GCL-OT: Graph Contrastive Learning with Optimal Transport for Heterophilic Text-Attributed Graphs

Yating Ren, Yikun Ban, Huobin Tan*

Beihang University
renyating@buaa.edu.cn, yikunb@buaa.edu.cn, thbin@buaa.edu.cn

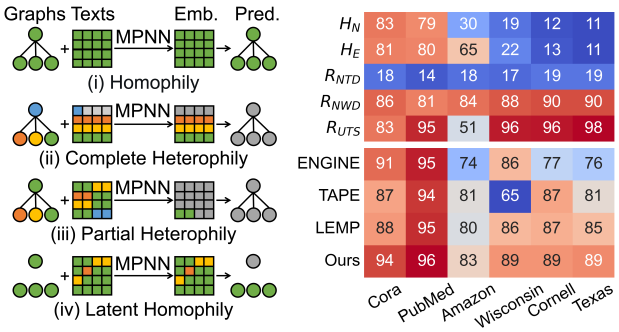
Abstract

Recently, structure–text contrastive learning has shown promising performance on text-attributed graphs by leveraging the complementary strengths of graph neural networks and language models. However, existing methods typically rely on homophily assumptions in similarity estimation and hard optimization objectives, which limit their applicability to heterophilic graphs. Although existing methods can mitigate heterophily through structural adjustments or neighbor aggregation, they usually treat textual embeddings as static targets, leading to suboptimal alignment. In this work, we identify the multi-granular heterophily in text-attributed graphs, including complete heterophily, partial heterophily, and latent homophily, which makes structure–text alignment particularly challenging due to mixed, noisy, and missing semantic correlations. To achieve flexible and bidirectional alignment, we propose GCL-OT, a novel graph contrastive learning framework with optimal transport, equipped with tailored mechanisms for each type of heterophily. Specifically, for partial heterophily, we design a RealSoftMax-based similarity estimator to emphasize key neighbor-word interactions while easing background noise. For complete heterophily, we introduce a prompt-based filter that adaptively excludes irrelevant noise during optimal transport alignment. Furthermore, we incorporate OT-guided soft supervision to uncover potential neighbors with similar semantics, enhancing the learning of latent homophily. Theoretical analysis shows that GCL-OT can improve the mutual information bound and Bayes error guarantees. Extensive experiments on nine benchmarks show that GCL-OT outperforms state-of-the-art methods, demonstrating its effectiveness and robustness.

Code, datasets and extended version — github.com/users-01/GCL-OT

Introduction

Text-attributed graphs (TAGs) represent text entities as nodes and their relationships as edges, widely used across diverse real-world domains, such as academic citations, web hyperlinks, and e-commerce recommendations (Yan et al. 2023). Recent advances in Large Language Models (LLMs) have shown substantial capability to capture the semantic richness, prompting researchers to combine Language Models (LMs) with Graph Neural Networks (GNNs) on TAGs for performance enhancement (He et al. 2024; Pan et al. 2024; Liu et al. 2025a). Among them, joint training strategies using Graph Contrastive Learning (GCL) (Brannon et al. 2024; Fang et al. 2024) enhance alignment between textual and structural representations by optimizing mutual information to distinguish positive and negative examples.



(a) Examples (b) Empirical analysis

Figure 1: Analysis of multi-granular heterophily in TAGs. Node colors denote categories. H_N/H_E: node/edge heterophily, $R_{\text{NTD}}/R_{\text{NWD}}$: neighbor token/sentence dissimilarity, R_{UTS}: similarity of unconnected nodes.

However, heterophilic TAGs are prevalent in reality due to the principle of opposites attracting, e.g., dating networks. Early GNN and GCL methods typically adopt the non-local neighbor extension and architectural refinement strategies (Chen, Lei, and Wei 2024; Wang et al. 2024a). Despite effectiveness, these approaches may be limited by the shallow embedding (Mikolov et al. 2013; Zhang, Jin, and Zhou 2010). Few methods study heterophily by combining deep semantic representations from LMs. Only LLM4HeG (Wu et al. 2025) and LEMP4HG (Wang and Cheng 2025) explore utilizing LLMs/LMs for heterophilic edge discrimination and reweighting, but face sub-optimal challenges with sparse or noisy data in the cascading pipelines (Pan et al. 2024; Liu et al. 2025a).

In this paper, we identify that TAGs exhibit multi-granular heterophily from three node-centric perspectives grounded

*Corresponding author

in the text-feature view, as illustrated in Figure 1a: (i) *Partial heterophily* occurs when only part of a node's text semantically aligns with its neighbors, and vice versa. Existing methods still struggle to capture subtle semantic mismatches (Song et al. 2023). (ii) *Complete heterophily* appears when a node's text has no relevance with its neighbors at all, typically caused by meaningless texts or connectivity (such as random co-purchases). Even state-of-the-art models can be misled by such signals (Li et al. 2025). (iii) *Latent homophily* denotes potential semantic neighbors that are disconnected, often due to missing or implicit links. While multi-hop propagation (Zhu et al. 2020) and potential neighbor discovery (Pei et al. 2020) exploit such signals, fully leveraging them remains open. As shown in Figure 1b, these patterns are prevalent in real-world TAGs and naturally induce many-to-many (N:N) alignment between structural and textual neighborhoods, whereas DGIs nearly N:1 summary pairing and InfoNCEs approximately 1:1 correspondence fail to capture this. Fortunately, optimal transport (OT) can fractionally allocate mass for soft N:N alignment when 1:1 is ambiguous, enabling mutual coverage via OT constraints to avoid one-sided greedy matches(Xu et al. 2023).

