LEMD: Latent Environment Extrapolation and Message Disentanglement for Dynamic Graph Under Distribution Shift

Xiaoran Wei[1], **Chen Zhao**[2], **Minglai Shao**[3], **Xintao Wu**[4], **Zhong Chen**[5], **Wenjun Wang**[6], **Qin Tian**[6], **Chang Liu**[1]

[1]School of Computer Science and Technology, Tianjin University, Tianjin, China
[2]Department of Computer Science, Baylor University, Waco, Texas, USA
[3]School of New Media and Communication, Tianjin University, Tianjin, China
[4]Department of Electrical Engineering and Computer Science, University of Arkansas, Fayetteville, Arkansas, USA
[5]School of Computing, Southern Illinois University, Carbondale, IL, USA
[6]College of Artificial Intelligence, Tianjin University, Tianjin, China
{wxr_by, shaoml, wjwang, tianqin123, changliu0425}@tju.edu.cn, chen_zhao@baylor.edu, xintaowu@uark.edu, zhong.chen@cs.siu.edu

Abstract

Dynamic graph neural networks (DyGNNs) are widely used to model evolving interactions, but may fail under data distribution shift. Due to limited and unreliable interventions and insufficient disentanglement, the existing dynamic graph domain generalization approaches lead to suboptimal results. We formalize a message sufficiency causal view: a node representation is fully mediated by its received message multiset. Building on this perspective, we propose **L**atent environment **E**xtrapolation and **M**essage **D**isentanglement (**LEMD**), a novel robust representation learning framework for dynamic graph domain generalization. A message extrapolation mechanism under soft uncertainty constraints is proposed to obtain the diverse counterfactual message distributions. Causal information is disentangled fully from the messages to suppress shortcuts via a recoverable evolving disentanglement module. We further provide rigorous theoretical analysis and proofs to ensure the effectiveness of LEMD. Across all six datasets and two tasks, LEMD consistently improves over state-of-the-art dynamic graph generalization baselines under distribution shift, and achieves the best performance increase of 7.7% relative compared to the suboptimal baseline. The code of LEMD for reviewer is available at https://github.com/W-WuJi/LEMD.

1 Introduction

By coupling time, interaction topology, and features, dynamic graphs can naturally represent many systems in the real world, such as social networks [Berger-Wolf and Saia, 2006], commercial networks [Cheng *et al.*, 2022], and collaboration networks [Hughes *et al.*, 2010]. Dynamic graph neural networks (DyGNNs) have become an advanced paradigm in dynamic graph representation learning by combining temporal modeling with spatial message passing [Kumar *et al.*, 2019; Barros *et al.*, 2021]. Most DyGNNs rely on the implicit assumption that training and testing data are identically distributed [Zhang *et al.*, 2022]. But in practice, the process of data generation is affected by unobserved environments that vary across domains and evolve over time [Creager *et al.*, 2021]. Such environmental variations may affect both the topology and features of the dynamic graph [Yuan *et al.*, 2023]. Many studies on generalization have recognized the challenge of data out of distribution (OOD) and attempted to solve it [Ye *et al.*, 2025; Arjovsky *et al.*, 2020]. However, most existing methods are mainly developed for Euclidean data [Li *et al.*, 2021]. Due to the evolving of non-Euclidean topology, extending these techniques to dynamic graphs is non-trivial [Tian *et al.*, 2024; Zhang *et al.*, 2022].

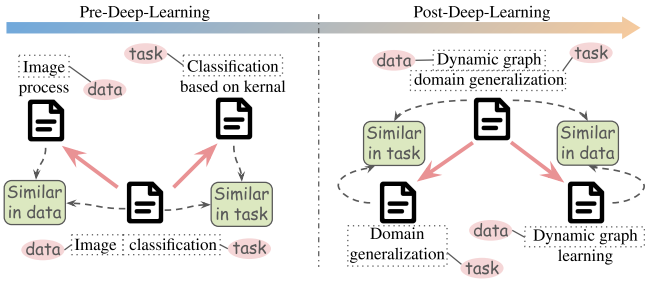

Figure 1: A toy paper citation dataset illustrating data shift: both node features and interactions may vary (pink) from pre–DL computer vision (left) to post–DL graph mining (right), while the potential relations remain invariant (green).

Although some work has made important progress in learning robust representations of dynamic graphs [Zhang *et al.*, 2024], there are still two open issues that have not been solved. Firstly, limited and unreliable environment modeling.

Considering the domain shift, the available training distribution may only cover a narrow range of conditions, while the target distribution may lie outside this support [Rosenfeld *et al.*, 2021; Lu *et al.*, 2024]. While unconstrained internal mixup of data or inferring the environment through pseudo-labeling may be effective in some cases [Zhang *et al.*, 2024; Yuan *et al.*, 2023], the invariance of learning only from a limited source may fail when the target distribution exceeds the training support [Christiansen *et al.*, 2022; Vuong *et al.*, 2025]. It is still challenging to perform environment extrapolation with effective constraints. Secondly, insufficient disentanglement across temporal and spatial dimensions. As the toy dataset shown in Figure 1, the distribution shift of dynamic graph can jointly affect features and topology across temporal and spatial dimensions, while the potential relations remain invariant. Static disentanglement only for observable factors (e.g., edges) is incomplete and leads to shortcuts leakage. A proxy object that can contain real domain information and invariant relations should be designed and allow the process of disentanglement to evolve over time.

To address these limitations, a novel approach named **LEMD** (**L**atent environment **E**xtrapolation and **M**essage **D**isentanglement) is developed for dynamic graph under distribution shift. We start from the message-sufficiency view of DyGNNs: A multiset of messages received by a node v at time t can fully mediate its features and topology while containing sufficient potential relationships between nodes. Accordingly, the distribution shift of dynamic graphs can be studied as the shift in dynamic message distributions. Thereby, we first design an uncertainty-guided dynamic message extrapolation mechanism. By focusing on the sufficient proxy of distribution shift, this mechanism produces a diverse extrapolation set of counterfactual messages under uncertainty constraints, thereby improving the model's ability to generalize to unseen domains without requiring any explicit domain labels. Fine-grained disentanglement of causal and spurious information is proposed at the message level. For each node, each message from neighbor is disentangled, taking into account both topology and features. An evolving gating network is introduced to allow the disentanglement mechanism to vary over time. Finally, a unified training objective is proposed with theoretical guarantees. Extensive experiments on six real-world datasets show that our method achieves state of the art performance in two downstream tasks under distribution shift. Our contributions are as follows:

- A novel framework for dynamic graph OOD generalization. We propose LEMD to adapt to dynamic graph data shifts that are widely observed in the real world.
- Message extrapolation constrained by uncertainty. We extrapolate counterfactual message distributions to construct diverse reliable environments without requiring explicit domain labels.
- Evolving message disentanglement. We disentangle neighbor messages into causal and spurious components identifiably, and employ a time-evolving gating mechanism to allow the temporal shift of disentanglement.
- Theoretical guarantees and empirical superiority. We provide theory for the proposed method and objective and achieve state-of-the-art on six real-world datasets across two downstream tasks under distribution shift.

