TRANSFER LEARNING FOR EDGE CLASSIFICATION ON DYNAMIC TEXT-ATTRIBUTED GRAPHS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning transferable representations for dynamic text-attributed graphs (DyTAGs) requires models to capture underlying interaction dynamics that persist across domains. However, existing methods tend to overfit to domain-specific structural, temporal, and semantic patterns, limiting edge classification performance under distribution shifts. To expose and address this, we formally establish a leave-one-domain-out (LODO) transfer learning protocol for edge classification on DyTAGs. Under this protocol, we demonstrate that state-of-the-art self-supervised methods for dynamic graph learning perform poorly when transferred to unseen domains. Strikingly, existing methods underperform a structurally and temporally unaware Bag of Events (BoE) model we introduce, which inputs only unordered sequences of node and edge text features. Proceeding from the BoE, we propose Spatio-Temporal Semantic Alignment (STSA), which integrates a spatio-temporal encoder that fuses representations of time deltas and node occurrence frequencies into a unified manifold. STSA is trained with a Contrastive Semantic Forecasting objective, which anchors edge representations to a multi-domain textual latent space initialized by a pretrained language model, providing a robust prior that outperforms BoE and all existing methods we evaluate.

1 INTRODUCTION

Dynamic text-attributed graphs (DyTAGs) serve as useful representations for complex systems where inherently linguistic or semantically annotated interactions between entities evolve over time, appearing frequently in domains such as e-commerce, temporal knowledge graphs, and communication networks. Standard frameworks model interactions through either discrete snapshot-based methods (Sankar et al., 2020; Pareja et al., 2020) or continuous-time methods that process data as a streaming sequence (Kumar et al., 2019; Trivedi et al., 2019; Rossi et al., 2020). Transformer architectures (Wang et al., 2021; Yu et al., 2023) advance this sequential paradigm, using self-attention over historical events to capture complex dependencies without relying on the restrictive compression of recurrent memory states. Such models were primarily designed for link prediction, a task where training labels are inherently abundant because they are derived from the topology of the graph. In contrast, the task of edge classification involves predicting the nature of an interaction, which frequently relies on scarce extrinsic annotations. Training highly parameterized models on limited data often causes them to overfit or collapse to majority class predictors. Self-supervised learning on unlabeled data becomes essential to train more robust models (Xu et al., 2023; Zhu et al., 2024; Liu et al., 2025; Gao et al., 2025; Bonnet & Rei, 2026). Capturing transferable interaction patterns across diverse domains improves generalizability further and allows developers to avoid the computational costs of pretraining a new model for every domain (Wang et al., 2023; Shamsi et al., 2025; Huang et al., 2025; Qiao et al., 2025; Chatterjee et al., 2025).

However, the field has yet to formally address cross-domain transfer learning for edge classification, as current frameworks rely on problem formulations limited to in-domain distributions. To address this, our **first contribution** is to formally establish a leave-one-domain-out (LODO) transfer learning protocol for edge classification on DyTAGs, which requires models to predict the labels of future interactions in an unseen target domain after self-supervised pretraining on a set of source domains.

To investigate model behavior under this protocol, our **second contribution** establishes a two-part diagnostic framework to expose critical vulnerabilities of existing approaches to modeling time and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

local structure. First, we conduct a representational analysis, demonstrating that when temporal and structural features are processed without non-linear fusion, their signals fail to properly integrate. Upon transfer to an unseen domain, this causes the representations to fragment, severely restricting the effective capacity of the model. Second, we introduce a Bag of Events (BoE) baseline. Pretrained solely on unordered sequences of node and edge text features, BoE serves as a control to determine whether structural or temporal inputs are sources of performance degradation. Strikingly, all existing methods we evaluate underperform BoE on average.

To address these vulnerabilities, our **third contribution** proposes Spatio-Temporal Semantic Alignment (STSA), which integrates a spatio-temporal encoder (STE) that fuses representations of time deltas and node frequencies into a unified manifold to resolve the fragmentation and expand representational capacity. We propose Contrastive Semantic Forecasting (CSF) to pretrain STSA. CSF allows more organizational freedom for representations than point-for-point reconstruction while anchoring the model to the multi-domain latent space of a pretrained language model (PLM).

Our **fourth contribution** lies in extensive empirical evaluations on the real-world datasets of the Dynamic Text-Attributed Graph Benchmark (DTGB) (Zhang et al., 2024). STSA consistently achieves statistically significant performance gains over BoE and all other existing methods across all DTGB domains. Furthermore, we isolate the quality of the pretrained representations of frozen encoders with a K-shot evaluation, demonstrating that STSA outperforms existing methods across all values of K. Finally, an ablation study confirms the necessity of STE and the superiority of CSF over InfoNCE (Oord et al., 2018) and Scaled Cosine Error (SCE) (Hou et al., 2022) losses.

2 Related Work

Research on pretraining generalizable dynamic graph models has largely proceeded along two parallel tracks: self-supervised learning within domains and transfer learning across domains. Importantly, frameworks in both tracks operate with pretraining steps that are decoupled from downstream evaluation. This distinguishes them from recent LLM-driven models for DyTAGs (Zhang et al., 2025a; Roy et al., 2025; Xu et al., 2026; Wang et al., 2026), which use joint-optimization paradigms that rely heavily on downstream task labels and cannot be easily isolated for standalone pretraining.

In the self-supervised track, pretraining strategies generally divide into generative reconstruction and contrastive alignment. Generative architectures formulate pretraining as a missing data recovery task by using masking strategies to reconstruct graph topology or temporal features (Liu et al., 2025; Gao et al., 2025). Conversely, contrastive methods optimize for mutual information by aligning varied structural or temporal views of the graph. This is often done by tracking topological changes or temporal neighborhood alignments (Xu et al., 2023; Zhu et al., 2024). More recently, models have used Transformer architectures with specialized embedding tables for directional role alignment (Bonnet & Rei, 2026). While these methods learn generalizable representations, they are typically evaluated within-domain and target intra-graph distribution shifts rather than bridging new domains.

To address domain shifts, the transfer learning track attempts to capture domain-agnostic interaction patterns. One line of research scales pretraining across multiple graphs to prevent encoders from memorizing graph-specific biases through training order shuffling or temporal normalization (Shamsi et al., 2025; Huang et al., 2025). Another relies on algorithmic adaptation to align distributions. This includes adversarial frameworks extracting domain-invariant features (Wang et al., 2023), continual learning strategies condensing synthetic memories (Qiao et al., 2025), and structural mapping approaches initializing target memory states from topological statistics (Chatterjee et al., 2025).

Although progress has been made in transfer learning for link prediction, node classification, and graph property prediction, none of the reviewed works address the significant challenges of transferring specifically for edge classification. STSA departs from the reviewed methods to tackle this directly. While other Transformer-based models exist, they introduce additional parameters for explicit role modeling, exacerbating overfitting (Bonnet & Rei, 2026), discard all node and edge attributes (Huang et al., 2025), or enforce domain invariance through adversarial alignment (Wang et al., 2023). Moreover, none of the methods fuse temporal and structural representations to mitigate overfitting to these modalities (STE) or align with a multi-domain PLM embedding space to improve transferability (CSF).

3 METHODOLOGY

This section details the formal setup of the leave-one-domain-out (LODO) transfer learning protocol for dynamic edge classification, our proposed Bag of Events (BoE) baseline, and our Spatio-Temporal Semantic Alignment (STSA) framework.

3.1 PRELIMINARIES

We represent a dynamic text-attributed graph as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{T})$, where $\mathcal{V}$ is the node set, $\mathcal{E}$ is the edge set, and $\mathcal{T}$ is the set of timestamps. Each interaction $\epsilon \in \mathcal{E}$ is a chronologically ordered tuple $\epsilon = (u, v, t)$ for $u, v \in \mathcal{V}$ and $t \in \mathcal{T}$. Nodes and interactions are associated with text attributes. For any given timestamp t, the temporal history of the graph is the set of all prior interactions $\mathcal{E}_{<t} = \{\epsilon \in \mathcal{E} \mid t_\epsilon < t\}$. Interactions $\mathcal{E}_{label} \subset \mathcal{E}$ are provided with a ground-truth label $y_\epsilon \in \mathcal{Y}$.

3.2 THE LEAVE-ONE-DOMAIN-OUT (LODO) PROTOCOL

To rigorously investigate transfer learning for edge classification on DyTAGs, we formally establish the leave-one-domain-out (LODO) protocol. Under this protocol, a domain can consist of one or more dynamic graphs. We define a set of source domains $\mathcal{D}$ for pretraining, which collectively contain a set of source graphs $\mathcal{S} = \{\mathcal{G}_1, \ldots, \mathcal{G}_K\}$. For evaluation, we hold out an entirely unseen target domain and evaluate on a constituent target graph $\mathcal{G}_{target}$. The learning process is divided into two phases. First, we learn an encoder h_θ over $\mathcal{S}$ without using edge labels. Second, we introduce a task-specific classification head g_ϕ for the dynamic edge classification task on $\mathcal{G}_{target}$. The model is then adapted using the labeled subset $\mathcal{E}_{label} \subset \mathcal{E}_{target}$, accommodating downstream paradigms such as end-to-end finetuning or frozen-encoder linear probing. For a new interaction between u and v arriving at time t in $\mathcal{G}_{target}$, the composite function may use any information about the query nodes or interaction time t, along with the available history $\mathcal{E}_{target<t}$, to predict the edge label $\hat{y}_\epsilon$. All information about the query edge itself must be masked from the input:

$$\hat{y}_\epsilon = g_\phi(h_\theta(u, v, t, \mathcal{E}_{target<t})). \tag{1}$$

3.3 BAG OF EVENTS (BOE)

The BoE model serves as a control for structural and temporal overfitting in existing models (illustrated in Appendix A). Conceptually, it treats the interaction history of a node as an unordered sequence of text-attributed events, disregarding topology and timestamps.