To address the multi-granular heterophily challenges in TAGs, we further propose GCL-OT, a novel GCL framework with tailored mechanisms grounded in OT theory, aiming to hierarchically and softly align structural and textual representations. Specifically, for partial heterophily, the RealSoftMax-based similarity estimation can identify and emphasize the most relevant word–neighbor pairs bidirectionally. For complete heterophily, the global filter-prompt strategy is employed to mitigate the negative impact of irrelevant embeddings during alignment. For latent homophily, OT assignment serves as the auxiliary supervision within the GCL objective to discover hidden neighbors. We analyze GCL-OT using mutual information (MI) theory, showing it can tighten the InfoNCE MI lower bound and reduce Bayes error in downstream node classification. Extensive experiments on nine TAG benchmarks demonstrate that GCL-OT outperforms strong baselines, highlighting its robustness and potential in both homophilic and heterophilic settings. Our main contributions are summarized as follows:

- To our knowledge, this is the first work to incorporate OT into GCL for heterophilic TAGs, enabling flexible bidirectional alignment of structural and textual views.
- Three mechanisms tailored to multi-granular heterophily patterns: RealSoftMax-based similarity for partial heterophily, filter-prompt for complete heterophily, and OT-guided supervision for latent homophily. The theoretical analysis shows tighter MI guarantees.
- Extensive experiments on nine benchmarks, in both homophilic and heterophilic scenarios, demonstrate the effectiveness and robustness of the proposed GCL-OT compared with state-of-the-art methods.

Related Work

Text-Attributed Graph Learning. Representation learning on Text-Attributed Graphs (TAGs) has attracted significant attention in graph machine learning. Traditional approaches typically couple shallow textual embeddings, such as bag-of-words or FastText, with GNNs (Kipf and Welling 2017; Hamilton, Ying, and Leskovec 2017; Veličković et al. 2018). Recent efforts combine GNNs with LMs for richer textual encoding in cascading, iterative, and parallel manners. Cascading methods freeze GNN or LM as an auxiliary to enrich the other, while the two-stage, separately encoding often produce sub-optimal integration (He et al. 2024; Wu et al. 2025; Wang and Cheng 2025). Iterative methods alternately optimize LMs and GNNs under the shared objective, improving label efficiency but potentially increasing computational costs (Yang et al. 2021; Zhao et al. 2023). Parallel methods, mostly based on GCL, align LMs and GNNs in a shared space (Zhu et al. 2024; Fang et al. 2024). Congrat (Brannon et al. 2024) and GraphGPT (Tang et al. 2024) align them by node-level InfoNCE objectives, which are intuitive but vulnerable to complex relations and noise. HASE-code (Zhang et al. 2024a) and G2P2 (Wen and Fang 2023) extend the alignment granularity to subgraphs, but may risk over-smoothing due to high-order structures. Few approaches consider heterophily in TAGs. LLM4HeG (Wu et al. 2025) uses LLMs to label heterophilic edges from node-pair texts for reweighting. LEMP4HG (Wang and Cheng 2025) refines it via key-pair selection and attention integration. However, their two-stage pipelines make them sensitive to noisy text (Pan et al. 2024; Liu et al. 2025a).

Heterophilic Graph Learning. Heterophily is generally considered a key challenge for GNN performance. Current researches typically fall into non-local neighbor extension and architectural refinement strategies (Zheng et al. 2022; Gong et al. 2024). The first extends non-local neighbors with similar attributes through mixing high-order neighbors (Abu-El-Haija et al. 2019; Song et al. 2023) or discovering potential neighbors based on various distances (Pei et al. 2020). The second refines GNN architecture by enhancing message aggregation from similar neighbors while minimizing the influence of dissimilar ones (Bo et al. 2021; Zhu et al. 2021; Liang et al. 2023), or mixing layers to capture information from different neighbor ranges (Xu et al. 2018; Zhu et al. 2020). Among them, recent GCL methods offer noteworthy advancements (Chen et al. 2025). PolyGCL (Chen, Lei, and Wei 2024) designs and contrasts learnable spectral polynomial filters for varying homophily levels, while HeterGCL (Wang et al. 2024a) incorporates structural and semantic modules to utilize label-inconsistent signals effectively. However, these methods rely on static shallow embeddings and struggle to capture context-aware information and complex semantic relationships, limiting their effectiveness in leveraging text attributes (He et al. 2024).

Graph Contrastive Learning with OT Optimal transport (OT) (Monge 1781; Villani 2009) is a mathematical framework for measuring distances between distributions, finding the most cost-efficient way to transform one distribution into another (Vincent-Cuaz et al. 2022; Lin et al. 2023; Xu et al. 2023). In Graph Contrastive Learning (GCL), traditional InfoNCE loss often relies on hard alignment with pairing positive and negative samples, which may introduce representation bias. To this end, researchers solve the problems caused

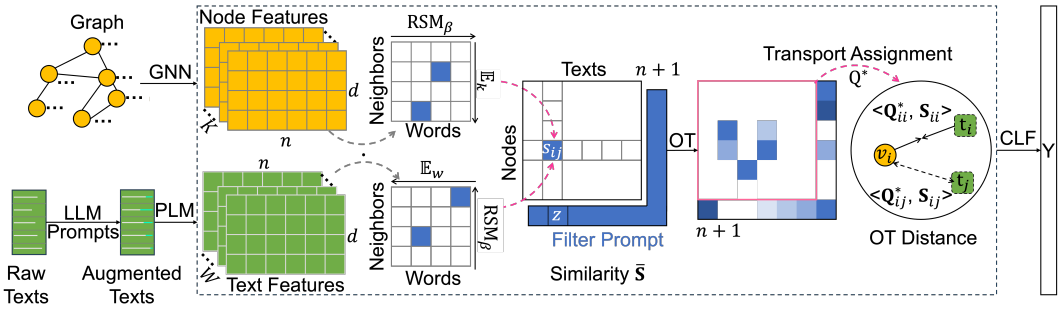


Figure 2: Overview of GCL-OT. Given a TAG, an LLM enriches node texts, a PLM encodes the enriched texts, and a GNN captures structure features. The text and structure views form a similarity matrix, where RealSoftMax highlights fine-grained interactions and the filter prompt suppresses coarse-grained noise. The contrastive module then aligns the two views and uncovers latent homophily. Finally, the fused embeddings drive node prediction.