2 Related Work

Dynamic graph representation learning. DyGNNs have developed into a class of temporal representation learners that unify message passing with temporal modeling [Ma *et al.*, 2026]. In snapshot settings, a prevalent paradigm is to encode each snapshot with a shared GNN and then capture temporal dependence of node states via recurrent units or temporal attention [Sankar *et al.*, 2020], offering a practical balance among expressivity, efficiency, and long range modeling. Recently, work such as EvolveGCN [Pareja *et al.*, 2020] and SEIGN [Qin *et al.*, 2023] have developed temporal modeling from the evolution of node representations to the evolution of both parameters and representations. Most dynamic graph representation learning methods follow the identical distribution assumption, which results in their limited generalization ability under the data distribution shift.

Causal learning for OOD generalization. A significant insight for OOD generalization is that robustness under distribution shift requires disentangling and d-separating the spurious factors, which means predictive shortcuts that hold in training but break in testing [Komanduri *et al.*, 2024]. Causal intervention methods generate counterfactual variations to discourage dependence on shortcuts [Sun *et al.*, 2021]. Invariant learning enforces stability of representations across environments, whether provided by domain labels [Ganin *et al.*, 2016] or inferred as latent partitions [Creager *et al.*, 2021]. Despite success on Euclidean data, the distribution shift of dynamic graphs affects evolving topology and features, making intervention and disentanglement challenging.

OOD generalization on graphs. Building on the intervention and invariance principles, recent work has begun to study OOD generalization for dynamic graphs [Tian *et al.*, 2026; Tian *et al.*, 2025; Ma *et al.*, 2025]. DIDA [Zhang *et al.*, 2022] disentangles spatial-temporal patterns via flipping the attention mechanism and constructs interventions by sampling and reassembling neighbors across spatial and temporal. SILD [Zhang *et al.*, 2023] learns disentangled spectral masks with invariant filtering to suppress domain sensitive frequency components. DyAug [Zhang *et al.*, 2024] extends graph rationalization to dynamic settings by separating rationale subgraphs over time and performing temporally consistent regularization. EAGLE [Yuan *et al.*, 2023] focuses on environments inference, combining disentanglement with a diverse instantiation mechanism. Despite this progress, existing methods often struggle to produce faithful, extrapolative counterfactuals and to disentangle causal and spurious factors sufficiently. Accordingly, we treat messages as the sufficient mediator and combine reasonable environment extrapolation with an evolving and identifiable message disentanglement to achieve OOD generalization of DyGNNs.

3 Preliminaries

3.1 Dynamic graph domain generalization

Dynamic graphs. A dynamic graph is a sequence of graphs $\mathcal{G}_{1:T} = \{G_t\}_{t=1}^{T} = \{(\mathbf{V}_t, \mathbf{A}_t, \mathbf{X}_t)\}_{t=1}^{T}$. Each snapshot G_t

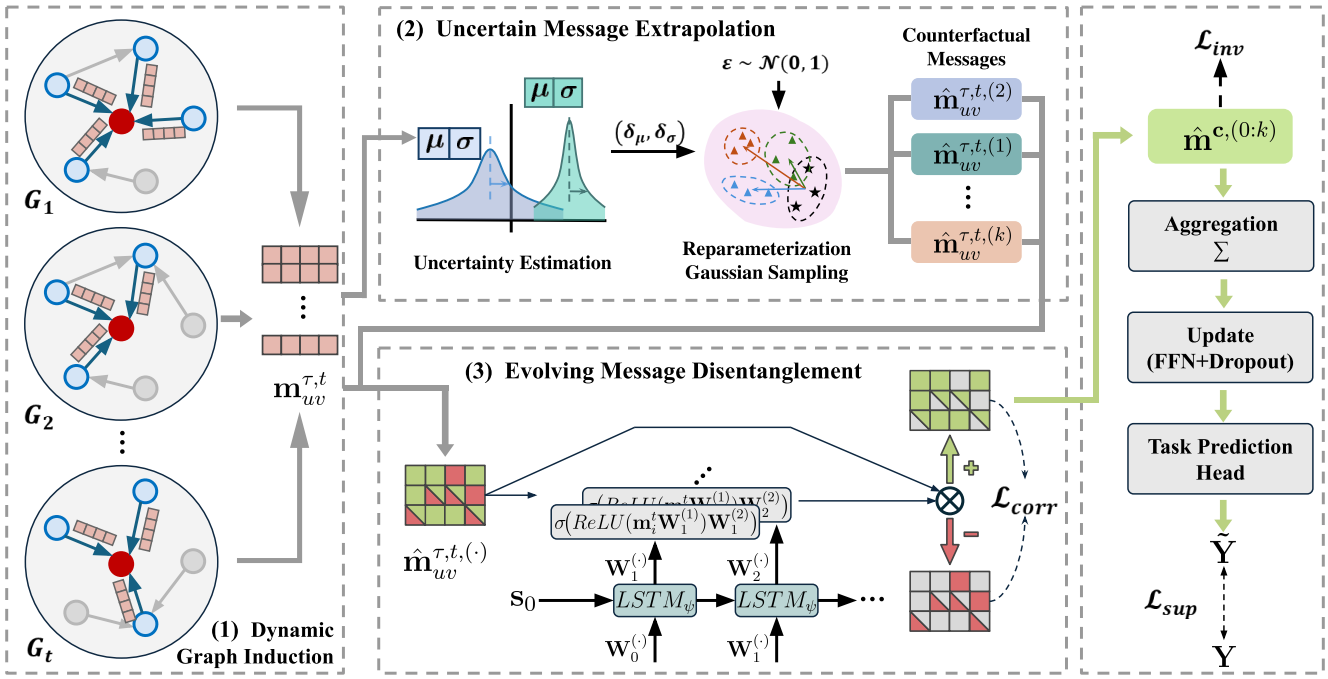

Figure 2: Overview of LEMD. (1) Messages are first induced from the dynamic graph sequence. Then, (2) K sets of counterfactual messages are obtained by extrapolating with uncertainty constraints. Combined with the original messages set, (3) all messages are divided into causal (green) and spurious (red) components through evolving message disentanglement with correlation loss. Finally, aggregation, updating, and prediction of causal messages are performed with invariance and supervised loss.

contains a node set $\mathbf{V}_t$, an adjacency matrix $\mathbf{A}_t$, and a node features matrix $\mathbf{X}_t \in \mathbb{R}^{|\mathbf{V}_t| \times d_x}$. The adjacency matrix is defined as: $(\mathbf{A}_t)_{uv} = \mathbb{I}[\, u \to v \text{ at time } t\,]$, for $u, v \in \mathbf{V}_t$. For a node $v \in \mathbf{V}_t$, we denote its incoming neighbor set at t as $\mathcal{N}_v^t = \{\, u \in \mathbf{V}_\tau : (\mathbf{A}_\tau)_{uv} = 1 \mid \exists \tau \le t\}$.

Domain generalization on dynamic graphs. Dynamic graph domain generalization posits latent environments $e \in \mathcal{E}$ that govern the data generation process and induce distribution shifts in both topology and features. Conditioned on e, samples follow the joint distribution $\mathbb{P}(\mathcal{G}_{1:T}, y \mid e)$. Unobserved environment distribution differs: $e \sim \mathbb{P}_{\text{tr}}(e)$ at training and $e \sim \mathbb{P}_{\text{te}}(e)$ at testing [Wu *et al.*, 2024].