For a query interaction between nodes u and v at time t, we retrieve their respective historical one-hop interactions prior to t, truncated or padded to a fixed length k. Prepending u and v to their histories produces distinct event sequences S_u and S_v. Textual features are the sole input to BoE. Node and edge text embeddings are precomputed using Qwen3-Embedding-4B (Zhang et al., 2025b) and held static. For each event $i \in S_u$ (and symmetrically for S_v), let $\mathbf{n}_i \in \mathbb{R}^{d_{node}}$ and $\mathbf{e}_i \in \mathbb{R}^{d_{edge}}$ denote the node and edge text embeddings, respectively. To prevent target leakage, the text embedding of the query edge itself is masked. The embeddings are independently projected via weight matrices $\mathbf{W}_n$ and $\mathbf{W}_e$, and concatenated to form the event representation $\mathbf{x}_i = [\mathbf{W}_n\mathbf{n}_i \parallel \mathbf{W}_e\mathbf{e}_i] \in \mathbb{R}^{2d_c}$, where $\parallel$ denotes concatenation and $d_c = 96$. The sequence of these representations forms the input matrix $\mathbf{X}_u \in \mathbb{R}^{k \times 192}$ for S_u (and $\mathbf{X}_v$ for S_v), which is encoded by a multi-layer Transformer (Vaswani et al., 2017) using a phase-dependent processing strategy.

During self-supervised pretraining, $\mathbf{X}_u$ and $\mathbf{X}_v$ are processed independently using shared weights to prevent trivial cross-attention shortcuts. To produce single vector representations, we apply masked mean pooling over the sequence dimension (excluding padding tokens), followed by L_2-normalization to yield final representations $\mathbf{z}_u$ and $\mathbf{z}_v$. BoE optimizes a symmetric InfoNCE loss (Oord et al., 2018) with temperature $\tau = 0.07$ across a minibatch to align $\mathbf{z}_u$ and $\mathbf{z}_v$ (formulation provided in Appendix B).

Conversely, during downstream finetuning for edge classification, the model must capture complex dependencies between the interacting neighborhoods. The sequences are therefore concatenated into a joint input $\mathbf{X}_{uv} = [\mathbf{X}_u \parallel \mathbf{X}_v]$ and processed jointly. After self-attention, the output sequence is

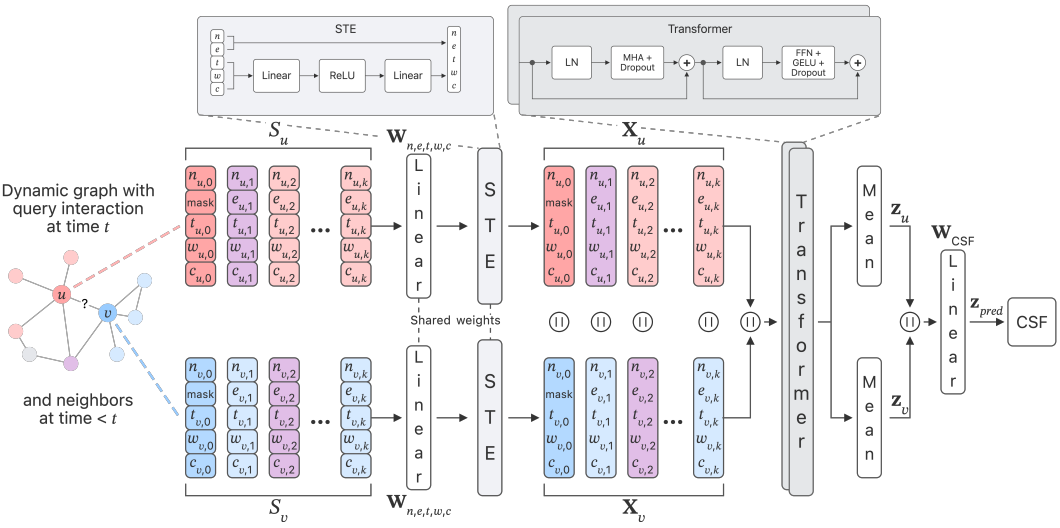


Figure 1: **Overview of the Spatio-Temporal Semantic Alignment (STSA) Framework. (Left)** For a query interaction between nodes u and v at time t, historical one-hop neighbors are retrieved. u and v combined with their respective neighbors form node sequences S_u and S_v. For each event i in these sequences, its node text embedding $\mathbf{n}_i$, edge text embedding $\mathbf{e}_i$, interaction time delta $\mathbf{t}_i$, within-sequence node frequency embedding $\mathbf{w}_i$, and cross-sequence node frequency embedding $\mathbf{c}_i$ are extracted, and the edge text embedding of the query nodes is masked. **(Center)** All five extracted features $\mathbf{n}_i$, $\mathbf{e}_i$, $\mathbf{t}_i$, $\mathbf{w}_i$, $\mathbf{c}_i$ are first independently projected via separate linear layers $\mathbf{W}_{n,e,t,w,c}$. The spatio-temporal encoder (STE) then fuses the projected temporal and structural features $\mathbf{t}_i$, $\mathbf{w}_i$, and $\mathbf{c}_i$ into a unified manifold. This fused spatio-temporal representation is concatenated with the projected node and edge text embeddings to form a single feature vector representing each event in S_u and S_v. **(Right)** These event sequences, corresponding to input matrices $\mathbf{X}_u$ and $\mathbf{X}_v$, are concatenated on the sequence dimension and jointly processed by a multi-layer Transformer encoder. After the output sequence is split back into distinct representations for S_u and S_v, mean pooling, with a mask to exclude padding tokens, is applied to form the final node representations $\mathbf{z}_u$ and $\mathbf{z}_v$. These representations are passed through a linear task head $\mathbf{W}_{\text{CSF}}$ to yield $\mathbf{z}_{pred}$, which is optimized via Contrastive Semantic Forecasting (CSF) to predict the masked query edge, anchoring the model to the multi-domain textual latent space.

split back into updated representations for S_u and S_v. Following masked mean pooling, the resulting $\mathbf{z}_u$ and $\mathbf{z}_v$ are concatenated and passed to a linear classification head.

3.4 Spatio-Temporal Semantic Alignment (STSA)

In this section, we introduce STSA (illustrated in Figure 1), a novel framework targeted at resolving structural, temporal, and semantic overfitting during transfer by integrating a spatio-temporal encoder (STE) and Contrastive Semantic Forecasting (CSF) into the architectural backbone of BoE.

3.4.1 Spatio-Temporal Encoder (STE)

A fundamental challenge in dynamic graph learning is the modality gap between structure and time. As we demonstrate with a representational analysis in Section 4.1, prior to Transformer encoding, these signals are severely fragmented and have restricted effective capacity under cross-domain transfer. To prevent this, STSA non-linearly fuses structural and temporal representations with an STE before sequence encoding.

For each event $i \in S_u$ (and symmetrically for S_v), let Δt be the time elapsed since the query interaction. We encode Δt via a Fourier time encoder (Xu et al., 2020) to produce $\mathbf{t}_i \in \mathbb{R}^{32}$. We map the occurrence count of the node within its own sequence and within the sequence of the paired node to embeddings $\mathbf{w}_i, \mathbf{c}_i \in \mathbb{R}^{32}$ (Bonnet & Rei, 2026), capturing local structure.

After projection via separate linear layers, the three 32-dimensional features representing Δt, within-sequence frequency, and cross-sequence frequency are concatenated to form a single 96-dimensional vector for each event. These vectors are processed by the STE, which is an MLP with a 256-dimensional hidden layer and a ReLU activation, to project the raw features into a unified 96-dimensional spatio-temporal latent space. This spatio-temporal representation is concatenated with the 192-dimensional semantic feature (detailed in Section 3.3) to form a 288-dimensional representation for each event in S_u and S_v, yielding $\mathbf{X}_u$ and $\mathbf{X}_v$.

3.4.2 CONTRASTIVE SEMANTIC FORECASTING (CSF)

For STSA pretraining, we propose CSF, an objective designed to predict the semantics of the interaction between u and v at time t. $\mathbf{X}_u$ and $\mathbf{X}_v$ are concatenated and processed jointly by the Transformer encoder. By forecasting the masked query edge rather than contrasting the nodes, CSF allows STSA to safely use joint sequence processing without target leakage. Following joint self-attention, the output sequence is split back into the distinct, updated representations for S_u and S_v. Finally, to produce single vector representations for u and v, we apply mean pooling over the sequence dimension excluding padding tokens from the calculation to yield the final representations $\mathbf{z}_u$ and $\mathbf{z}_v$.