by non-aligned contrast views or structural differences in the graph by introducing OT distance for soft alignment, subgraph generation, or cross-view comparison (Wang et al. 2023b; Deng et al. 2025), thereby enhance node representation learning, which subsequently benefits downstream tasks such as node clustering (Zhang et al. 2024b; Wang et al. 2024b; Deng et al. 2025), node classification (Xie and Giraldo 2024; SANGARE et al. 2025; Zhu et al. 2022; Liu et al. 2025b), and node anomaly detection (Wang et al. 2023a). Prior OT-based GCL methods (e.g., THESAURUS, FOSSIL) already treat OT as a soft-alignment remedy to InfoNCE's hard pairing, but they operate on homogeneous graphs or prototype graphs.

Method

To address multi-granular heterophily in text-attributed graphs, we propose GCL-OT, a graph contrastive learning framework with optimal transport designed to align structure and text representations gradually. The overall architecture of GCL-OT is shown in Figure 2.

Notations

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{T}, \mathbf{y})$ denotes a text-attributed graph, where $\mathcal{V} = \{v_i\}_{i=1}^N$ is a set of N nodes, $\mathcal{E}$ is a set of M edges, $\mathcal{T} = \{t_i\}_{i=1}^N$ is text attributes of nodes, and $\mathbf{y} = [y_i]_{i=1}^N$ denotes node label vector drawn from a fixed set of categories. Given labeled nodes for training/validation, the goal is to predict labels for the remaining nodes.

Prompt-Enhanced Multi-View Feature Encoding

Heterophilic text-attributed graphs (TAGs) present a significant challenge, as node features and structural neighbors often belong to different classes. To overcome this inconsistency, we construct a multi-view encoding scheme that integrates both textual and structural signals.

Raw textual attributes are typically noisy and semantically ambiguous. To enrich their expressiveness, we augment each node's text using a task-specific prompt issued to a frozen large language model (LLM) following (He et al. 2024). The original and prompted texts are concatenated to form an enhanced textual description t_i^{aug} for each node v_i.

Then, we encode t_i^{aug} with a lightweight PLM (e.g., DistilBERT (Sanh et al. 2020)) to obtain token-level embeddings $\mathbf{H}_i^{\varpi} \in \mathbb{R}^{W \times d_t}$ and the sentence-level embedding $\mathbf{h}_i^t \in \mathbb{R}^{d_t}$. In parallel, we encode structures using a graph neural network (GNN). Node features are initialized and propagated through the GNN to produce structural embeddings $\mathbf{H}^{\zeta} \in \mathbb{R}^{N \times d_s}$, where d_s is the structural dimension. For each node v_i, we derive a neighborhood-level embedding matrix $\mathbf{H}_i^{\mathcal{N}} \in \mathbb{R}^{K \times d_s}$ by aggregating and linear projecting the embeddings of its neighborhood $\mathcal{N}(v_i) = \{v_k \mid e_{ik} \in \mathcal{E}\} \cup \{v_i\}$, where $K = |\mathcal{N}(v_i)|$. These four views provide the foundation for contrastive objectives.

GCL with OT for Hierarchical Alignment

In partially heterophilic settings, simply averaging neighbor and word embeddings may weaken salient signals, while hard-max pooling is often extreme and unstable. To more balancedly emphasize meaningful interactions between neighbors and words, we introduce a soft maximum similarity estimation mechanism based on RealSoftMax.

Specifically, given the k-th neighbor embedding vector $\mathbf{h}_{i,k}^{\mathcal{N}}$ of node v_i, and the w-th word embedding vector $\mathbf{h}_{j,w}^{\varpi}$ of node v_j. The similarity between v_i and v_j is defined as

$$s_{ij} = \frac{1}{2}(\mathbb{E}_k[\text{RSM}_\beta(\{\mathbf{h}_{ik}^{\mathcal{N}} \cdot \mathbf{h}_{jw}^{\varpi}\}_{w=1}^W)] + \mathbb{E}_w[\text{RSM}_\beta(\{\mathbf{h}_{iw}^{\varpi} \cdot \mathbf{h}_{jk}^{\mathcal{N}}\}_{k=1}^K)]), \tag{1}$$

where $\mathbb{E}(\cdot)$ is the mean operator, $\text{RSM}_\beta(\{x_\ell\}) = \beta \log \sum_\ell \exp(x_\ell/\beta)$ smoothly interpolates between mean ($\beta \to \infty$) and maximum ($\beta \to 0$). The first term highlights the most relevant words for each neighbor, the second does the opposite, jointly emphasizing informative cross-view interactions and downweighting background noise.

To address the challenge of complete heterophily, we introduce a prompt-based filtering mechanism that reduces noise that can not align with others. Specifically, the global similarity matrix $\hat{\mathcal{S}}$ between structural embedding $\mathbf{H}^{\zeta}$ and textual embedding $\mathbf{H}^t$ is merged into $\mathbf{S}$, and then $\mathbf{S}$ is expanded with additional row and column prompt vectors,

$$\bar{\mathbf{S}} = \begin{bmatrix} \mathbf{S} & \mathbf{z} \\ \mathbf{z}^\top & z_{N+1} \end{bmatrix} \in \mathbb{R}^{(N+1)\times(N+1)}, \tag{2}$$

where $\mathbf{z}$ is a learnable vector. Embeddings whose maximum similarity falls below the corresponding $\mathbf{z}$ are instead aligned with the prompt vector during the subsequent alignment.