We learn f_θ that minimizes the expected testing loss:

$$\min_{\theta} \; \mathbb{E}_{(\mathcal{G}_{1:T}, y) \sim \mathbb{P}_{\text{te}}} \Big[\ell\big(f_\theta(\mathcal{G}_{1:T}), y\big) \Big], \tag{1}$$

where $\mathbb{P}_{\text{te}}$ is unknown during training and the environment index e is unobserved. Hence, the key challenge is to build robustness to unseen distribution shifts without relying on explicit environment labels.

3.2 A message passing pipeline of DyGNNs

Most DyGNNs implement a spatiotemporal message passing pipeline: exchange messages along edges and update node representations by aggregating messages [Gilmer *et al.*, 2017]. Let $\mathbf{h}_v^t \in \mathbb{R}^d$ be the hidden state of node v at time t. We apply an attention-style DyGNN layer, where the message is defined as:

$$\mathbf{m}_{uv}^{\tau,t} = \underset{u \in \mathcal{N}_v^t}{\text{softmax}} \left(\frac{\langle \mathbf{q}_v^t, \mathbf{k}_u^\tau \rangle}{\sqrt{d_k}} \right) \cdot \mathbf{v}_u^\tau, \tag{2}$$

where $\mathbf{q}_v^t = \mathbf{W}_q \mathbf{h}_v^t$, $\mathbf{k}_u^\tau = \mathbf{W}_k \mathbf{h}_u^\tau$ and $\mathbf{v}_u^\tau = \mathbf{W}_v \mathbf{h}_u^\tau$. And $\mathbf{W}_{(\cdot)}$ denotes learnable projections, d_k is the dimension of the latent variable $\mathbf{k}_u^\tau$. Collecting all directed messages received by v at time t gives a message multiset $\mathcal{M}_v^t := \{\mathbf{m}_{uv}^{\tau,t} \mid u \in \mathcal{N}_v^t\}$. Aggregation and update are then applied:

$$\tilde{\mathbf{h}}_v^t = \text{UPD}\left(\text{AGG}\left(\mathcal{M}_v^t\right)\right), \tag{3}$$

where $\text{AGG}(\mathcal{M}_v^t) \in \mathbb{R}^d$ and $\text{UPD}(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$.

4 Method

We propose LEMD as shown in Figure 2. Following the message fully mediated perspective in Sec. 4.1, the dynamic graph is induced into message multisets. Uncertain extrapolation (Sec. 4.2) and evolving causal disentanglement (Sec. 4.3) are applied for message multisets to generate diverse distributional shifts and suppress spurious shortcuts. Rigorous theories and proofs are given to support method validity.

4.1 A message-mediated causal view

Message sufficiency as explicit mediation. Given the message passing pipeline in Sec. 3.2, a DyGNN layer constructs the received message multiset $\mathcal{M}_v^t$ and then updates node

states by $\tilde{\mathbf{h}}_v^t = \mathrm{UPD}(\mathrm{AGG}(\mathcal{M}_v^t))$. Let $\mathrm{Pred}(\cdot)$ be an arbitrary task head applied on $\tilde{\mathbf{h}}_v^t$. Then for each prediction $\tilde{y}_v^t$:

$$\tilde{y}_v^t = g_\theta(\mathcal{M}_v^t), \qquad g_\theta \triangleq \mathrm{Pred} \circ \mathrm{UPD} \circ \mathrm{AGG}, \quad (4)$$

which formalizes a mediation principle: for fixed model parameters, the dependence on the dynamic graph history $(\mathbf{A}_{1:t}, \mathbf{X}_{1:t})$ is fully mediated by the message multiset $\mathcal{M}_v^t$.

Induced message distributions under latent environments. Recall that environments $e \in \mathcal{E}$ govern the data-generating process (Sec. 3.1). Since $\mathcal{M}_v^t$ is constructed from $(\mathbf{A}_{1:t}, \mathbf{X}_{1:t})$ given fixed model, each environment e induces a corresponding joint distribution in message space, denoted by $\mathbb{P}(\mathcal{M}_v^t, y \mid e)$. Accordingly, for node v at time t, the test message distribution is a mixture over environments:

$$\mathbb{P}_{\mathrm{te}}^{\mathcal{M}}(\mathcal{M}_v^t, y) = \int \mathbb{P}(\mathcal{M}_v^t, y \mid e)\, d\mathbb{P}_{\mathrm{te}}(e), \quad (5)$$

and similarly $\mathbb{P}_{\mathrm{tr}}^{\mathcal{M}}$ is defined by $\mathbb{P}_{\mathrm{tr}}(e)$. Therefore, Eq. (1) admits an equivalent message-space view:

$$\min_\theta \ \mathbb{E}_{(v,t)}\, \mathbb{E}_{(\mathcal{M}_v^t, y) \sim \mathbb{P}_{\mathrm{te}}^{\mathcal{M}}} \Big[\ell\big(g_\theta(\mathcal{M}_v^t), y\big) \Big]. \quad (6)$$

Eq. (6) illustrates that domain shifts on $(\mathbf{A}_{1:t}, \mathbf{X}_{1:t})$ are equivalently captured as shifts in the distributions $\mathbb{P}(\mathcal{M}_v^t \mid e)$.

A structural causal model with label invariance. A structural view is adopted in which a stable semantic factor and an environment-specific bias factor jointly shape the observed dynamic graph and thus the induced messages. For each t, let $\mathbf{C}_t$ denote stable semantics and $\mathbf{S}_t$ denote environment-dependent biases. The key objects in our setting admit the following factorization (with $\mathcal{M}_v^t$ as an explicit mediator):

$$\begin{aligned} &\mathbb{P}(y_v^t, \mathcal{M}_v^t, \mathbf{A}_{1:t}, \mathbf{X}_{1:t} \mid e) = \\ &\quad \mathbb{P}(y_v^t \mid \mathbf{C}_t)\, \mathbb{P}(\mathcal{M}_v^t \mid \mathbf{A}_{1:t}, \mathbf{X}_{1:t})\, \mathbb{P}(\mathbf{A}_{1:t}, \mathbf{X}_{1:t} \mid e, \mathbf{C}_t, \mathbf{S}_t). \end{aligned} \quad (7)$$

We assume counterfactual label invariance with respect to environment shifts:

$$\mathbb{P}(y \mid \mathbf{C}_t, e) = \mathbb{P}(y \mid \mathbf{C}_t), \quad (8)$$

i.e., changing environment can affect interactions and features, while the causal mechanism remains preserved.

4.2 Uncertainty-guided environment extrapolation

Counterfactuals in message space. Under the message-mediated view in Eq. (6), unseen environments perform as shifts in the induced message distributions. Since the environment is latent, we proxy such shifts by constructing counterfactuals in message space around the observed messages.