To generate the predicted semantic representation $\mathbf{z}_{pred}$, the final node representations $\mathbf{z}_u$ and $\mathbf{z}_v$ from the Transformer are concatenated and fed through a dedicated CSF linear task head, which projects the joint state to 256 dimensions via $\mathbf{z}_{pred} = \mathbf{W}_{\text{CSF}}[\mathbf{z}_u \parallel \mathbf{z}_v] + \mathbf{b}_{\text{CSF}}$, where $\mathbf{W}_{\text{CSF}}$ and $\mathbf{b}_{\text{CSF}}$ are learnable parameters. For the ground-truth target $\mathbf{z}_{target}$, we use precomputed Qwen3 edge text embeddings of the interactions truncated to 256 dimensions. Because Qwen3-Embedding supports matryoshka representation learning (Kusupati et al., 2022), it enables this truncation while retaining the vast majority of the most important semantic information.

The model is optimized using an InfoNCE approach to align the predicted and target representations. For a minibatch of size N, the loss is formulated as:

$$\mathcal{L}_{\text{CSF}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\frac{\mathbf{z}_{pred}^{(i)} \cdot \mathbf{z}_{target}^{(i)}}{\tau}\right)}{\sum_{j=1}^{N} \exp\left(\frac{\mathbf{z}_{pred}^{(i)} \cdot \mathbf{z}_{target}^{(j)}}{\tau}\right)}, \tag{2}$$

where τ is a temperature parameter set to 0.07, following standard practice in contrastive learning.

Instead of forcing point-for-point reconstruction with a mean squared error or SCE (Hou et al., 2022) loss, CSF allows more organizational freedom for the latent space by having the model produce a representation more similar to the true edge text embedding than to all in-batch negatives. Consequently, this objective incentivizes the model to learn the semantic trajectory of the interaction dynamics of the graph.

4 EXPERIMENTS

In this section, we first conduct a representational analysis to demonstrate the mechanistic basis of the spatio-temporal encoder (STE), revealing how it non-linearly fuses structural and temporal signals to mitigate representational fragmentation and expand effective capacity under cross-domain transfer. We then evaluate BoE and STSA alongside existing self-supervised and transfer learning baselines on dynamic edge classification under the LODO protocol. We report performance under both end-to-end finetuning and a parametric K-shot linear probe that assesses the inherent quality of the transferred representations. Finally, we conduct an ablation study to isolate the contributions of STE and CSF.

4.1 SPATIO-TEMPORAL REPRESENTATIONAL ANALYSIS

To demonstrate the mechanistic necessity of fusing temporal and structural features via STE, we conduct a representational analysis (Figure 2). We isolate these modalities by generating text-free dynamic graphs across four archetypal topologies (detailed in Appendix C).

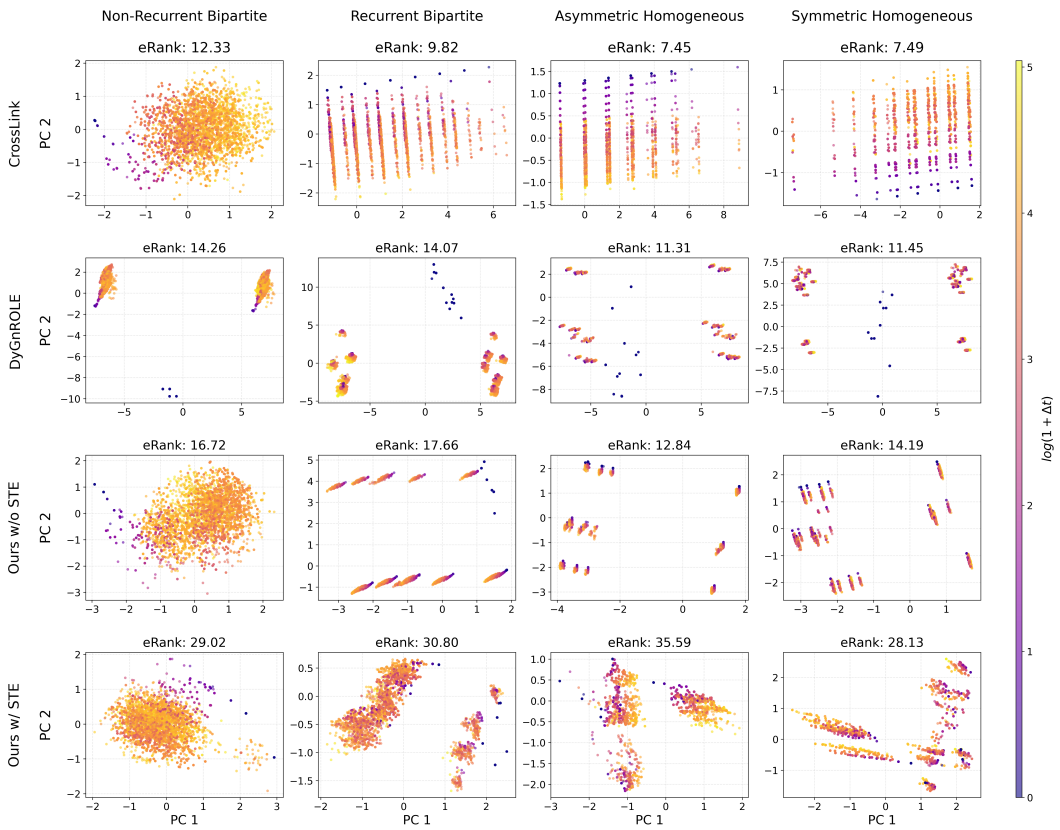


Figure 2: **Analysis of Spatio-Temporal Representations.** Principal Component Analysis (PCA) of spatio-temporal representations colored by $\log(1 + \Delta t)$ across four synthetic graph topologies. **Top 3**: Under domain transfer, concatenated time and node frequency features suffer from representational fragmentation and a restricted Effective Rank (eRank). **Bottom**: STE fuses the temporal and structural features into a unified manifold and substantially improves eRank.

Using the BoE/STSA backbone with independent Transformer processing (Section 3.3), we train two text-free variants (with and without STE) with InfoNCE under the LODO protocol. We also train two event-based Transformer architectures, DyGnROLE and CrossLink, on their native objectives. After validation early stopping with a patience of 10, we extract target-domain features immediately prior to Transformer encoding and evaluate the manifold geometry via Principal Component Analysis (PCA) and dimensional health via Effective Rank (eRank) (Roy & Vetterli, 2007).

As shown in Figure 2 (top 3), the time and node frequency features of DyGnROLE, CrossLink, and Ours w/o STE suffer severe representational fragmentation and a restricted eRank of 7.45–17.66 out of 96. Without fusion, a Transformer must expend self-attention capacity bridging these gaps. STE resolves this via non-linear fusion into a more expressive manifold (bottom) that approximately doubles eRank on average. The clusters are artifacts of the within- and cross-sequence frequency representations of the models. We include another version of this figure coloring points by count combinations instead of $\log(1 + \Delta t)$ in Appendix D. The separation of DyGnROLE representations for the Non-Recurrent Bipartite domain is a result of its additive role-semantic positional encodings, allowing the model to differentiate source-sequence nodes from destination-sequence nodes.

The Non-Recurrent Bipartite domain highlights a particular behavior of STE. In topologies with very little recurrence, within-sequence and cross-sequence node frequencies are almost always 1 and 0 (98.52% of nodes in our generated graph), respectively, meaning the spatio-temporal signal is largely temporal. Although PCA visualizes the representations of most models in this domain as visually similar linear gradients, the eRank with STE confirms a large capacity expansion from a baseline maximum of 16.72 to 29.02.

Table 1: **Dynamic Edge Classification Results.** Values report average Macro F1 (%) ± standard deviation over 10 runs. The best performance in each column is shown in **bold**. Statistical significance of the best-in-column result over the strongest baseline is indicated by $^{*}p < 0.10$, $^{**}p < 0.05$, and $^{***}p < 0.01$ under a two-tailed paired t-test for individual datasets and Wilcoxon signed-rank test for the average. In each method block, the superior variant for each dataset is underlined, comparing models with pretraining (*W/ PT*) and without pretraining (*W/O PT*).