Based on $\bar{\mathbf{S}}$, we adopt the OT distance as the similarity measure for robust and efficient alignment. Let $\mathbf{Q} \in \mathbb{R}_{+}^{(N+1)\times(N+1)}$ denote the transport assignment of $\bar{\mathbf{S}}$, where q_{ij} is the corresponding probabilities. OT aims to establish a flexible alignment by maximizing $\langle \mathbf{Q}, \bar{\mathbf{S}} \rangle = \mathrm{tr}(\mathbf{Q}^\top \bar{\mathbf{S}})$. The optimization problem can be defined as

$$\begin{aligned} &\max_{\mathbf{Q}\in\mathcal{Q}} \langle \mathbf{Q}, \bar{\mathbf{S}} \rangle + \varepsilon \, \mathrm{H}(\mathbf{Q}) \\ &\text{s.t. } \mathcal{Q} = \left\{ \mathbf{Q} \mid \mathbf{Q}\mathbf{1}_{N+1} = \boldsymbol{\mu}_1, \mathbf{Q}^\top \mathbf{1}_{N+1} = \boldsymbol{\mu}_2 \right\}, \end{aligned} \tag{3}$$

where $\mathbf{1}_{N+1} \in \mathbb{R}^{N+1}$ is an all-one vector, entropy term $\mathrm{H}(\mathbf{Q}) = -\sum_{ij} q_{ij} \log q_{ij}$ is a smooth convex regular, ε is weight of entropy, $\boldsymbol{\mu}_1 \in \mathbb{R}^{N+1}$ and $\boldsymbol{\mu}_2 \in \mathbb{R}^{N+1}$ are relative importance of structural and textual embeddings. To avoid bias, let $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ follow the uniform distribution $\frac{\mathbf{1}_{N+1}}{N+1}$.

As shown in Equation (3), OT can realign each node's structure or text with multiple related texts or structures based on similarity, effectively resolving potential issues between structures and texts. Let the kernel matrix $\mathbf{K} = \exp(\bar{\mathbf{S}}/\varepsilon)$, and we approximate $\mathbf{K}$ with a non-negative low-rank factorization

$$\mathbf{K} \approx \mathbf{U} \, \mathrm{diag}(\Upsilon) \mathbf{V}^\top, \tag{4}$$

where $\mathbf{U} \in \mathbb{R}_{+}^{(N+1)\times r}$ and $\mathbf{V} \in \mathbb{R}_{+}^{(N+1)\times r}$ are non-negative basis matrices, $\mathrm{diag}(\Upsilon) \in \mathbb{R}^{r\times r}$ turns the weight vector into a diagonal matrix, and target rank $r \leq N$. The optimal $\mathbf{Q}^*$ of Equation (3) has a simple normalized exponential matrix solution by LRSinkhorn (Scetbon, Cuturi, and Peyré 2021):

$$\mathbf{Q}^* = [\mathrm{Diag}(\boldsymbol{\kappa}_1) \mathbf{K} \, \mathrm{Diag}(\boldsymbol{\kappa}_2)]_{1:N,1:N}, \tag{5}$$

with iteratively updated

$$\begin{cases} \boldsymbol{\kappa}_1 \leftarrow \boldsymbol{\mu}_1 ./ (\mathbf{K} \cdot \boldsymbol{\kappa}_2), \\ \boldsymbol{\kappa}_2 \leftarrow \boldsymbol{\mu}_2 ./ (\mathbf{K}^\top \cdot \boldsymbol{\kappa}_1), \end{cases} \tag{6}$$

where $\boldsymbol{\kappa}_1 \in \mathbb{R}^N$, $\boldsymbol{\kappa}_2 \in \mathbb{R}^N$ are the non-negative left and right scaling vectors.

By utilizing OT distance as the similarity criterion, and let

$$d_{ij} = \exp(q_{ij}^* \bar{s}_{ij}/\tau), \tag{7}$$

where q_{ij}^* is the corresponding transport assignment of $\bar{s}_{ij}$, and τ is the temperature parameter. To encourage strong alignment of matching pairs while repelling mismatches, we minimize the bi-directional contrastive loss $\mathcal{L}_{\text{MHA}}$ as

$$\mathcal{L}_{\text{MHA}} = -\mathbb{E}_i \left[\log \frac{d_{ii}}{\sum_{j=1}^{r} d_{ij}} + \log \frac{d_{ii}}{\sum_{j=1}^{r} d_{ji}} \right], \tag{8}$$

where we retain r samples with the largest scores as negatives. For irrelevant inputs, the OT plan q_{ij}^* distributes mass more broadly, resulting in a lower alignment score s_{ij} and a decreased affinity d_{ij}. Thus, minimizing $\mathcal{L}_{\mathbf{MHA}}$ penalizes them by increasing distance in embedding space, enhancing the contrastive signal.

OT-Guided Latent Homophily Mining

Randomly sampled negative samples and one-hot labels may mistakenly penalize semantically similar but unconnected potential neighbors as negative samples. We leverage the OT assignment matrix as auxiliary supervision to reduce the distances between potential neighbors.

Specifically, given the global similarity matrix $\hat{\mathbf{S}}$ between structural embeddings $\mathbf{H}^{\zeta}$ and textual embeddings $\mathbf{H}^{t}$, the OT problem can be formulated and solved as illustrated in Equation (3)–(6), then it generates the assignment $\hat{\mathbf{Q}}^*$.