For node v at time t, we extrapolate from its received message multiset $\mathcal{M}_v^t$. Let $\mathcal{S}_v^t : \mathcal{M}_v^t \to \mathbb{R}^q$ be a parameter-free summary operator and $\mathcal{S}(\mathcal{M})$ be the collection of messages over all (v,t). We estimate an uncertainty descriptor $\boldsymbol{\rho} = \mathcal{U}(\mathcal{S}(\mathcal{M}))$ from observed summaries, where $\boldsymbol{\rho} \in \mathbb{R}_+^q$ and $\mathcal{U}$ quantifies the feasible variation scale supported by the observed messages. This yields an uncertainty guided soft region of counterfactual message distributions:

$$\Omega(\boldsymbol{\rho}) = \Big\{ \hat{\mathcal{M}} : \mathcal{S}(\hat{\mathcal{M}}) \sim \mathbb{Q}(\cdot \mid \mathcal{S}(\mathcal{M}), \boldsymbol{\rho}) \Big\}, \quad (9)$$

where $\mathbb{Q}$ is a density distribution centered at the factual summary with scale $\boldsymbol{\rho}$. We sample K counterfactual message multisets $\{\hat{\mathcal{M}}^{(k)}\}_{k=1}^K$ from $\Omega(\boldsymbol{\rho})$ via the reparameterized trick with moment matching.

Reparameterized moment matching construction. According to Theorem 1, the smaller dimension q of $\boldsymbol{\rho}$, the lower the upper bound of expectation. The low order moments of a feature can effectively summarize its information. Studies have shown that low order moments of messages can reflect environmental information to a certain extent [Li *et al.*, 2020]. The low order moments of $\mathcal{M}_v^t$ are used as its summary:

$$\mathcal{S}_v^t(\mathcal{M}_v^t) = (\mu_v^t, \sigma_v^t), \quad (10)$$

where μ_v^t and σ_v^t are the mean and standard deviation within $\mathcal{M}_v^t$ respectively. Then, the scale of feasible extrapolation $\boldsymbol{\rho}$ is estimated by analyzing the dispersion of $\mathcal{S}_v^t(\mathcal{M}_v^t)$:

$$\boldsymbol{\rho} = (\delta_{\boldsymbol{\mu}}, \delta_{\boldsymbol{\sigma}}), \quad (11)$$

where the $\delta_{\boldsymbol{\mu}}$ and $\delta_{\boldsymbol{\sigma}}$ are the standard deviation of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ across all time and nodes. $\boldsymbol{\rho}$ measures how heterogeneous all the incoming messages are.

According to the central limit theorem, $\mathbb{Q}$ is naturally realized as a bivariate diagonal Gaussian distribution in the summary space:

$$\mathbb{Q}(\cdot \mid \mathcal{S}(\mathcal{M}), \boldsymbol{\rho}) = \mathcal{N}\left((\mu_v^t, \sigma_v^t), \mathrm{diag}(\delta_{\boldsymbol{\mu}}^2, \delta_{\boldsymbol{\sigma}}^2)\right). \quad (12)$$

K counterfactual summaries are sampled from $\mathbb{Q}$, i.e. $(\tilde{\mu}_v^{t,(k)}, \tilde{\sigma}_v^{t,(k)}) \sim \mathbb{Q}$. In actual implementation, to ensure gradient propagation, the reparameterization trick is applied:

$$\tilde{\mu}_v^{t,(k)} = \mu_v^t + \delta_{\boldsymbol{\mu}} \epsilon_{\mu,v}^{t,(k)}, \qquad \tilde{\sigma}_v^{t,(k)} = \sigma_v^t + \delta_{\boldsymbol{\sigma}} \epsilon_{\sigma,v}^{t,(k)}, \quad (13)$$

where $\epsilon_{\mu,v}^{t,(k)}, \epsilon_{\sigma,v}^{t,(k)} \sim \mathcal{N}(0,1)$. σ is cut off by a certain minimum positive value to avoid $\sigma < 0$.

Given the sampled k-th counterfactual summary $(\tilde{\mu}_v^{t,(k)}, \tilde{\sigma}_v^{t,(k)})$, the counterfactual message multiset is generated by an affine moment transform:

$$\hat{\mathcal{M}}_v^{t,(k)} = \left\{ \left(\frac{\mathbf{m}_i - \mu_v^t}{\sigma_v^t} \right) \tilde{\sigma}_v^{t,(k)} + \tilde{\mu}_v^{t,(k)} \right\}_{i=1}^{|\mathcal{M}_v^t|}, \quad (14)$$

where i is the index of message within $\mathcal{M}_v^t$. This counterfactual generation method extrapolates messages under control while preserving their normalized structure, thereby providing a proxy for shifts induced by unseen environments.

Counterfactual invariance regularization. Let $\mathbf{m}_i^{(0)}$ denote a factual directed message and $\hat{\mathbf{m}}_i^{(k)}$ $(k \geq 1)$ its counterfactual variants generated by Eq. (9)–(14). Let $\mathbf{m}_i^{c,(k)}$ be the corresponding causal component produced by the disentangler in Sec. 4.3. We enforce counterfactual invariance by penalizing the variance across sampled environments:

$$\mathcal{L}_{\mathrm{inv}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \mathrm{Var}_k\Big(\mathbf{m}_i^{c,(k)}\Big), \quad (15)$$

where $\mathrm{Var}_k(\cdot)$ denotes the mean squared deviation over $k \in \{0, \ldots, K\}$ and $\mathcal{M}$ is the collection of all messages.

4.3 Evolving message disentanglement

Disentangling the messages with evolving parameters. Our causal formulation assumes a stable semantic factor $\mathbf{C}_t$ that fully determines the target mechanism as shown in

Eq. (8), while the latent environment e affects topology and features and hence the induced messages through spurious factors. Meanwhile, message sufficiency view states that predictions are mediated by the received messages, $\tilde{y}_v^t = g_\theta(\mathcal{M}_v^t)$. This means that the message space can be used as a proxy for observing distributional shifts. And the causal component that is predictive of y but independent of environmental changes can be disentangled from each message.

Concretely, given the message indexed by i at target time t, a counterfactual family $\{\mathbf{m}_i^{(k)}\}_{k=0}^K$ is constructed according to Sec. 4.2, where $\mathbf{m}_i^{(0)} \triangleq \mathbf{m}_i$ is factual and $\mathbf{m}_i^{(k)}$ $(k \geq 1)$ are counterfactual messages. Then a time-evolving network D_t is introduced to disentangle a causal component $\mathbf{m}_i^{c,(k)}$ that remains stable across k and a complementary spurious component that practically independent of y.

A time varying generation model in message space. We set that each message is generated from a time-dependent mixing between stable semantics and spurious biases:

$$\mathbf{m}_i^{(k)} = \Psi_t(\mathbf{C}_t, \mathbf{S}_t^{(k)}), \tag{16}$$

where $\mathbf{C}_t$ is shared across k and $\mathbf{S}_t^{(k)}$ varies with the counterfactuals. The mixing mechanism Ψ_t is indexed by t to capture temporal distribution shift.

Recoverable causal–spurious disentanglement. A time specific network D_t is constructed to disentangle each message into a causal and a spurious component:

$$(\mathbf{m}_i^{c,(k)}, \mathbf{m}_i^{s,(k)}) = D_t\Big(\mathbf{m}_i^{(k)}\Big), \tag{17}$$

with recover operator R_t:

$$R_t\Big(\mathbf{m}_i^{c,(k)}, \mathbf{m}_i^{s,(k)}\Big) = \mathbf{m}_i^{(k)}. \tag{18}$$

Recoverability of D_t prevents trivial solutions that discard information, while counterfactual invariance (Eq. (15)) encourages $\mathbf{m}^{c,(k)}$ to represent stable semantics.

