph federated learning. In *International conference on machine learning*, 1396–1415. PMLR.

Chang, S.; Zhang, Y.; Yu, M.; and Jaakkola, T. 2020. Invariant Rationalization. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 1448–1458. PMLR.

Collins, L.; Hassani, H.; Mokhtari, A.; and Shakkottai, S. 2021. Exploiting shared representations for personalized federated learning. In *International conference on machine learning*, 2089–2099. PMLR.

Gao, M.; Xu, K.; Gao, X.; Cai, T.; and Ge, H. 2025. Spatial-Temporal Heterogenous Graph Contrastive Learning for Microservice Workload Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(11): 11681–11689.

Grover, A.; and Leskovec, J. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 855–864.

Guo, S.; Lin, Y.; Feng, N.; Song, C.; and Wan, H. 2019. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 922–929.

Huang, W.; Ye, M.; Shi, Z.; and Du, B. 2023. Generalizable Heterogeneous Federated Cross-Correlation and Instance Similarity Learning. *TPAMI*.

Huang, W.; Ye, M.; Shi, Z.; Wan, G.; Li, H.; Du, B.; and Yang, Q. 2024. A Federated Learning for Generalization, Robustness, Fairness: A Survey and Benchmark. *TPAMI*.

Jiang, R.; Wang, Z.; Yong, J.; Jeph, P.; Chen, Q.; Kobayashi, Y.; Song, X.; Fukushima, S.; and Suzumura, T. 2023. Spatio-temporal meta-graph learning for traffic forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 8078–8086.

Kim, S.; Lee, Y.; Oh, Y.; Lee, N.; Yun, S.; Lee, J.; Kim, S.; Yang, C.; and Park, C. 2025. Subgraph Federated Learning for Local Generalization. arXiv:2503.03995.

Li, Q.; He, B.; and Song, D. 2021. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10713–10722.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2: 429–450.

Li, X.; Wu, Z.; Zhang, W.; Sun, H.; Li, R.-H.; and Wang, G. 2024. AdaFGL: A New Paradigm for Federated Node Classification with Topology Heterogeneity. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, 2517–2530.

Li, X.; Zhu, Y.; Pang, B.; Yan, G.; Yan, Y.; Li, Z.; Wu, Z.; Zhang, W.; Li, R.-H.; and Wang, G. 2025. OpenFGL: A Comprehensive Benchmark for Federated Graph Learning. *Proceedings of the VLDB Endowment*, 18(5): 1305–1320. Publisher Copyright: © 2025, VLDB Endowment. All rights reserved.; 51st International Conference on Very Large Data Bases, VLDB 2025 ; Conference date: 01-09-2025 Through 05-09-2025.

Li, Y.; Yu, R.; Shahabi, C.; and Liu, Y. 2017. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*.

Liang, Y.; Xia, Y.; Ke, S.; Wang, Y.; Wen, Q.; Zhang, J.; Zheng, Y.; and Zimmermann, R. 2023. Airformer: Predicting nationwide air quality in china with transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 14329–14337.

Lin, W.; Lan, H.; and Li, B. 2021. Generative Causal Explanations for Graph Neural Networks. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 6666–6679. PMLR.

Lin, W.; Lan, H.; Wang, H.; and Li, B. 2022. OrphicX: A Causality-Inspired Latent Variable Model for Interpreting Graph Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13729–13738.

Liu, J.; Cheng, J.; Han, R.; Tu, W.; Wang, J.; and Peng, X. 2025a. Federated Graph-Level Clustering Network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 18870–18878.

Liu, J.; Han, R.; Tu, W.; Wang, H.; Wu, J.; and Cheng, J. 2025b. Federated Node-Level Clustering Network with Cross-Subgraph Link Mending. In *Proceedings of the International Conference on Machine Learning*, 38540–38556.

Liu, Q.; Sun, S.; Liang, Y.; Liu, M.; and Xue, J. 2025c. Personalized Federated Learning for Spatio-Temporal Forecasting: A Dual Semantic Alignment-Based Contrastive Approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(11): 12192–12200.

Liu, Y.; He, Z.; Wang, S.; Wang, Y.; Wang, P.; Huang, Z.; and Sun, Q. 2025d. Federated Subgraph Learning via Global-Knowledge-Guided Node Generation. *Sensors*, 25(7).

Liu, Y.; Wang, H.; Wang, S.; He, Z.; Xu, W.; Zhu, J.; and Yang, F. 2024. Disentangle Estimation of Causal Effects from Cross-Silo Data. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6290–6294. IEEE.

Liu, Y.; Xu, C.; Zhao, Z.; Wang, Y.; and He, Z. 2025e. Federated Graph Learning under Expressiveness Heterogeneity with Personalized Subgraph Selection. *Knowledge-Based Systems*, 115106.

Lyu, W.; Zhong, S.; Yang, G.; Wang, H.; Ding, Y.; Wang, S.; Liu, Y.; He, T.; and Zhang, D. 2025. InCo: Exploring Inter-Trip Cooperation for Efficient Last-mile Delivery. In *Proceedings of the ACM on Web Conference 2025*, 5183–5191.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282. PMLR.

Meng, L.; Qi, Z.; Wu, L.; Du, X.; Li, Z.; Cui, L.; and Meng, X. 2024. Improving global generalization and local personalization for federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36.