Considering the diagonal matrix $\mathbf{I}$ encoding self-positives, and $\hat{\mathbf{Q}}^*$ encoding latent positives. The contrastive target $\mathbf{P}$ can be defined as their combination:

$$\mathbf{P} = \mathbf{I} + \hat{\mathbf{Q}}^*, \tag{9}$$

where $\mathbf{P}$ is normalized to obtain a valid probability distribution for each anchor. Based on the soft target $\mathbf{P}$, the Latent Homophily Mining (LHM) loss can be defined as

$$\mathcal{L}_{\mathbf{LHM}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{r} p_{ij} (\log \hat{k}_{ij}^{'} + \log \hat{k}_{ij}^{''}) \tag{10}$$

where $\hat{k}_{ij}^{'} = \mathrm{softmax}_j(\hat{s}_{ij}/\tau)$ denotes the row-wise softmax over index j, $\hat{k}_{ij}^{''} = \mathrm{softmax}_i(\hat{s}_{ij}/\tau)$ denotes denotes the row-wise softmax over index i. By minimizing $\mathcal{L}_{\mathbf{LHM}}$, GCL-OT can implicitly increase the mutual attraction among similar embeddings, thereby strengthening the contrastive learning objective.

Optimization Objective

To achieve both effective alignment between textual and structural information and strong discriminative power for node classification, GCL-OT is trained with a composite objective that couples a contrastive component with the standard node-classification term,

$$\mathcal{L} = \mathcal{L}_{\text{NC}} + \lambda \mathcal{L}_{\text{GCL-OT}}, \tag{11}$$

where $\mathcal{L}_{\text{NC}}$ is the cross-entropy loss for node classification, λ balances contributions of the contrastive term.

Theoretical Analysis

In this section, we explain the effectiveness of GCL-OT from the view of Mutual Information (MI) theory.

Proposition 1. *Let $\mathcal{L}_{InfoNCE}$ denote the standard InfoNCE loss between structural embeddings $\mathbf{H}^{\zeta}$ and textual embeddings $\mathbf{H}^{t}$. According to the standard InfoNCE MI lower bound (van den Oord, Li, and Vinyals 2019), $\mathcal{L}_{LHM}$ provides a tighter variational lower bound than InfoNCE*

$$\mathrm{MI}(\mathbf{H}^{\zeta}, \mathbf{H}^{t}) \geq \log N - \mathcal{L}_{MHA} \geq \log N - \mathcal{L}_{InfoNCE}. \tag{12}$$

Proposition 2. *According to the standard InfoNCE MI lower bound (van den Oord, Li, and Vinyals 2019), the latent-homophily objective $\mathcal{L}_{LHM}$ also satisfies*

$$\mathrm{MI}(\mathbf{H}^{\zeta}, \mathbf{H}^{t}) \geq \log N - \mathcal{L}_{LHM} \geq \log N - \mathcal{L}_{InfoNCE}. \tag{13}$$

Propositions 1 and 2 claim that minimizing $\mathcal{L}_{\text{MHA}}$ and $\mathcal{L}_{\text{LHM}}$ for optimization maximizes a variational lower bound on mutual information than $\mathcal{L}_{\text{InfoNCE}}$. Based on these propositions, we can derive Theorem 1.

Theorem 1. *Let $f^{\zeta}(\cdot)$ and $f^{t}(\cdot)$ denote deterministic encoder functions for structural and textual information, respectively. For each node v_i, let $\mathbf{x}_i^{(k)} = \{\mathbf{x}_j\}_{j\in\mathcal{N}(\mathbf{x}^{(k)},i)}$ denote its k-hop neighborhood that collectively maps to its high-level features, $\mathbf{h}_i^{\zeta} = f^{\zeta}(\mathbf{x}_i^{(k)})$, $\mathbf{h}_i^{t} = f^{t}(\mathbf{x}_i^{(k)})$. Let $\mathbf{h}_i = \text{Linear}(\mathbf{h}_i^{\zeta}, \mathbf{h}_i^{t})$ denotes the final embedding. Assume that the weight matrix of the $\text{Linear}(\cdot,\cdot)$ has full column rank. Then minimising loss in Equation (11) maximizes the variational lower bounds on $\text{MI}(\mathbf{H}^{\zeta}, \mathbf{H}^{t})$ and thus $\text{MI}(\mathbf{x}_i^{(k)}, \mathbf{h}_i)$.*

Remark. For each view $i \in \{\zeta, t\}$, let P^i_{pos} and P^i_{neg} denote the positive and negative pair distributions, $\text{JS}^{\zeta}(P^{\zeta}_{\text{pos}}, P^{\zeta}_{\text{neg}})$ and $\text{JS}^{t}(P^{t}_{\text{pos}}, P^{t}_{\text{neg}})$ denote the corresponding Jensen-Shannon divergences (JSD). Assume that under the standard one positive and many negative construction and the optimal discriminators, we can rewrite $\mathcal{L}_{\text{GCL-OT}} \geq 4\log r - 2(\text{JS}^{\zeta} + \text{JS}^{t})$, which proves that minimizing $\mathcal{L}_{\text{GCL-OT}}$ not only maximizes MI, but also encourages a larger JSD between positive and negative distributions in both structural and textual views.

Connection with downstream tasks. Optimizing objective in Equation (11), GCL-OT provides a tighter upper bound on the downstream Bayes error (Tsai et al. 2021; Xiao et al. 2022) compared to $\mathcal{L}_{\textbf{InfoNCE}}$ and $\mathcal{L}_{\textbf{NC}}$. This suggests that downstream tasks can benefit from the learned representations obtained through our loss.