Practical instantiation. Considering complexity and recoverability requirements, D_t is instantiated as an element-wise gate $\boldsymbol{\alpha}_i^t \in (0,1)^d$ which is calculated from the message by an evolving gate network:

$$\boldsymbol{\alpha}_i^{t,(k)} = \sigma\big(\mathrm{ReLU}(\mathbf{m}_i^{(k)}\mathbf{W}_t^{(1)})\mathbf{W}_t^{(2)}\big), \tag{19}$$

where σ is the Sigmoid function and $\mathbf{W}_t^{(1)}, \mathbf{W}_t^{(2)}$ are time-specific parameters. This gate is then applied to disentangle every counterfactual variant:

$$\mathbf{m}_i^{c,(k)} = \boldsymbol{\alpha}_i^{t,(k)} \odot \mathbf{m}_i^{(k)}, \qquad \mathbf{m}_i^{s,(k)} = \Big(\mathbf{1} - \boldsymbol{\alpha}_i^{t,(k)}\Big) \odot \mathbf{m}_i^{(k)}, \tag{20}$$

where $\odot$ is the Hadamard product. Eq. (18) holds with R_t being simple addition $\mathbf{m}_i^{(k)} = \mathbf{m}_i^{c,(k)} + \mathbf{m}_i^{s,(k)}$.

To track the shift in Ψ_t, we let the gate parameters evolve with t by an evolving parameter generator. $\{\mathbf{W}_t^{(1)}, \mathbf{W}_t^{(2)}\}$ are generated through a recurrent generator in parameter space:

$$\mathbf{W}_t^{(\cdot)},\ \mathbf{s}_t^{(\cdot)} = \mathrm{LSTM}_\psi(\mathbf{W}_{t-1}^{(\cdot)}, \mathbf{s}_{t-1}^{(\cdot)}), \tag{21}$$

where $\mathbf{s}_t$ is the recurrent state and ψ are generator parameters. This induces a smooth and associated temporal prior over the disentanglement mechanism without requiring an independent disentangler at each t.

Correlation regularization. To enforce the independence between $\mathbf{m}^c$ and $\mathbf{m}^s$, we penalize their dependence using a centered correlation:

$$\mathcal{L}_{\mathrm{cor}} = \frac{1}{|\mathcal{M}|}\sum_{i=1}^{|\mathcal{M}|} \frac{\Big|\big(\mathbf{m}_i^c - \bar{\mathbf{m}}^c\big)^\top \big(\mathbf{m}_i^s - \bar{\mathbf{m}}^s\big)\Big|}{\sqrt{\|\mathbf{m}_i^c - \bar{\mathbf{m}}^c\|_2^2}\,\sqrt{\|\mathbf{m}_i^s - \bar{\mathbf{m}}^s\|_2^2}}. \tag{22}$$

with $\bar{\mathbf{m}}^c = \frac{1}{|\mathcal{M}|}\sum_{i=1}^{|\mathcal{M}|}\mathbf{m}_i^c$ and $\bar{\mathbf{m}}^s = \frac{1}{|\mathcal{M}|}\sum_{i=1}^{|\mathcal{M}|}\mathbf{m}_i^s$. Theorem 2 explains the contribution of independence to identifiable causal components.

The final objective. For each (v,t), we feed the raw causal message multiset $\mathcal{M}_v^{c,t,(0)}$ to the DyGNN update via $\tilde{\mathbf{h}}_v^t = \mathrm{UPD}(\mathrm{AGG}(\mathcal{M}_v^{c,t,(0)}))$. Then the supervised loss function is computed as

$$\mathcal{L}_{\mathrm{sup}} = \frac{1}{|V||T|}\sum_{v,t}\ell(\mathrm{Pred}(\tilde{\mathbf{h}}_v^t), y_v^t), \tag{23}$$

Hence the complete training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{sup}} + \lambda_{\mathrm{inv}}\mathcal{L}_{\mathrm{inv}} + \lambda_{\mathrm{cor}}\mathcal{L}_{\mathrm{cor}}, \tag{24}$$

where λ_{inv} and λ_{cor} are coefficients of counterfactual invariance and causal–spurious separation respectively. Theorem 2 provides theoretical guarantees for the approximate identifiability of disentanglement under this optimization objective.

5 Analysis

Two theories are proposed to support the effectiveness of messages extrapolation and evolving disentanglement. All the theories are rigorously proved in Appendix A.3.

Theorem 1 (Counterfactual causal-message invariance with uncertainty-guided probabilistic coverage)**.** *Assume: (i) the loss is bounded $0 \leq \ell \leq B_\ell$; (ii) for all y, the mapping $\mathcal{M} \mapsto \ell(g_\theta(\mathcal{M}), y)$ is L-Lipschitz w.r.t. the message input; (iii) the statistic-to-distribution regularity and target typicality assumptions hold. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$ over the draw of N training instances and the uncertainty-guided sampling of K counterfactual summaries, the following holds:*

$$\begin{aligned}\mathbb{E}_{(\mathcal{M}^c,y)\sim\mathbb{P}_{\mathrm{te}}}\left[\ell\left(g_\theta(\mathcal{M}^c), y\right)\right] \ \leq\ \mathcal{L}_{\mathrm{sup}}(\theta) + \\ L\sqrt{\mathcal{L}_{\mathrm{inv}}(\theta)} + \varepsilon_N(\delta/2) + L\,C_{\mathcal{S}}\,\|\boldsymbol{\rho}\|_\infty\, r_{K,\delta/2}(r),\end{aligned} \tag{25}$$

where $\varepsilon_N(\cdot)$ is the standard uniform-convergence term and $r_{K,\delta}(r)$ is a Gaussian K-sample covering radius.

Remark. The bound above decomposes into: the empirical supervised risk on factual causal messages, a stability term controlled by the counterfactual invariance loss, a standard generalization term, and a probabilistic coverage term quantifying how likely the K sampled counterfactual summaries contain one that is close to the target distribution. For fixed q and typicality radius r, the covering radius satisfies $r_{K,\delta}(\tau) = \tilde{O}(K^{-1/q})$. As a result, increasing K and decreasing L_{sup} and L_{inv} may tighten the bound.

Theorem 1 gives perspective from the upper bound on test loss expectation, which explains the necessity for L_{sup} and L_{inv}, and illustrates the effectiveness of uncertainty-constrained message extrapolation for reducing the coverage error term and thus the upper bound on test loss expectation.