Pearl, J. 2009. *Causality*. Cambridge university press.

Qi, Z.; Meng, L.; Chen, Z.; Hu, H.; Lin, H.; and Meng, X. 2023. Cross-silo prototypical calibration for federated learning with non-iid data. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3099–3107.

Qi, Z.; Meng, L.; Li, Z.; Hu, H.; and Meng, X. 2025a. Cross-Silo Feature Space Alignment for Federated Learning on Clients with Imbalanced Data. In *The 39th Annual AAAI Conference on Artificial Intelligence (AAAI-25)*, 19986–19994.

Qi, Z.; Zhou, S.; Meng, L.; Hu, H.; Yu, H.; and Meng, X. 2025b. Federated Deconfounding and Debiasing Learning for Out-of-Distribution Generalization. In *The 34th International Joint Conference on Artificial Intelligence*, 6084–6092.

Tan, Y.; Liu, Y.; Long, G.; Jiang, J.; Lu, Q.; and Zhang, C. 2023. Federated learning on non-iid graphs via structural knowledge sharing. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 9953–9961.

Wan, G.; Cheng, X.; Liu, R.; Huang, W.; Shi, Z.; Jin, P.; Zhang, G.; Du, B.; and Ye, M. 2025a. Multi-order Orchestrated Curriculum Distillation for Model-Heterogeneous Federated Graph Learning. In *NeurIPS*.

Wan, G.; Qian, J.; Huang, W.; Xu, Q.; Guo, X.; Li, B.; Zhang, G.; Du, B.; and Ye, M. 2025b. OASIS: One-Shot Federated Graph Learning via Wasserstein Assisted Knowledge Integration. In *NeurIPS*.

Wang, H.; Wang, S.; Li, J.; Ren, H.; Han, X.; Xu, W.; Guo, S.; Zhang, T.; and Li, R. 2025a. BSemiFL: Semi-supervised Federated Learning via a Bayesian Approach. In *Forty-second International Conference on Machine Learning*.

Wang, H.; Xu, H.; Li, Y.; Xu, Y.; Li, R.; and Zhang, T. 2024. FedCDA: Federated Learning with Cross-rounds Divergence-aware Aggregation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Wang, L.; Tu, W.; Cheng, J.; Wang, J.; Tang, X.; and Wang, C. 2025b. Discovering Maximum Frequency Consensus: Lightweight Federated Learning for Medical Image Segmentation. In *Proceedings of the ACM International Conference on Multimedia*, 1900–1909.

Wang, P.; Feng, L.; Zhang, W.; and Hui, K. 2025c. TWIST: An Efficient Spatial-Temporal Transformer With Temporal Window and Sparse Attention for Traffic Forecasting. *IEEE Internet of Things Journal*.

Xie, H.; Ma, J.; Xiong, L.; and Yang, C. 2021. Federated Graph Classification over Non-IID Graphs. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 18839–18852. Curran Associates, Inc.

Yao, H.; Tang, X.; Wei, H.; Zheng, G.; and Li, Z. 2019. Revisiting spatial-temporal similarity: A deep learning framework for traffic prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 5668–5675.

Yu, B.; Yin, H.; and Zhu, Z. 2017. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*.

Yuan, X.; Chen, J.; Yang, J.; Zhang, N.; Yang, T.; Han, T.; and Taherkordi, A. 2022. Fedstn: Graph representation driven federated learning for edge computing enabled urban traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*, 24(8): 8738–8748.

Zhang, K.; Yang, C.; Li, X.; Sun, L.; and Yiu, S. M. 2021. Subgraph federated learning with missing neighbor generation. *Advances in Neural Information Processing Systems*, 34: 6671–6682.

Zhang, Y.; Wang, X.; Wang, P.; Wang, B.; Zhou, Z.; and Wang, Y. 2024. Modeling Spatio-Temporal Mobility Across Data Silos via Personalized Federated Learning. *IEEE Transactions on Mobile Computing*, 23(12): 15289–15306.

Zhang, Y.; Wang, X.; Yu, X.; Sun, Z.; Wang, K.; and Wang, Y. 2025a. Drawing informative gradients from sources: A one-stage transfer learning framework for cross-city spatiotemporal forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1147–1155.

Zhang, Y.; Yu, X.; Wang, X.; Sun, Z.; Zhang, C.; Wang, P.; and Wang, Y. 2025b. COFlownet: Conservative constraints on flows enable high-quality candidate generation. In *The Thirteenth International Conference on Learning Representations*.

Zhao, K.; and Zhang, L. 2024. Causality-Inspired Spatial-Temporal Explanations for Dynamic Graph Neural Networks. In *The Twelfth International Conference on Learning Representations*.

Zheng, C.; Fan, X.; Wang, C.; and Qi, J. 2020. Gman: A graph multi-attention network for traffic prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 1234–1241.

Zhu, Y.; Li, X.; Wu, Z.; Wu, D.; Hu, M.; and Li, R.-H. 2024. FedTAD: Topology-aware Data-free Knowledge Distillation for Subgraph Federated Learning. In Larson, K., ed., *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 5716–5724. International Joint Conferences on Artificial Intelligence Organization. Main Track.