Method	Amazon	Enron	GDELT	Google	ICEWS	Stack E	Stack U	Yelp	Avg
Majority	$15.76_{\pm 0.00}$	$4.06_{\pm 0.00}$	$0.06_{\pm 0.00}$	$15.93_{\pm 0.00}$	$0.12_{\pm 0.00}$	$42.82_{\pm 0.00}$	$44.65_{\pm 0.00}$	$13.99_{\pm 0.00}$	$17.17_{\pm 0.00}$
Random	$15.26_{\pm 0.04}$	$7.00_{\pm 0.06}$	$0.16_{\pm 0.01}$	$15.02_{\pm 0.06}$	$0.13_{\pm 0.01}$	$46.72_{\pm 0.09}$	$44.81_{\pm 0.10}$	$16.95_{\pm 0.03}$	$18.26_{\pm 0.05}$
GCAL									
W/O PT	$15.76_{\pm 0.00}$	$4.20_{\pm 1.60}$	$0.04_{\pm 0.03}$	$15.93_{\pm 0.00}$	$0.09_{\pm 0.04}$	$42.82_{\pm 0.00}$	$44.65_{\pm 0.00}$	$13.99_{\pm 0.00}$	$17.19_{\pm 0.21}$
W/ PT	$15.76_{\pm 0.00}$	$3.59_{\pm 2.61}$	$0.05_{\pm 0.02}$	$15.93_{\pm 0.00}$	$0.10_{\pm 0.05}$	$42.82_{\pm 0.00}$	$44.65_{\pm 0.00}$	$13.99_{\pm 0.00}$	$17.11_{\pm 0.34}$
IDOL									
W/O PT	$15.76_{\pm 0.00}$	$4.06_{\pm 0.01}$	$0.12_{\pm 0.03}$	$15.93_{\pm 0.00}$	$0.13_{\pm 0.04}$	$42.82_{\pm 0.00}$	$44.65_{\pm 0.00}$	$13.99_{\pm 0.00}$	$17.18_{\pm 0.01}$
W/ PT	$15.76_{\pm 0.00}$	$4.07_{\pm 0.02}$	$0.12_{\pm 0.03}$	$15.93_{\pm 0.00}$	$0.13_{\pm 0.02}$	$42.82_{\pm 0.00}$	$44.65_{\pm 0.00}$	$13.99_{\pm 0.00}$	$17.18_{\pm 0.01}$
DyGMAE									
W/O PT	$15.76_{\pm 0.00}$	$4.34_{\pm 0.88}$	$0.05_{\pm 0.03}$	$15.93_{\pm 0.00}$	$0.08_{\pm 0.04}$	$42.82_{\pm 0.00}$	$44.65_{\pm 0.00}$	$13.99_{\pm 0.00}$	$17.20_{\pm 0.12}$
W/ PT	$15.76_{\pm 0.00}$	$4.95_{\pm 1.75}$	$0.07_{\pm 0.03}$	$15.93_{\pm 0.00}$	$0.09_{\pm 0.05}$	$42.82_{\pm 0.00}$	$44.65_{\pm 0.00}$	$13.99_{\pm 0.00}$	$17.28_{\pm 0.23}$
TLP									
W/O PT	$15.76_{\pm 0.00}$	$4.10_{\pm 1.26}$	$0.04_{\pm 0.02}$	$15.93_{\pm 0.00}$	$0.09_{\pm 0.04}$	$42.82_{\pm 0.00}$	$44.65_{\pm 0.00}$	$13.99_{\pm 0.00}$	$17.17_{\pm 0.17}$
W/ PT	$15.76_{\pm 0.00}$	$5.37_{\pm 2.12}$	$0.06_{\pm 0.02}$	$15.94_{\pm 0.00}$	$0.08_{\pm 0.04}$	$42.82_{\pm 0.00}$	$44.65_{\pm 0.00}$	$13.99_{\pm 0.00}$	$17.33_{\pm 0.27}$
DVGMAE									
W/O PT	$11.91_{\pm 5.93}$	$3.84_{\pm 2.46}$	$0.02_{\pm 0.02}$	$11.45_{\pm 6.16}$	$0.02_{\pm 0.03}$	$38.45_{\pm 9.22}$	$39.35_{\pm 11.20}$	$9.37_{\pm 5.31}$	$14.30_{\pm 5.04}$
W/ PT	$15.76_{\pm 0.00}$	$5.66_{\pm 2.07}$	$0.05_{\pm 0.03}$	$15.93_{\pm 0.00}$	$0.07_{\pm 0.05}$	$42.82_{\pm 0.00}$	$44.65_{\pm 0.00}$	$13.99_{\pm 0.00}$	$17.37_{\pm 0.27}$
CLDG									
W/O PT	$15.76_{\pm 0.00}$	$5.64_{\pm 1.71}$	$0.04_{\pm 0.02}$	$15.93_{\pm 0.00}$	$0.05_{\pm 0.03}$	$42.82_{\pm 0.00}$	$44.65_{\pm 0.00}$	$13.99_{\pm 0.00}$	$17.36_{\pm 0.22}$
W/ PT	$15.76_{\pm 0.01}$	$5.94_{\pm 2.08}$	$0.06_{\pm 0.04}$	$15.94_{\pm 0.01}$	$0.07_{\pm 0.03}$	$42.82_{\pm 0.00}$	$44.65_{\pm 0.00}$	$13.99_{\pm 0.00}$	$17.40_{\pm 0.27}$
EvoluNet									
W/O PT	$13.44_{\pm 5.19}$	$4.19_{\pm 2.03}$	$0.02_{\pm 0.02}$	$12.78_{\pm 5.84}$	$0.02_{\pm 0.02}$	$38.29_{\pm 9.54}$	$38.96_{\pm 11.99}$	$12.57_{\pm 4.38}$	$15.03_{\pm 4.88}$
W/ PT	$16.15_{\pm 0.46}$	$5.18_{\pm 1.98}$	$0.04_{\pm 0.03}$	$16.06_{\pm 0.38}$	$0.06_{\pm 0.04}$	$44.41_{\pm 2.15}$	$45.58_{\pm 1.43}$	$14.61_{\pm 0.56}$	$17.76_{\pm 0.88}$
CrossLink									
W/O PT	$15.76_{\pm 0.00}$	$5.27_{\pm 0.68}$	$0.08_{\pm 0.04}$	$15.93_{\pm 0.00}$	$0.12_{\pm 0.00}$	$42.82_{\pm 0.00}$	$44.65_{\pm 0.00}$	$17.28_{\pm 2.24}$	$17.74_{\pm 0.37}$
W/ PT	$15.76_{\pm 0.00}$	$6.65_{\pm 1.11}$	$0.09_{\pm 0.04}$	$15.93_{\pm 0.00}$	$0.12_{\pm 0.03}$	$42.97_{\pm 0.17}$	$49.19_{\pm 0.85}$	$18.47_{\pm 1.43}$	$18.65_{\pm 0.46}$
DyGnROLE									
W/O PT	$15.76_{\pm 0.00}$	$4.14_{\pm 0.14}$	$0.07_{\pm 0.01}$	$15.93_{\pm 0.00}$	$0.12_{\pm 0.00}$	$42.82_{\pm 0.00}$	$44.65_{\pm 0.00}$	$13.99_{\pm 0.00}$	$17.18_{\pm 0.02}$
W/ PT	$17.06_{\pm 0.30}$	$8.67_{\pm 0.54}$	$0.27_{\pm 0.02}$	$16.13_{\pm 0.24}$	$0.43_{\pm 0.08}$	$47.56_{\pm 3.13}$	$46.54_{\pm 0.97}$	$16.94_{\pm 1.40}$	$19.20_{\pm 0.83}$
BoE									
W/O PT	$17.44_{\pm 1.55}$	$7.78_{\pm 0.18}$	$0.15_{\pm 0.05}$	$15.95_{\pm 0.05}$	$0.16_{\pm 0.05}$	$45.79_{\pm 3.84}$	$45.28_{\pm 0.91}$	$16.80_{\pm 4.00}$	$18.67_{\pm 1.33}$
W/ PT	$24.61_{\pm 0.86}$	$7.71_{\pm 0.38}$	$0.28_{\pm 0.03}$	$18.11_{\pm 1.16}$	$0.54_{\pm 0.11}$	$54.22_{\pm 1.34}$	$51.61_{\pm 1.27}$	$26.34_{\pm 2.34}$	$22.93_{\pm 0.94}$
STSA									
W/O PT	$16.68_{\pm 1.75}$	$7.95_{\pm 0.31}$	$0.17_{\pm 0.02}$	$16.51_{\pm 0.67}$	$0.21_{\pm 0.03}$	$46.07_{\pm 2.54}$	$45.44_{\pm 1.14}$	$19.77_{\pm 3.21}$	$19.10_{\pm 1.21}$
W/ PT	$\mathbf{25.22}_{\pm 0.40}{}^{*}$	$\mathbf{9.12}_{\pm 0.23}{}^{**}$	$\mathbf{0.46}_{\pm 0.02}{}^{***}$	$\mathbf{20.40}_{\pm 0.72}{}^{***}$	$\mathbf{1.24}_{\pm 0.08}{}^{***}$	$\mathbf{55.52}_{\pm 0.55}{}^{**}$	$\mathbf{52.84}_{\pm 0.76}{}^{**}$	$\mathbf{28.96}_{\pm 0.75}{}^{***}$	$\mathbf{24.22}_{\pm 0.44}{}^{***}$

4.2 Benchmark Evaluation Setup

Datasets. We evaluate on eight datasets from DTGB (Zhang et al., 2024), spanning four diverse domains: e-commerce reviews (Amazon, Google, Yelp), corporate communication (Enron), technical Q&A forums (Stack E, Stack U), and political events (GDELT, ICEWS). These domains exhibit high topological, temporal, and linguistic variance, alongside widely varied dynamic edge classification tasks, making them an excellent testbed for transfer learning. Comprehensive graph statistics, label distributions, and text attribute descriptions are provided in Appendix E.

LODO on DTGB. To apply the LODO protocol on DTGB, each model undergoes four separate pretraining runs, each excluding the datasets of one domain. For example, the model evaluated on the e-commerce domain is pretrained only on datasets from corporate communication (Enron), technical Q&A forums (Stack E and Stack U), and political events (GDELT and ICEWS), excluding Amazon, Google, and Yelp. This ensures no data from the target domain is observed during pretraining.

Pretraining and Finetuning Procedure. We implement a chronological interleaving strategy that alternates batches from source datasets in a round-robin sequence (details in Appendix F). We follow DTGB in partitioning graphs into chronological 70-15-15 training, validation, and test splits. We cap the total pretraining data at 1M training edges and 150k validation edges distributed evenly across the source graphs. Pretraining is performed until convergence, with early stopping based on validation performance using a patience of 10 epochs. During finetuning, the model is adapted using the latest 500 labeled edges from the training split for a fixed 20 epochs.