	YELP		MOOC-ACT		COLLAB		WIKI		REDDIT	
	w/o OOD	**OOD**	w/o OOD	**OOD**	w/o OOD	**OOD**	w/o OOD	**OOD**	w/o OOD	**OOD**
GCRN	68.59±1.05	54.68±7.59	76.28±0.51	64.35±1.24	82.78±0.54	69.72±0.45	80.52±0.21	70.97±0.67	81.08±0.70	73.69±0.33
DySAT	78.87±0.57	66.09±1.42	78.52±0.40	66.55±1.21	88.77±0.23	76.59±0.20	91.34±0.64	74.33±0.30	92.16±1.62	79.78±0.42
EGCN	78.21±0.03	53.82±2.06	74.55±0.33	63.17±1.05	86.62±0.95	76.15±0.91	91.12±0.75	72.54±0.33	93.41±0.99	78.23±0.19
SEIGN	80.72±0.39	67.19±0.84	84.57±0.18	75.90±0.82	92.19±0.21	80.68±0.72	95.23±1.02	78.91±0.81	97.81±0.96	82.80±0.31
IRM	66.49±10.8	56.02±16.1	80.02±0.57	69.19±1.35	87.96±0.90	75.42±0.87	80.17±1.89	71.91±2.54	77.62±2.48	70.37±1.88
VREx	79.04±0.16	66.41±1.87	81.72±0.29	76.24±1.09	88.31±0.32	76.24±0.77	84.21±0.66	73.72±0.95	83.04±0.49	75.90±0.59
GroupD	79.38±0.42	66.97±0.61	81.50±0.53	75.92±1.62	88.76±0.12	76.33±0.29	82.67±0.17	72.81±0.61	82.17±0.85	75.55±0.92
DIDA	78.22±0.40	75.92±0.90	89.84±0.82	78.64±0.97	91.97±0.05	81.87±0.40	94.81±0.62	79.71±0.59	98.09±0.78	85.17±0.76
IB-D^2	78.73±1.91	76.11±1.14	90.16±0.44	78.77±0.65	91.92±0.36	80.46±0.46	94.95±1.06	79.02±0.79	97.92±0.43	85.99±0.91
EAGLE	78.97±0.31	77.26±0.74	92.37±0.53	82.70±0.72	92.45±0.21	84.41±0.87	95.16±0.94	81.84±0.26	89.47±0.57	85.22±0.81
EvoGD	79.71±1.26	78.74±1.50	92.68±0.01	82.94±0.03	93.44±0.05	84.46±0.03	95.89±0.14	82.16±0.19	93.55±0.22	88.61±0.28
DGIB	76.88±0.20	72.56±0.74	94.49±0.20	81.32±0.44	92.17±0.20	83.09±0.56	90.08±0.81	84.02±0.60	90.15±0.79	89.58±0.27
SILD	83.34±0.52	78.65±2.22	93.13±0.91	83.20±1.24	**94.02**±0.37	84.09±0.16	96.69±1.06	83.87±0.79	98.55±0.82	89.21±0.12
DyAug	82.84±0.35	76.50±0.65	85.46±0.29	82.05±1.18	93.62±0.42	83.11±0.56	95.03±0.46	80.09±0.76	92.26±0.85	86.94±0.94
LEMD	**87.91**±0.45	**82.98**±0.88	**94.95**±0.50	**89.64**±0.32	93.91±0.26	**84.83**±0.37	**98.03**±0.48	**90.37**±0.21	**99.06**±0.08	**95.29**±0.14

Table 1: Results (AUC%) with standard deviation of link prediction. Blue background with bold font indicates the best results, and light blue background with underline font indicates the suboptimal results. LEMD achieve state of the art on five the datasets under the OOD setting.

Theorem 2 (Approximate identifiability of invariant causal messages)**.** *Fix a time step t and assume the augmented messages follow an invertible mixing model Ψ_t. Let D_t be any recoverable disentangler as shown in Eq.* (17)*. If the learned causal part is approximately counterfactually invariant,*

$$\mathbb{E}\Big[\mathrm{Var}_{k\in\{0,\dots,K\}}\Big(\mathbf{m}_t^{c,(k)} \mid \mathbf{C}_t\Big)\Big] \le \varepsilon_{\mathrm{inv}}, \tag{26}$$

and regularity holds non-degenerate spurious support and continuity, then there exists a decoder h_t such that

$$\mathbb{E}\left[\|h_t(\mathbf{m}_t^{c,(k)}) - \mathbf{C}_t\|_2\right] = O(\sqrt{\varepsilon_{\mathrm{inv}}})\,. \tag{27}$$

In particular, as $\varepsilon_{\mathrm{inv}} \to 0$, $\mathbf{m}_t^c$ identifies $\mathbf{C}_t$ up to an invertible transform on its support.

Remark. The result is pointwise in t, hence compatible with time-evolving disentanglement; full constants and proof are provided in Appendix.

Theorem 2 formalizes that counterfactual invariance is the key condition for recovering stable semantics from the learned causal messages. The correlation penalty $\mathcal{L}_{\mathrm{cor}}$ in Eq. (22) serves a complementary purpose: it discourages linear leakage between $\mathbf{m}^c$ and $\mathbf{m}^s$, thereby improving the operational cleanliness of the split and stabilizing optimization under drift. A formal statement is given in Appendix.

6 Experiments

6.1 Experiment Setup

Datasets. We evaluate the performance of LEMD on dynamic graphs by six real world datasets under two tasks: future link prediction and node classification. For each dataset, we construct a dedicated distribution shift setting. `COLLAB` [Tang *et al.*, 2012] is a citation network dataset for 16 years, comprising paper authors and their citation relations. Edges associated with the *Data Mining* category are regarded as distribution shift data. `YELP` [Sankar *et al.*, 2020] is a business-review dataset covering 24 months, consisting of interactions (reviews) between customers and businesses. Reviews for the *Pizza* category are regarded as distribution shift data. `MOOC-ACT` [Kumar *et al.*, 2019] contains student interaction records on a MOOC platform over 30 days. `WIKI` [Kumar *et al.*, 2019] records page-editing activities over 31 days on Wikipedia. `REDDIT` [Kumar *et al.*, 2019] consists of posts made by users on Reddit subreddits over a month. According to general settings of strong baselines [Yuan *et al.*, 2023], for MOOC-ACT, WIKI, and REDDIT, we perform K-means clustering on the interactions therein and randomly select a class as distribution shift. The above five datasets are used to evaluate future link prediction under distribution shifts. In addition, we adopt `AMINER` [Sinha *et al.*, 2015; Tang *et al.*, 2008] for node classification. AMINER is a long-term citation network spanning 18 years. Since the dynamics of the graph may evolve over time, we regard the final three years as distribution shift data.

Baselines. As in Sec. 2, our baseline includes three classes of methods. (i) Domain generalization methods contain IRM [Arjovsky *et al.*, 2020], V-REx [Krueger *et al.*, 2021] and GroupDRO [Sagawa *et al.*, 2020]. (ii) Dynamic graph learning methods contain GCRN [Seo *et al.*, 2018], DySAT [Sankar *et al.*, 2020], EvolveGCN [Pareja *et al.*, 2020] and SEIGN [Qin *et al.*, 2023]. (iii) OOD generalization methods on dynamic graphs contain DIDA [Zhang *et al.*, 2022], IB-D^2 [Wang *et al.*, 2025], EAGLE [Yuan *et al.*, 2023], EvoGD [Sun *et al.*, 2025], DGIB [Yuan *et al.*, 2024], SILD [Zhang *et al.*, 2023] and DyAug [Zhang *et al.*, 2024]. The specific method is introduced in the Appendix.