Time Complexity Analysis

Training in GCL-OT mainly consists of feature encoding and contrastive learning. The GNN encoder (GCN) costs $\mathcal{O}(|E|D)$, the PLM encoder costs $\mathcal{O}(NW^2D)$. RealSoftMax similarity over K neighbors and W tokens costs $\mathcal{O}(NKWD)$, with $K \ll N$. Projecting both embeddings onto r-rank bases and keeping r similarities per row costs $\mathcal{O}(NrD)$. The rank-r LR Sinkhorn costs $\mathcal{O}(Nr + r^2)$. The contrastive losses add $2\mathcal{O}(Nr)$. Consequently, the overall per-epoch complexity is $\mathcal{O}(|E|D + NW^2D + NrD + Nr)$. In typical benchmarks, the $\mathcal{O}(NW^2D)$ term dominates, so GCL-OT scales nearly linearly with the number of nodes and feature dimension, and quadratically with text length.

Experiments

Experimental Settings

To evaluate the performance of GCL-OT, the node classification tasks are conducted on 9 real-world datasets across different homophily degrees. Cora, PubMed (Sen et al. 2008), ArXiv, and Products (Hu et al. 2020) generally regarded as homophilic, whereas Amazon (Platonov et al. 2023), ArXiv23 (He et al. 2024), Wisconsin, Cornell, and Texas (Pei et al. 2020) fall into heterophilic category. Original textual information for heterophilic graphs was collected from publicly available sources (Leskovec, Adamic, and Huberman 2007; Craven et al. 1998; Wu et al. 2025).

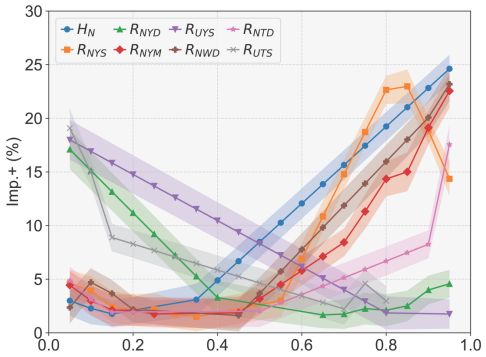


Figure 3: Improvements over InfoNCE across various heterophily metrics on Cora.

Experiments are conducted on one NVIDIA 4090 GPU and three NVIDIA 3080 GPUs. Text augmentation uses GPT-3.5 following (He et al. 2024). The text encoder employs a partially frozen DistilBERT (Sanh et al. 2020) model with six layers. Graph encoder adopts GCN (Kipf and Welling 2017), GAT (Veličković et al. 2017), and GraphSAGE (Hamilton, Ying, and Leskovec 2017), each configured with two layers. GAT utilizes eight attention heads, and GraphSAGE samples all neighbors. The classifier adopts a three-layer MLP with ReLU. Dimension of Text and Graph encoder is set to 768, with learning rates of 0.01 and 2e-5, weight decay at 0.0005, dropout from 0 to 0.8, and a batch size of 512. ArXiv and Products follow the standard splits from (Hu et al. 2020), the other datasets use random splits (60/20/20) following (He et al. 2024). Baseline hyperparameters follow prior work. All experiments report the average accuracy and standard deviation over 10 random seeds.

Experimental Results

Case Study. We first study the improvements over InfoNCE across various heterophily metrics by perturbing edges and texts. As shown in Figure 3, GCL-OT achieves substantial gains over the InfoNCE variant, obtained by replacing $\mathcal{L}_{\text{GCL-OT}}$ with $\mathcal{L}_{\text{InfoNCE}}$, with significant improvements under mixed-label neighborhoods and strong semantic heterophily. These trends indicate that GCL-OT can handle label mixing and semantic disagreement within neighborhoods, while the abundance of purely heterophilic neighborhoods and unlinked similar pairs may limit its benefits.

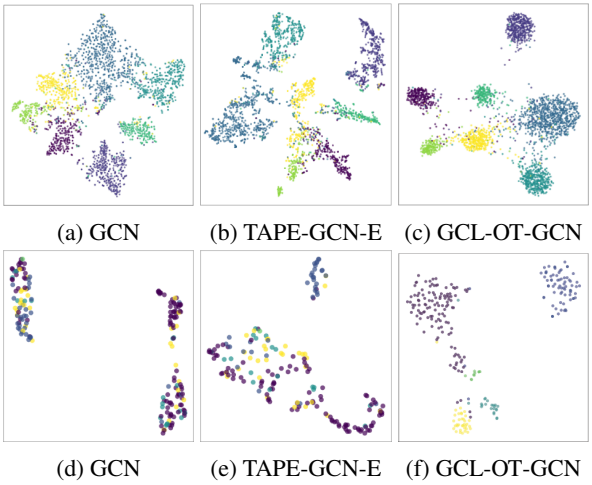
Results on Node Classification. Table 1 shows that GCL-OT substantially outperforms MLP, LM-only, and classical GNN baselines under the supervised setting, highlighting the benefit of jointly modeling text and structure. It also surpasses TAG methods and heterophilic GNNs, indicating that GCL-OT can effectively mitigate heterophily rather than overfitting to homophilic patterns. In the unsupervised setting, we train GCL-OT on raw node texts with no labels, then freeze the model and train a one-layer linear classifier on 60%/20%/20% splits of the embeddings and labels. As