Baselines and Implementation. We compare BoE and STSA alongside diverse state-of-the-art methods. Our selection includes recent self-supervised methods by Liu et al. (2025), Gao et al. (2025), Xu et al. (2023), Zhu et al. (2024), and Bonnet & Rei (2026), as well as transfer learn-

ing methods by Wang et al. (2023), Huang et al. (2025), Qiao et al. (2025), and Chatterjee et al. (2025). Implementation details, source code links (where available), and hyperparameter settings are provided in Appendix G. Hardware specifications are provided in Appendix H.

4.3 Main Results

State-of-the-Art Comparison. Table 1 reveals that existing models struggle to transfer effectively to unseen domains. STSA demonstrates superior performance, consistently achieving substantial performance gains. It also yields the highest average positive transfer (5.12%), defined as the performance difference between models with and without pretraining. While nearly all models achieve some positive transfer on average, over half collapse to majority class predictors or degrade below random chance on multiple datasets. This widespread degradation suggests that existing mechanisms (e.g., continual memory replay in GCAL, node-level temporal consistency in IDOL) anchor too rigidly to source-domain topologies and cause interference when transferring to target domains.

Vulnerabilities of Existing Methods. Perhaps most strikingly, the minimalist BoE baseline outperforms all existing state-of-the-art methods. Despite being structurally and temporally unaware, operating solely on unordered sequences of historical text features, it achieves an average Macro F1 score of 22.93%, surpassing the next best model, DyGnROLE, by 3.73 points. This exposes a critical vulnerability in current methods: they overfit to the domain-specific structural, temporal, and linguistic patterns of their pretraining data, and consequently their learned representations fail to provide transferable discriminative power when moved to disjoint domain distributions.

Impact of Label Scarcity. Many DTGB datasets suffer from severe class imbalance (see Figure 6 in Appendix E). We observed a strong, statistically significant positive correlation ($r_s = +0.905, p = 0.002$) between training split class imbalance (Gini coefficient) and the relative performance gain of STSA over the BoE baseline, demonstrating that the more a dataset relies on transfer learning due to label scarcity, the greater the performance improvement. In this setting, STSA outperforms BoE by relative margins of 64.29% on GDELT and 129.63% on ICEWS. The low absolute scores on GDELT and ICEWS occur because 147 and 169 classes, respectively, are entirely absent from the finetuning split, severely diluting the unweighted Macro F1. Evaluating solely on observed classes significantly expands the absolute margins of STSA over all baselines (Appendix I).

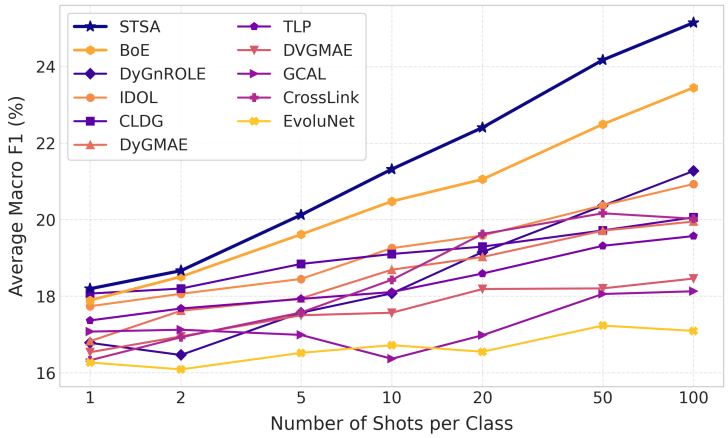


Figure 3: **Parametric K-Shot Linear Probe Performance.** Average Macro F1 across all target datasets under the leave-one-domain-out (LODO) protocol using a frozen pretrained backbone and a Logistic Regression classifier fitted on K shots per class.

4.4 Linear Probing of Frozen Representations

While our main results evaluate models under a finetuning paradigm, end-to-end adaptation can obscure the inherent quality of the pretrained representations. To rigorously evaluate the linear separability of the transferred representations without the confounding factor of backbone adaptation, we conduct a parametric K-shot linear probing experiment.

Table 2: **Ablation study of STSA finetuning for dynamic edge classification.** Values report average Macro F1 (%) ± standard deviation for individual datasets over 10 runs with different random seeds. Best performance in each column is shown in **bold**.

Variant	Amazon	Enron	GDELT	Google	ICEWS	Stack E	Stack U	Yelp	Avg
Reference Baseline									
BoE	$24.61_{\pm 0.86}$	$7.71_{\pm 0.38}$	$0.28_{\pm 0.03}$	$18.11_{\pm 1.16}$	$0.54_{\pm 0.11}$	$54.22_{\pm 1.34}$	$51.61_{\pm 1.27}$	$26.34_{\pm 2.34}$	$22.93_{\pm 0.94}$
STSA Variants									
w/o STE	$24.14_{\pm 0.54}$	$8.18_{\pm 0.38}$	$0.31_{\pm 0.02}$	$18.07_{\pm 0.56}$	$0.70_{\pm 0.09}$	$51.76_{\pm 2.69}$	$51.16_{\pm 1.72}$	$28.16_{\pm 0.53}$	$22.81_{\pm 0.82}$
CSF → SF	$23.79_{\pm 0.47}$	$8.34_{\pm 0.42}$	$0.41_{\pm 0.02}$	$20.23_{\pm 0.72}$	$1.12_{\pm 0.06}$	$54.38_{\pm 2.29}$	$52.71_{\pm 0.78}$	$28.16_{\pm 0.38}$	$23.64_{\pm 0.64}$
CSF → InfoNCE	$21.72_{\pm 0.62}$	$\mathbf{10.53}_{\pm 1.21}$	$0.37_{\pm 0.01}$	$18.47_{\pm 0.79}$	$0.95_{\pm 0.06}$	$54.08_{\pm 1.22}$	$50.88_{\pm 0.94}$	$26.97_{\pm 0.54}$	$22.99_{\pm 0.68}$
Complete STSA	$\mathbf{25.22}_{\pm 0.40}$	$9.12_{\pm 0.23}$	$\mathbf{0.46}_{\pm 0.02}$	$\mathbf{20.40}_{\pm 0.72}$	$\mathbf{1.24}_{\pm 0.08}$	$\mathbf{55.52}_{\pm 0.55}$	$\mathbf{52.84}_{\pm 0.76}$	$\mathbf{28.96}_{\pm 0.75}$	$\mathbf{24.22}_{\pm 0.44}$

In this setup, we freeze the pretrained encoders, and for each target graph, we extract node representations and fit a simple ℓ_2-regularized multinomial Logistic Regression classifier. To mitigate the effects of class imbalance and evaluate label efficiency, we randomly sample K instances per class from the target training split to train the classifier, where $K \in \{1, 2, 5, 10, 20, 50, 100\}$. The model is then evaluated on the full test split.

The results, illustrated in Figure 3, reinforce the findings of our finetuning evaluation while exposing weaknesses in existing methods. STSA consistently achieves superior performance across K. As the number of examples increases, the performance gap between STSA and the other models widens.

4.5 Ablation Study

In this section, we evaluate the contribution of the STE module and CSF pretraining objective in our STSA baseline model (Table 2).

STE. Bypassing the STE and feeding concatenated structural and temporal features directly into the Transformer (w/o STE) reduces the average Macro F1 from 24.22 to 22.81. This confirms the findings of our representational analysis: non-linear fusion via the STE is necessary to mitigate representational fragmentation and prevent a drop in eRank when transferring to unseen domains.

CSF. Swapping the CSF objective for a standard InfoNCE loss (CSF → InfoNCE) results in a drop in average Macro F1 from 24.22 to 22.99, exposing the limitations of purely structural objectives that ignore query edge semantics. However, a notable exception occurs on the Enron dataset, where InfoNCE outperforms CSF. This suggests that in networks defined by organizational cliques, the structural identity of the dyad can act as a stronger predictor of the interaction topic than the semantic trajectory of past interactions. Crucially, this structural alignment only succeeds in combination with the STE. InfoNCE without the STE (the BoE baseline) falls to 7.71. Finally, replacing CSF with a direct semantic forecasting objective using an SCE (Hou et al., 2022) loss (CSF → SF) degrades average performance to 23.64, demonstrating that contrastive alignment yields a more robust transfer prior than point-by-point vector reconstruction. We provide the formulation of SCE in Appendix B.

5 Conclusion

We formally established a LODO protocol to investigate transfer learning for edge classification on DyTAGs. Our evaluations exposed a critical vulnerability in current dynamic graph methods: they overfit to the structural, temporal, and semantic distributions of their pretraining domains. Consequently, complex models fail to generalize to unseen networks, consistently underperforming our structurally and temporally unaware BoE baseline. To address this, we proposed STSA, a framework that uses a spatio-temporal encoder to non-linearly fuse temporal and structural representations, mitigating fragmentation and expanding effective capacity. This is paired with a CSF objective that anchors edge representations to a multi-domain PLM latent space. STSA demonstrates superior performance and linear separability across all evaluated DTGB domains. Ultimately, this work underscores the importance of evaluating models under severe distribution shifts and provides a robust foundation for learning generalizable interaction patterns.

AI USE STATEMENT

Generative AI was used to improve text readability.

ETHICS STATEMENT

This work focuses on foundational methodology for dynamic graph forecasting. However, predicting future interactions in networks, such as political events or corporate communications, presents dual-use risks. In downstream deployments, such predictive models could be used for unwarranted surveillance of individuals or groups. Mitigating these risks in applied settings requires careful consideration of the deployment context and the implementation of appropriate access controls when using human-centric datasets.