Implementation details. For the future link prediction task, the dataset is first divided into training, validation, and test sets based on the temporal order, and then categorized into an OOD section and a without OOD section based on the different domains of the edges. All pairs of nodes with existing edges and equally sampled non-existing edges in a

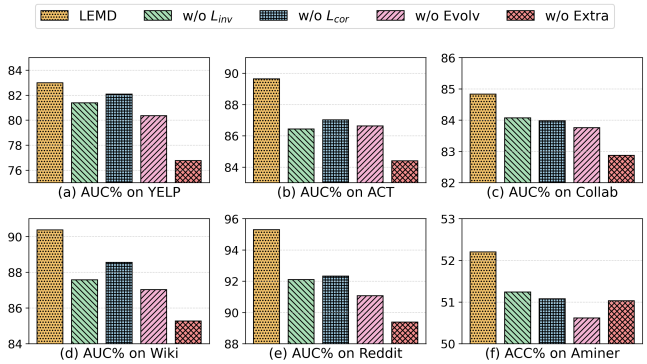

Figure 3: Ablation study on six datasets with two tasks.

	AMINER15	AMINER16	AMINER17
GCRN	47.96±1.12	51.33±0.62	42.93±0.71
EGCN	44.14±1.12	46.28±1.84	37.71±1.84
DySAT	48.41±0.81	49.76±0.96	42.39±0.62
IRM	48.44±0.13	50.18±0.73	42.40±0.27
VREx	48.70±0.73	49.24±0.27	42.59±0.37
GroupD	48.73±0.61	49.74±0.26	42.80±0.36
DIDA	50.34±0.81	51.43±0.27	44.69±0.06
SILD	52.35±1.04	54.11±0.62	45.54±1.19
LEMD	**53.79**±0.25	**55.70**±0.32	**47.12**±0.13

Table 2: Results (ACC%) with standard deviation of node classification. LEMD achieved the state of the art on all OOD settings.

future time slice are used as the ground truth of positive and negative samples. Test set results for the OOD section and w/o OOD section are reported, using AUC% as the evaluation metric. For the node classification task, model is used to predict the venues of the papers. Results of future three years are reported separately as OOD data, using percentage of accuracy (ACC%) as the evaluation metric.

6.2 Results

Link prediction under domain shifts. Experiments on future link prediction task under domain shifts are conducted on five widely used real world dynamic graph datasets. The results are shown in Table **??**, where blue background with bold font indicates the optimal results, and light blue background with underline indicates the suboptimal results. Across all five datasets, SILD achieves the largest number of suboptimal results, which may be attributed to its strong capability of spectral disentanglement. Comparing all the most advanced baselines, LEMD achieves state of the art on all datasets under OOD setting. On the MOOC-ACT dataset, LEMD recorded the largest performance increase of 7.7% compared to the suboptimal baseline. LEMD achieves state-of-the-art performance on four datasets under the w/o OOD setting. Only under the w/o OOD setting on the COLLAB dataset does LEMD underperform the best baseline, SILD, by a margin of 0.1%. These results suggest that LEMD has strong representation and generalization ability across domain.

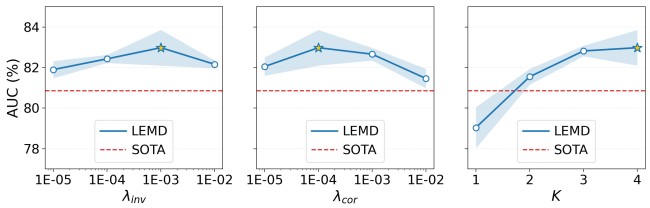

Figure 4: Sensitivity of λ_{inv}, λ_{cor} and K.

Node classification under domain shifts. Experiments on node classification task under domain shifts on AMINER dataset. Considering the impact of the outbreak of deep learning, the citation pattern of papers since 2015 has major changed. As the results shown in Table 2, LEMD achieves the state of the art in accuracy of node classification prediction in all three years. Averaged over three years, LEMD's node classification accuracy is 3% higher than the suboptimal SILD. These results demonstrate LEMD's universal representational capacity and its strong generalization across domain on a broader range of tasks.

Ablation Study. Ablation experiments are performed on our proposed module on six datasets within two tasks. Each subfigure in Figure 3 shows, in turn, complete LEMD, without $\mathcal{L}_{\text{inv}}$, without $\mathcal{L}_{\text{cor}}$, without to **Evolv**e disentanglement, and without message **Extra**polation. The ablation results show that the removal of the two loss constraints and the two modules will respectively drop the model performance. For link prediction tasks under distribution shifts, it is usually the removal of message extrapolation impacts most severely, which reveals the effective modeling of message distribution extrapolation for data across domains. For node classification tasks under distribution shifts, the removal of evolve disentanglement impact most severely. This result is consistent with the context that the AMINER dataset is primarily affected by deep learning breakout, which creates shifts alone time.

Hyperparameter Sensitivity. We performed sensitivity analysis on all three most important hyperparameters λ_{inv}, λ_{cor} and K on the most representative dataset YELP, which results are shown in Figure 4. It can be found that moderate invariance regularity and correlation regularity (e.g. $\lambda_{\text{inv}} = 1e-3$, $\lambda_{\text{cor}} = 1e-4$) give the best model performance. Moreover, the model performance increases with increasing K, but there is a marginal decreasing effect.

7 Conclusion

This paper investigates dynamic graph domain generalization under distribution shifts, and motivates a message-sufficiency causal perspective in which predictions are fully mediated by received messages. We propose LEMD, a framework that improves robustness by intervening in message space. LEMD combines message extrapolation under uncertainty constraints to sample counterfactual message distributions with evolving causal–spurious message disentanglement via a recoverable gated network. Supported by theoretical analysis, LEMD achieves consistent OOD gains across six datasets, including +7.7% on future link prediction task and +3.0% average accuracy on node classification task.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62272338). Chen Zhao, Xintao Wu, and Zhong Chen did not receive any financial support for this work and contributed only by developing the research ideas, participating in discussions, and providing feedback on the manuscript.

Contribution Statement

Minglai Shao is the corresponding author.

References

[Arjovsky *et al.*, 2020] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, March 2020.

[Barros *et al.*, 2021] Claudio D. T. Barros, Matheus R. F. Mendonça, Alex B. Vieira, and Artur Ziviani. A survey on embedding dynamic graphs. *ACM Comput. Surv.*, 55(1):10:1–10:37, November 2021.

[Berger-Wolf and Saia, 2006] 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, New York, NY, USA, August 2006. Association for Computing Machinery.

[Cheng *et al.*, 2022] Dawei Cheng, Xiaoyang Wang, Ying Zhang, and Liqing Zhang. Graph neural network for fraud detection via spatial-temporal attention. *IEEE Transactions on Knowledge and Data Engineering*, 34(8):3800–3813, August 2022.

[Christiansen *et al.*, 2022] Rune Christiansen, Niklas Pfister, Martin Emil Jakobsen, Nicola Gnecco, and Jonas Peters. A causal framework for distribution generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6614–6630, October 2022.

[Creager *et al.*, 2021] Elliot Creager, Joern-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *Proceedings of the 38th International Conference on Machine Learning*, pages 2189–2200. PMLR, July 2021.

[Ganin *et al.*, 2016] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

[Gilmer *et al.*, 2017] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1263–1272. PMLR, July 2017.