Methods	Cora	PubMed	Products	ArXiv	Amazon	ArXiv23	Wisconsin	Cornell	Texas
$H_{\mathbf{N}}$	82.52	79.24	78.97	63.53	37.57	29.66	18.68	12.12	10.68
MLP	63.88±2.1	86.35±0.3	61.06±0.1	53.36±0.4	47.87±0.9	62.02±0.6	85.85±6.8	72.05±7.1	80.26±7.3
DistilBERT	76.06±3.8	94.94±0.5	72.97±0.2	73.61±0.0	65.42±0.7	73.58±0.1	86.79±5.5	79.49±4.5	62.50±14.5
GPT3.5	67.69	93.42	74.40	73.50	37.71	73.56	62.26	69.23	65.78
GCN	89.11±0.2	85.33±1.0	75.64±0.2	71.82±0.3	45.14±0.6	63.41±0.6	46.98±9.3	44.36±5.6	54.21±7.9
GAT	88.24±0.9	88.75±0.9	79.45±0.6	73.95±0.1	43.55±0.4	64.53±0.5	44.91±9.9	46.67±12.3	51.32±9.9
SAGE	88.24±0.1	88.81±0.0	78.29±0.2	71.71±0.2	45.00±0.6	64.30±0.4	81.13±8.4	69.49±7.8	78.16±7.3
H2GCN	87.81±1.4	89.59±0.3	48.62±0.1	72.50±0.2	41.64±0.7	77.61±2.9	87.74±6.3	83.68±5.1	84.21±5.4
FAGCN	88.85±1.4	89.98±0.5	67.94±0.2	71.83±0.2	44.20±0.7	74.46±0.4	49.43±4.2	49.74±9.3	53.95±6.5
GIANT-BERT	85.52±0.7	85.02±0.5	74.06±0.4	74.26±0.2	52.31±0.6	72.18±0.3	86.38±3.6	75.74±3.6	77.87±2.8
GLEM-DeBERTa	85.60±0.1	94.71±0.2	73.77±0.1	74.69±0.3	50.18±0.8	78.58±0.1	87.14±4.1	84.76±4.6	86.76±5.2
ENGINE-LLAMA	91.48±0.3	95.24±0.4	80.05±0.5	76.02±0.3	54.60±0.9	79.76±0.1	85.50±4.0	77.36±4.5	75.68±5.0
TAPE-GCN	87.41±1.6	94.31±0.4	79.96±0.4	75.20±0.0	48.14±4.1	80.80±2.2	61.79±14.3	87.32±1.8	81.36±0.4
TAPE-RevGAT	92.80±2.8	96.04±0.5	79.76±0.1	77.50±0.1	47.22±1.0	79.95±0.6	87.77±7.0	88.46±5.3	85.90±3.3
TAPE-SAGE	92.90±3.1	94.80±0.4	81.37±0.4	76.72±0.0	46.39±2.4	80.23±0.3	88.89±3.4	87.18±4.7	85.26±6.4
SimTeG-e5	88.04±1.4	94.84±0.8	74.51±1.5	75.29±0.2	48.95±1.2	79.51±0.5	87.12±2.8	85.04±3.9	84.36±3.0
LEMP-TAPE	88.26±1.2	94.85±0.3	–	–	46.75±2.1	80.03±0.2	85.65±4.6	86.54±3.2	85.09±5.3
Ours-GCN	93.54±1.3	96.08±0.9	81.50±0.1	**78.15±0.1**	**66.40±0.4**	82.60±2.2	88.68±7.1	**88.64±6.0**	89.47±3.4
Ours-GAT	92.98±1.6	96.14±1.2	**82.24±0.4**	77.50±0.1	66.12±0.3	81.63±2.6	81.13±12.3	87.69±5.6	88.16±5.6
Ours-SAGE	**93.73±1.9**	**96.62±0.6**	81.73±0.5	78.13±0.2	66.28±0.5	**82.72±2.8**	**89.26±4.9**	88.21±6.0	**90.01±4.4**

Table 1: Mean node classification accuracy (%). The best and second-best results are in **bold** and underlined, respectively. – denotes unavailable results. H_N is the node-level homophily metric.

Methods	Wisconsin	Cornell	Texas
DGI	55.21±1.0	45.33±6.1	58.5±3.0
GRACE	47.11±3.5	41.48±3.8	53.55±3.5
Congrat	63.00±4.9	57.90±4.1	62.60±4.3
PolyGCL	70.24±4.3	62.45±5.5	66.77±5.5
HeterGCL	71.93±4.2	64.02±4.8	72.79±2.5
Ours-GCN	**72.83±7.1**	**64.10±4.4**	**73.68±3.4**

Table 2: Mean node classification accuracy (%) under the unsupervised setting.



(a) GCN (b) TAPE-GCN-E (c) GCL-OT-GCN

(d) GCN (e) TAPE-GCN-E (f) GCL-OT-GCN

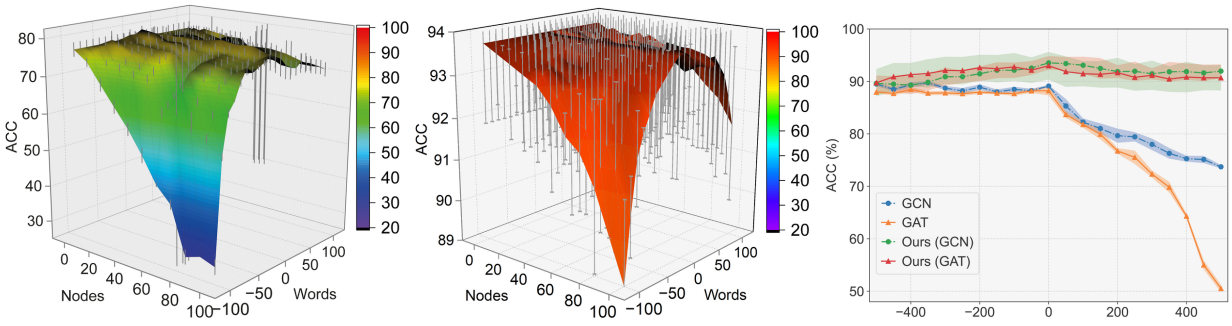
Figure 4: T-SNE visualization of node representations learned by different models on Cora (a–c) and Texas (d–f), with colors indicating ground-truth class labels.

shown in Table 2, GCL-OT outperforms classical and heterophilic GCL methods, demonstrating that GCL-OT can learn discriminative representations even without supervision

Visualization of Learned Representations. We further conduct t-SNE visualization on learned node embeddings. The results on Cora, Actor, and Texas datasets are shown in Figure 4. It can be observed that compared to baselines, GCL-OT produces more semantically coherent and better-separated clusters.