REPRODUCIBILITY STATEMENT

We detail the leave-one-domain-out protocol in Section 4.2 and provide all hyperparameters in Appendix G. The datasets are accessible via the DTGB repository (`https://github.com/zjs123/DTGB`), and code will be made public upon publication. Detailed architectural descriptions and hyperparameter settings in the main text and appendices support reproducibility in the interim.

REFERENCES

Tyler Bonnet and Marek Rei. Dygnrole: Modeling asymmetry in dynamic graphs with node-role-oriented latent encoding. *arXiv preprint arXiv:2602.23135*, 2026.

Ayan Chatterjee, Barbara Ikica, Babak Ravandi, and John Palowitch. Transfer learning for temporal link prediction. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2025.

Mengzhou Gao, Xinxun Zhang, Pengfei Jiao, Tianpeng Li, and Zhidong Zhao. Dvgmae: Self-supervised dynamic variational graph masked autoencoder. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. Graphmae: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 594–604, 2022.

Xuanwen Huang, Wei Chow, Yize Zhu, Yang Wang, Ziwei Chai, Chunping Wang, Lei Chen, and Yang Yang. Enhancing cross-domain link prediction via evolution process modeling. In *Proceedings of the ACM on Web Conference 2025*, pp. 2158–2171, 2025.

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*, pp. 1269–1278, Anchorage, AK, USA, July 2019. ACM.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249, 2022.

Weixiong Liu, Junwei Cheng, Zhongyu Pan, Chaobo He, and Quanlong Guan. Dygmae: A novel dynamic graph masked autoencoder for link prediction. In *The 41st Conference on Uncertainty in Artificial Intelligence*, 2025.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Aldo Pareja, Giacomo Domeniconi, Jie Chen, Tengfei Ma, Toyotaro Suzumura, Hiroki Kanezashi, Tim Kaler, Tao B. Schardl, and Charles E. Leiserson. Evolvegcn: Evolving graph convolutional networks for dynamic graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5363–5370. AAAI Press, 2020.

Ziyue Qiao, Qianyi Cai, Hao Dong, Jiawei Gu, Pengyang Wang, Meng Xiao, Xiao Luo, and Hui Xiong. GCAL: Adapting graph models to evolving domain shifts. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 50113–50129. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/qiao25a.html.

Emanuele Rossi, Benjamin P. Chambers, Fabrizio Frasca, and et al. Temporal graph networks for deep learning on dynamic graphs. *arXiv preprint arXiv:2006.10637*, 2020.

Amit Roy, Ning Yan, and Masood Mortazavi. Llm-driven knowledge distillation for dynamic text-attributed graphs. *arXiv preprint arXiv:2502.10914*, 2025.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pp. 606–610. IEEE, 2007.

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 (WSDM)*, pp. 519–527. ACM, 2020.

Kiarash Shamsi, Tran Gia Bao Ngo, Razieh Shirzadkhani, Shenyang Huang, Farimah Poursafaei, Poupak Azad, Reihaneh Rabbany, Baris Coskunuzer, Guillaume Rabusseau, and Cuneyt Gurcan Akcora. Mint: Multi-network transfer benchmark for temporal graph learning. In *Thirty-Ninth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Rakshit Trivedi, Mehrdad Farajtabar, Prasenjeet Biswal, and Hongyuan Zha. Dyrep: Learning representations over dynamic graphs. In *International Conference on Learning Representations (ICLR)*, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Haohui Wang, Yuzhen Mao, Jianhui Sun, Si Zhang, and Dawei Zhou. Dynamic transfer learning across graphs. *CoRR*, abs/2305.00664, 2023. doi: 10.48550/ARXIV.2305.00664. URL https://doi.org/10.48550/arXiv.2305.00664.

Lu Wang, Xiaofu Chang, Shuang Li, Yunfei Chu, Hui Li, Wei Zhang, Xiaofeng He, Le Song, Jingren Zhou, and Hongxia Yang. Tcl: Transformer-based dynamic graph modelling via contrastive learning. *arXiv preprint arXiv:2105.07944*, 2021.

Yunan Wang, Jianxin Li, and Ziwei Zhang. Global-recent semantic reasoning on dynamic text-attributed graphs with large language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=puocvrFZRl.

Da Xu, Chao Ruan, Evren Korpeoglu, Srijan Kumar, and Kannan Achan. Inductive representation learning on temporal graphs. In *International Conference on Learning Representations (ICLR)*, 2020.

Yiming Xu, Bin Shi, Teng Ma, Bo Dong, Haoyi Zhou, and Qinghua Zheng. Cldg: Contrastive learning on dynamic graphs. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pp. 696–707. IEEE, 2023.

Yuanyuan Xu, Wenjie Zhang, Ying Zhang, Xuemin Lin, and Xiwei Xu. Unlocking multi-modal potentials for link prediction on dynamic text-attributed graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 27386–27394, 2026.

Le Yu, Leilei Sun, Bowen Du, and Weifeng Lv. Towards better dynamic graph learning: New architecture and unified library. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 67686–67700, 2023.

Jiasheng Zhang, Jialin Chen, Menglin Yang, Aosong Feng, Shuang Liang, Jie Shao, and Rex Ying. Dtgb: A comprehensive benchmark for dynamic text-attributed graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 91405–91429. Curran Associates, Inc., 2024.

Siwei Zhang, Yun Xiong, Yateng Tang, Jiarong Xu, Xi Chen, Zehao Gu, Xuehao Zheng, Zi'an Jia, and Jiawei Zhang. Unifying text semantics and graph structures for temporal text-attributed graphs with large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=9env0BdcDV.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025b.

Zulun Zhu, Kai Wang, Haoyu Liu, Jintang Li, and Siqiang Luo. Topology-monitorable contrastive learning on dynamic graphs. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4700–4711, 2024.

A BoE Architecture Illustration

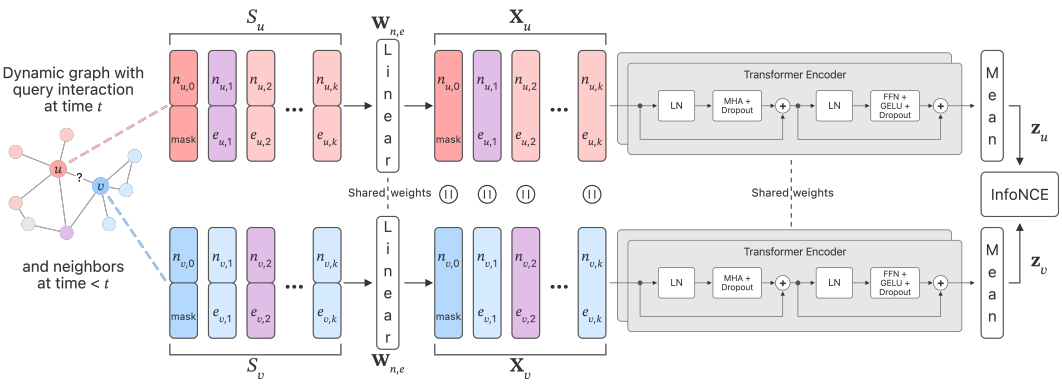


Figure 4: **Overview of the Bag of Events (BoE) Baseline. (Left)** For a query interaction between nodes u and v at time t, historical one-hop neighbors are retrieved. u and v combined with their respective neighbors form node sequences S_u and S_v. For each event i in these sequences, its node text embedding $\mathbf{n}_i$ and edge text embedding $\mathbf{e}_i$ are extracted, and the edge text embedding of the query nodes is masked. **(Center)** Both extracted features $\mathbf{n}_i$ and $\mathbf{e}_i$ are first independently projected via separate linear layers $\mathbf{W}_{n,e}$. These projected embeddings are then concatenated to form a single feature vector representing each event in S_u and S_v. **(Right)** These event sequences, corresponding to input matrices $\mathbf{X}_u$ and $\mathbf{X}_v$, are processed independently by a multi-layer Transformer encoder. Mean pooling, with a mask to exclude padding tokens, is applied to form the final node representations $\mathbf{z}_u$ and $\mathbf{z}_v$. These representations are optimized via a symmetric InfoNCE loss.

B Self-Supervised Pretraining Objectives

This section provides the formal mathematical formulations for the self-supervised pretraining objectives evaluated in this work, specifically the symmetric InfoNCE loss used for the Bag of Events (BoE) baseline and the Scaled Cosine Error (SCE) used in our ablation study.

Symmetric InfoNCE Loss For the BoE baseline, the model optimizes a symmetric contrastive objective to align the independent sequence representations of interacting nodes u and v. Given a minibatch of size N, let $\mathbf{z}_u^{(i)}$ and $\mathbf{z}_v^{(i)}$ denote the L_2-normalized, pooled representations for the i-th interaction in the batch. The symmetric InfoNCE loss (Oord et al., 2018) is formulated as:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{2N}\sum_{i=1}^{N}\left(\log\frac{\exp(\mathbf{z}_u^{(i)}\cdot\mathbf{z}_v^{(i)}/\tau)}{\sum_{j=1}^{N}\exp(\mathbf{z}_u^{(i)}\cdot\mathbf{z}_v^{(j)}/\tau)} + \log\frac{\exp(\mathbf{z}_v^{(i)}\cdot\mathbf{z}_u^{(i)}/\tau)}{\sum_{j=1}^{N}\exp(\mathbf{z}_v^{(i)}\cdot\mathbf{z}_u^{(j)}/\tau)}\right), \quad (3)$$

where τ is a temperature hyperparameter set to 0.07. The first term matches the representation of node u with its true interacting partner v against all other in-batch negative representations, while the second term performs the symmetric matching for node v.