[Hughes *et al.*, 2010] Michael E. Hughes, John Peeler, and John B. Hogenesch. Network dynamics to evaluate performance of an academic institution. *Science Translational Medicine*, 2(53):53ps49–53ps49, October 2010.

[Komanduri *et al.*, 2024] Aneesh Komanduri, Chen Zhao, Feng Chen, and Xintao Wu. Causal diffusion autoencoders: Toward counterfactual generation via diffusion probabilistic models. *arXiv preprint arXiv:2404.17735*, 2024.

[Krueger *et al.*, 2021] 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), February 2021.

[Kumar *et al.*, 2019] 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, New York, NY, USA, July 2019. Association for Computing Machinery.

[Li *et al.*, 2020] Guohao Li, Chenxin Xiong, Ali Thabet, and Bernard Ghanem. Deepergcn: All you need to train deeper gcns, June 2020.

[Li *et al.*, 2021] Xiaotong Li, Yongxing Dai, Yixiao Ge, Jun Liu, Ying Shan, and Lingyu Duan. Uncertainty modeling for out-of-distribution generalization. In *International Conference on Learning Representations*, October 2021.

[Lu *et al.*, 2024] Bin Lu, Ze Zhao, Xiaoying Gan, Shiyu Liang, Luoyi Fu, Xinbing Wang, and Chenghu Zhou. Graph out-of-distribution generalization with controllable data augmentation. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6317–6329, November 2024.

[Ma *et al.*, 2025] Xiaoxu Ma, Chen Zhao, Minglai Shao, and Yujie Lin. Hypergraph-based dynamic graph node classification. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[Ma *et al.*, 2026] Xiaoxu Ma, Dong Li, Minglai Shao, Xintao Wu, and Chen Zhao. Llm-enhanced energy contrastive learning for out-of-distribution detection in text-attributed graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 24308–24316, 2026.

[Pareja *et al.*, 2020] Aldo Pareja, Giacomo Domeniconi, Jie Chen, Tengfei Ma, Toyotaro Suzumura, Hiroki Kanezashi, Tim Kaler, Tao Schardl, and Charles Leiserson. Evolvegcn: Evolving graph convolutional networks for dynamic graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5363–5370, April 2020.

[Qin *et al.*, 2023] Xiao Qin, Nasrullah Sheikh, Chuan Lei, Berthold Reinwald, and Giacomo Domeniconi. Seign: A simple and efficient graph neural network for large dynamic graphs. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 2850–2863, April 2023.

[Rosenfeld *et al.*, 2021] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization, March 2021.

[Sagawa *et al.*, 2020] 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, April 2020.

[Sankar *et al.*, 2020] 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, New York, NY, USA, January 2020. Association for Computing Machinery.

[Seo *et al.*, 2018] Youngjoo Seo, Michaël Defferrard, Pierre Vandergheynst, and Xavier Bresson. Structured sequence modeling with graph convolutional recurrent networks. In Long Cheng, Andrew Chi Sing Leung, and Seiichi Ozawa, editors, *Neural Information Processing*, pages 362–373, Cham, 2018. Springer International Publishing.

[Sinha *et al.*, 2015] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th International Conference on World Wide Web*, pages 243–246, New York, NY, USA, May 2015. Association for Computing Machinery.

[Sun *et al.*, 2021] Xinwei Sun, Botong Wu, Xiangyu Zheng, Chang Liu, Wei Chen, Tao Qin, and Tie-Yan Liu. Recovering latent causal factor for generalization to distributional shifts. In *Advances in Neural Information Processing Systems*, volume 34, pages 16846–16859. Curran Associates, Inc., 2021.

[Sun *et al.*, 2025] Qingyun Sun, Jiayi Luo, Haonan Yuan, Xingcheng Fu, Hao Peng, Jianxin Li, and Philip S. Yu. Evolving graph learning for out-of-distribution generalization in non-stationary environments, November 2025.

[Tang *et al.*, 2008] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 990–998, New York, NY, USA, August 2008. Association for Computing Machinery.

[Tang *et al.*, 2012] 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, New York, NY, USA, August 2012. Association for Computing Machinery.

[Tian *et al.*, 2024] Qin Tian, Chen Zhao, Minglai Shao, Wenjun Wang, Yujie Lin, and Dong Li. Mldgg: Meta-learning for domain generalization on graphs, November 2024.

[Tian *et al.*, 2025] Qin Tian, Chen Zhao, Minglai Shao, Wenjun Wang, Yujie Lin, and Dong Li. Mldgg: Meta-learning for domain generalization on graphs. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 1361–1372, 2025.

[Tian *et al.*, 2026] Qin Tian, Chen Zhao, Xintao Wu, Dong Li, Minglai Shao, Xujiang Zhao, and Wenjun Wang. Class-domain incremental learning on graphs via disentangled knowledge distillation. In *Proceedings of the ACM Web Conference 2026*, pages 452–462, 2026.

[Vuong *et al.*, 2025] Long-Tung Vuong, Vy Vo, Hien Dang, Van-Anh Nguyen, Thanh-Toan Do, Mehrtash Harandi, Trung Le, and Dinh Phung. Why domain generalization fail? a view of necessity and sufficiency, February 2025.

[Wang *et al.*, 2025] Xin Wang, Haoyang Li, Zeyang Zhang, Haibo Chen, Tong Xiao, Kehan Li, and Wenwu Zhu. Uncertainty-aware disentangled dynamic graph attention network for out-of-distribution generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–18, 2025.

[Wu *et al.*, 2024] Qitian Wu, Hengrui Zhang, Junchi Yan, and David Wipf. Handling distribution shifts on graphs: An invariance perspective, August 2024.

[Ye *et al.*, 2025] Wenqian Ye, Luyang Jiang, Eric Xie, Guangtao Zheng, Yunsheng Ma, Xu Cao, Dongliang Guo, Daiqing Qi, Zeyu He, Yijun Tian, Megan Coffee, Zhe Zeng, Sheng Li, Ting-hao, Huang, Ziran Wang, James M. Rehg, Henry Kautz, and Aidong Zhang. The clever hans mirage: A comprehensive survey on spurious correlations in machine learning, October 2025.

[Yuan *et al.*, 2023] 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. *Advances in Neural Information Processing Systems*, 36:49715–49747, December 2023.

[Yuan *et al.*, 2024] Haonan Yuan, Qingyun Sun, Xingcheng Fu, Cheng Ji, and Jianxin Li. Dynamic graph information bottleneck. In *Proceedings of the ACM Web Conference 2024*, pages 469–480, New York, NY, USA, May 2024. Association for Computing Machinery.

[Zhang *et al.*, 2022] 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, December 2022.

[Zhang *et al.*, 2023] Zeyang Zhang, Xin Wang, Ziwei Zhang, Zhou Qin, Weigao Wen, Hui Xue', Haoyang Li, and Wenwu Zhu. Spectral invariant learning for dynamic graphs under distribution shifts. *Advances in Neural Information Processing Systems*, 36:6619–6633, December 2023.

[Zhang *et al.*, 2024] Guibin Zhang, Yiyan Qi, Ziyang Cheng, Yanwei Yue, Dawei Cheng, and Jian Guo. Rationalizing and augmenting dynamic graph neural networks. In *The Thirteenth International Conference on Learning Representations*, October 2024.