Scaled Cosine Error (SCE) In the ablation study (Section 4.5), we evaluate a direct semantic forecasting (SF) variant of our model that replaces the CSF objective with a vector reconstruction loss, SCE, proposed by Hou et al. (2022). Let $\mathbf{z}_{pred}^{(i)}$ be the predicted representation and $\mathbf{z}_{target}^{(i)}$ be the true edge text embedding. The SCE loss is defined as:

$$\mathcal{L}_{\text{SCE}} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{\mathbf{z}_{pred}^{(i)}\cdot\mathbf{z}_{target}^{(i)}}{\|\mathbf{z}_{pred}^{(i)}\|_2\|\mathbf{z}_{target}^{(i)}\|_2}\right)^{\gamma}, \quad (4)$$

where γ is a scaling factor that down-weights easily reconstructed samples, incentivizing the model to focus training capacity on interaction dynamics that are more difficult to predict.

C Synthetic Graph Generation Parameters

For the representational analysis detailed in Section 4.1, we generate synthetic dynamic graphs for four distinct archetypes: Non-Recurrent Bipartite, Recurrent Bipartite, Asymmetric Homogeneous, and Symmetric Homogeneous. These archetypes simulate distinct structural patterns: one-off user-item reviews, repeated user-post interactions, imbalanced communication networks, and highly reciprocal event networks, respectively. Each graph contains 60 nodes and 1,500 interactions. We control structural variance across topologies by altering power-law degree distributions, community boundaries, node recurrence probabilities, and edge reciprocity. We control temporal variance by sampling interaction intervals from exponential distributions with varying scale parameters to produce both dense temporal bursts and sparse activity. Edges are partitioned into an 80% pretraining split and a 20% validation split for the InfoNCE objective. Below are the generation parameters.

Non-Recurrent Bipartite. Nodes are partitioned into 36 sources and 24 destinations. Source selection probability is proportional to $i^{-0.5}$, and destination probability to $i^{-0.8}$. Recurrence is strictly prohibited until a source node exhausts all 24 available destinations, at which point its history resets. This finite pool exhaustion by high-activity nodes creates a marginal fraction (1.48%) of recurrent edges where the root node appears in the destination sequence, producing $(w = 1, c = 1)$ tokens. Time steps Δt between interactions are drawn from $\text{Exp}(5.0) + 1$, producing slow, sparse activity.

Recurrent Bipartite. Nodes are equally partitioned into 30 sources and 30 destinations. Source selection probability is proportional to $i^{-1.1}$. There is a 45% probability of recurrence. Time steps are fast, drawn from $\text{Exp}(0.4)$ plus an additional step of 1 with a 60% probability.

Asymmetric Homogeneous. Nodes are divided into four 15-node communities. Out-degree probability is proportional to $i^{-0.9}$. In-degree probability is a random permutation of the out-degree distribution, breaking symmetry. We apply a 75% intra-community bias for destination selection. Time increments are binary (1 with a 15% probability, 0 otherwise) to simulate simultaneous interactions. Additionally, there is a 10% probability of immediate edge reciprocity ($v \to u$).

Symmetric Homogeneous. Similar to the asymmetric variant, nodes form four 15-node communities with out-degree probability proportional to $i^{-0.9}$. However, in-degree weights exactly match out-degree weights. The intra-community selection bias is increased to 80%. Time increments are 1 with a 12% probability. The immediate edge reciprocity rate is heavily increased to 55%, simulating dense, highly communicative subgraphs.

D SPATIO-TEMPORAL REPRESENTATIONS COLORED BY NODE FREQUENCY

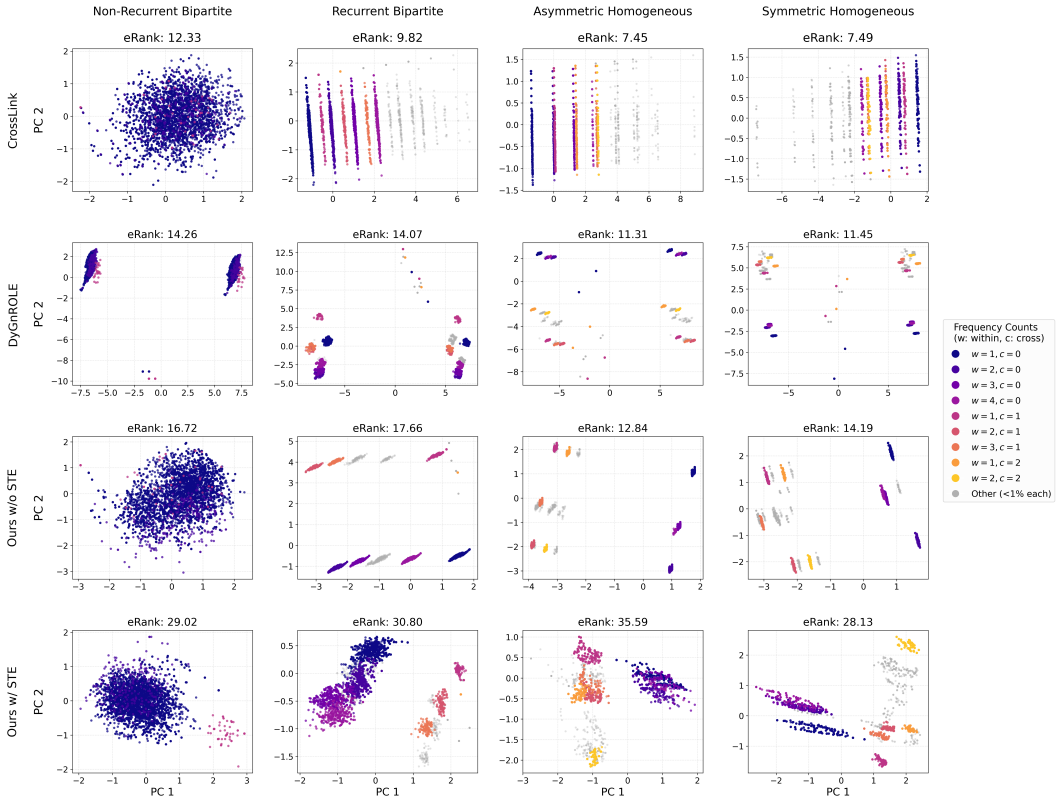


Figure 5: **Analysis of Spatio-Temporal Representations.** Principal Component Analysis (PCA) of spatio-temporal representations colored by combinations of within- and cross-sequence node frequency counts across four synthetic graph topologies. **Top 3**: Under domain transfer, concatenated time and node frequency features suffer from representational fragmentation and a restricted Effective Rank (eRank). **Bottom**: STE fuses the temporal and structural features into a unified manifold and substantially improves eRank.

E DATASET DETAILS

This section provides details for the datasets from the Dynamic Text-Attributed Graph Benchmark (Zhang et al., 2024) used in our evaluation, including summary statistics in Table 3. The underlying raw data for these benchmarks originates from the publicly available sources cited below.

Table 3: Summary statistics for the DTGB datasets.

Dataset	#Nodes	#Edges	#Classes	Domain
Enron	42,711	797,907	10	Corporate communication
GDELT	6,786	1,339,245	236	Political events
ICEWS1819	31,796	1,100,071	266	Political events
Stack elec	397,702	1,262,225	2	Technical Q&A forum
Stack ubuntu	674,248	1,497,006	2	Technical Q&A forum
Amazon movies	293,566	3,217,324	5	E-commerce reviews
Googlemap CT	111,168	1,380,623	5	E-commerce reviews
Yelp	2,138,242	6,990,189	5	E-commerce reviews

Enron[1] represents a network of email communications between employees of the Enron corporation from 1999 to 2002.

- **Nodes**: Employees
- **Node texts**: Professional email addresses
- **Edges**: Email exchanges between employees
- **Edge texts**: Raw email content
- **Edge labels**: 10 distinct communication categories (e.g., 'deal', 'calendar')

GDELT[2] is derived from the Global Database of Events, Language, and Tone. It records international political events and interactions.

- **Nodes**: Political entities or actors
- **Node texts**: Entity names
- **Edges**: Relationships or actions between two political entities
- **Edge texts**: Concise descriptions of the relationship
- **Edge labels**: 236 unique relationship types (e.g., 'provide economic aid')

ICEWS1819 (ICEWS)[3] originates from the Integrated Crisis Early Warning System and models political events occurring between 2018 and 2019.

- **Nodes**: Political entities
- **Node texts**: Entity names
- **Edges**: Documented relationships between political actors
- **Edge texts**: Short descriptions of the event or relationship
- **Edge labels**: 266 unique interaction categories (e.g., 'Engage in negotiation')

Stack elec (Stack E)[4] consists of data from the Stack Exchange forum focused on electronics and electrical engineering.

- **Nodes**: Users or specific questions
- **Node texts**: User introduction and location: question title and post content
- **Edges**: The submission of an answer or a comment on a post
- **Edge texts**: Textual content of answers or comments
- **Edge labels**: Binary classification ('useful' or 'useless') based on a voting score where $score = upvotes - downvotes$. An edge is 'useful' if $score > 1$.

Stack ubuntu (Stack U)[5] follows the same structure as Stack E but focuses on the Ubuntu operating system community.

- **Nodes**: Users or questions
- **Node texts**: User introduction and location: question title and post content
- **Edges**: Answer submissions or comments on existing posts
- **Edge texts**: Textual content of answers or comments
- **Edge labels**: Binary classification ('useful' or 'useless') where $score = upvotes - downvotes$ and 'useful' requires $score > 1$.

[1] https://www.cs.cmu.edu/~enron/
[2] https://www.gdeltproject.org/
[3] https://dataverse.harvard.edu/dataverse/icews
[4] https://archive.org/details/stackexchange
[5] https://archive.org/details/stackexchange

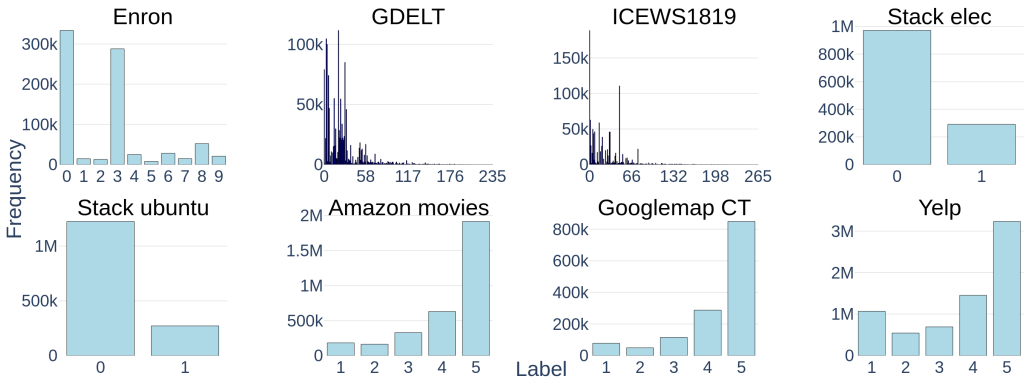


Figure 6: Label distributions of the DTGB datasets.

Googlemap CT (Google)[6] consists of business reviews from Google for the state of Connecticut.

- **Nodes**: Users or local businesses
- **Node texts**: Usernames for individuals: business name, address, category, and introduction
- **Edges**: Review interactions between users and businesses
- **Edge texts**: The textual content of user reviews
- **Edge labels**: 1 to 5 based on the standard five-star rating system

Amazon movies (Amazon)[7] consists of review data for the Movies and TV category on Amazon.

- **Nodes**: Users or products in the Movies and TV category
- **Node texts**: Usernames for individuals: product title, description, category, and rank score
- **Edges**: Reviews submitted by users for specific products
- **Edge texts**: The textual content of user reviews
- **Edge labels**: 1 to 5 based on the five-star rating system

Yelp[8] consists of reviews for businesses such as hotels and restaurants.

- **Nodes**: Users or businesses
- **Node texts**: Name, register date, review count, and average rating for users: business name, city, address, and category
- **Edges**: Review interactions between users and businesses
- **Edge texts**: The textual content of user reviews
- **Edge labels**: 1 to 5 based on the five-star rating system

F Multi-Dataset Pretraining and Interleaving Strategy

To facilitate cross-domain pretraining across the heterogeneous datasets in the LODO protocol, we implement a unified data orchestration layer. This layer ensures that the model is exposed to a balanced distribution of interaction dynamics from all source domains simultaneously, preventing the encoder from oscillating between domain-specific local minima.

[6] https://mcauleylab.ucsd.edu/public_datasets/gdrive/googlelocal/
[7] https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/
[8] https://www.yelp.com/dataset

Feature Stacking via Virtual Offsetting. Because each dataset in the DTGB benchmark uses an independent coordinate system for node and edge identifiers, naive concatenation leads to collision. We use a virtual feature stacking mechanism that creates a unified global index. For a set of source graphs $\{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_K\}$, we compute a cumulative offset $O_k = \sum_{i=1}^{k-1} |\mathcal{V}_i|$ for nodes and similarly for edges. During pretraining, the raw identifiers in $\mathcal{G}_k$ are mapped to a global space as $ID_{global} = ID_{local} + O_k$. This allows STSA and baseline models to treat the collection of source graphs as a single, large super-graph without requiring expensive physical merges of the feature tensors.

Chronological Batch Interleaving. To facilitate cross-domain pretraining, we use a deterministic round-robin interleaving strategy. While each individual mini-batch contains interactions from a single source dataset to ensure structural consistency, we interleave these batches during training. For N source datasets across the included pretraining domains, the model processes a sequence of N batches, one from each dataset, before repeating the cycle. This ensures that the global gradient trajectory is informed by the collective variance of all source datasets, preventing the encoder from overfitting to the idiosyncratic distribution of any single graph while adhering to the chronological order within each dataset.

G Model Hyperparameters

We implement all baselines using PyTorch. To ensure a fair comparison, we use standardized pretraining hyperparameters across all models unless a specific architecture requires otherwise.

Shared Settings

- Optimizer: AdamW
- Learning rate: 1×10^{-4}
- Batch size (N): 256
- Pretraining epochs: 1000
- Early stopping patience: 10 epochs
- Global random seed: 42
- Maximum number of neighbors (S): 10

BoE

- Transformer layers: 2
- Transformer heads: 2
- Transformer dropout rate: 0.1
- Text channel dimension: 96

STSA

- Transformer layers: 2
- Transformer heads: 2
- Transformer dropout rate: 0.1
- Text channel dimension: 96
- Time dimension: 32
- Node frequency embedding dimension: 32
- Spatio-temporal encoder hidden dimension: 256
- Spatio-temporal encoder latent dimension: 96

IDOL[9]

[9] https://github.com/ZulunZhu/dynamic-contrastive-learning

- Hidden dimension: 512
- Embedding dimension: 128
- Neighborhood samples (K): 4
- Temporal similarity weight (α): 0.1
- Maximum bias (r^b_{max}): 0.01
- Loss trade-off (λ): 0.01

CLDG[10]

- InfoNCE temperature: 0.07
- Hidden dimension: 128
- Embedding dimension: 64
- Encoder layers: 2
- Dropout rate: 0.1

TLP[11]

- Alignment weight (α): 1.0
- Alignment objective: Mean Squared Error
- Link prediction objective: Binary Cross-Entropy
- Topological normalization: Offline (global statistics)

GCAL[12]

- Entropy weight: 5.0
- Replay weight: 10.0
- VAE weight: 1.0
- Edge weight: 1.0
- Teacher EMA α: 0.9
- Bilevel optimization inner loops: 10
- Memory optimizer learning rate: 1×10^{-3}
- Warmup epochs: 2

DVGMAE

- Hidden dimension: 128
- Embedding dimension: 32
- Dropout rate: 0.1
- Epsilon (ϵ): 0.25
- Mask ratio: 0.5
- Loss lambda weight (λ): 0.5

DyGMAE

- Hidden dimension: 128
- Encoder layers: 2
- Dropout rate: 0.1

[10] https://github.com/yimingxu24/CLDG
[11] https://github.com/google-research/google-research/tree/master/fm4tlp
[12] https://github.com/joe817/GCAL

- Number of masks: 2
- Mask ratios: 0.2, 0.4
- Mask types: path, random
- Fusion type: attention
- Walk length: 2
- Loss gamma (γ): 0.1
- Loss delta (δ): 0.1
- Loss lambda c (λ_c): 0.1

DyGnROLE

- Transformer layers: 2
- Transformer heads: 2
- Transformer dropout rate: 0.1
- Text channel dimension: 32
- Time dimension (d_t): 32
- Node frequency embedding dimension (d_c): 32

H Computing Resources

Experiments were performed using an Intel Xeon Silver 4214 CPU (2.20GHz), 128 GB of system RAM, four NVIDIA GeForce RTX 2080 Ti GPUs (11 GB VRAM each) for all tasks.

I Unseen Class Dilution of Finetuning Results

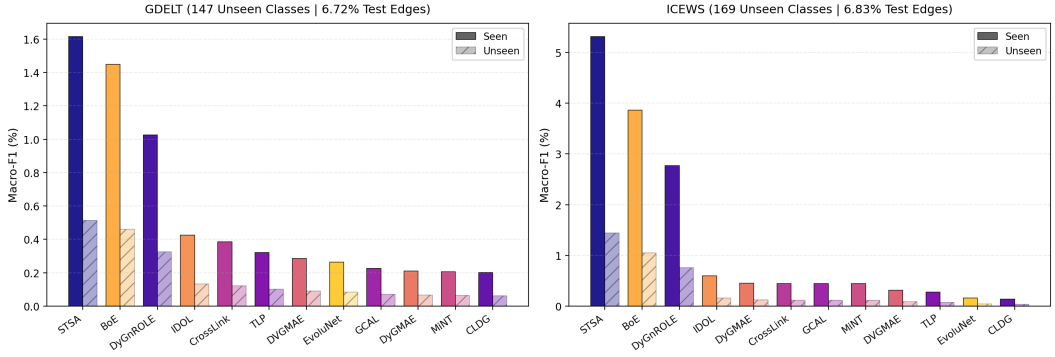


Figure 7: **Impact of Unseen-Class Dilution on Macro-F1.** Because Macro F1 weights all classes equally, the unobserved classes in the 500 training edges of the finetuning experiment performed in Section 4 drag down the overall score on GDELT and ICEWS. Isolating evaluation to seen classes removes this dilution artifact, revealing larger absolute gains for STSA on GDELT and ICEWS.