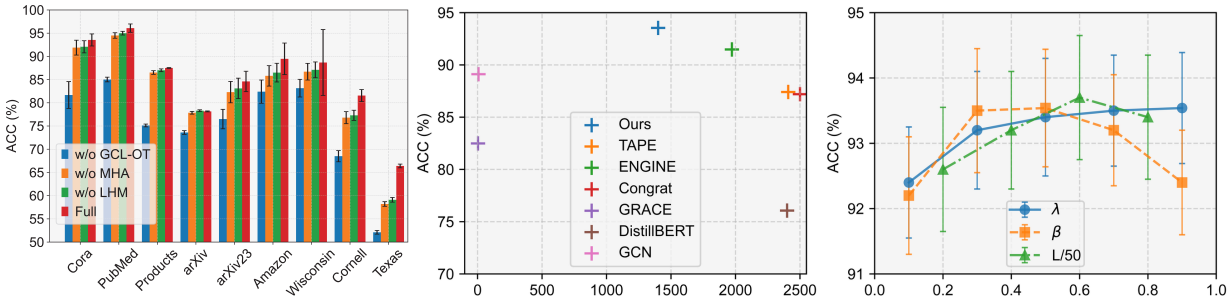
Robustness Analysis. We evaluate robustness under random edge and text perturbations. Edge perturbation ranges from -500 (removal) to +500 (addition) in increments of 50. Text perturbations range from 0% to 100% of nodes, with word perturbation varying from -100% (removal) to +100% (addition) at intervals of 10%. Results on Cora are shown in Figure 5. In the edge perturbation setting, vanilla GCN and GAT suffer sharp accuracy drops under edge deletions, falling approx 50% and 75% at the -500 worst condition. At the same time, GCL-OT with GCN and GCL-OT with GAT achieve relative gains of 24.74% and 79.44% respectively. In the text perturbation scenario, DistilBERT shows significant degradation, while our model consistently retains nearly 85%, exhibiting a smoother performance curve. These experiments highlight the robustness of our model against both structural and text perturbations.

Ablation Study. To evaluate the contribution of each component in GCL-OT, we conduct an ablation study by selectively contribution three main losses: contrastive loss $\mathcal{L}$GCL-OT, alignment loss $\mathcal{L}_{\text{MHA}}$, and latent homophily mining loss $\mathcal{L}_{\text{LHM}}$. Results in Figure 6a show that removing $\mathcal{L}_{\text{GCL-OT}}$ significantly degrades performance, highlight-



(a) DistilBERT on text perturbation (%) (b) GCL-OT on text perturbation (%) (c) Models on edge perturbation

Figure 5: Evaluation of model robustness under text and edge perturbations on Cora.



(a) Ablation study. (b) Training time (sec.) (c) Parameters

Figure 6: Analysis of ablation, computational cost, and hyperparameter sensitivity on Cora.

ing its importance for consistency and alignment of structure and text information. Excluding $\mathcal{L}_{\text{MHA}}$ results in significant drops in heterophilic datasets, such as Texas and Cornell. Removing $\mathcal{L}_{\text{LHM}}$ leads to poor performance in datasets like Amazon with weak structural signals. The complete model consistently achieves the best results, confirming the complementary strengths of all components.

Efficiency Comparison. The efficiency of GCL-OT is evaluated against baselines by plotting training time against ACC. As shown in Figure 6b, GCL-OT with GCN (blue) achieves the highest ACC while maintaining a relatively short training time, demonstrating a favorable trade-off between efficiency and performance. Although GCN and GRACE train faster, they fail to capture deep textual semantics, resulting in inferior performance. These results highlight the practicality of GCL-OT in real-world scenarios.

PLM Comparison. We further replace the default DistilBERT encoder with stronger PLMs. As shown in 3, using DeBERTa or RoBERTa yields comparable or slightly better accuracy across most datasets, indicating that GCL-OT can benefit from stronger encoders without relying on them.

Parameter Sensitivity Analysis. GCL-OT includes three key hyperparameters: Weight of contrastive loss λ, temperature of RSM β, and number of LRSinkhorn iterations. Results with GCN on Cora are shown in Figure 6c. The values of λ and β are varied within $\{0.0, 0.1, \cdots, 1.0\}$, and the number of LRSinkhorn iterations is tested over $\{10, 20, 30, 40\}$. The results indicate that suitable values of these parameters can enhance performance, and λ influences the stability, as reflected in the standard deviation, with only mild fluctuations beyond these values.

	Wisconsin	Cornell	Texas
DeBERTa	89.16 ± 4.7	87.35 ± 6.2	87.30 ± 11.8
RoBERTa	86.42 ± 6.6	88.21 ± 3.9	86.82 ± 14.6

Table 3: GCL-OT with different PLMs.

Conclusion

This paper addresses the multi-granular heterophily challenges in TAGs. We propose GCL-OT, a contrastive learning framework with OT with tailored mechanisms to hierarchically align and fuse textual and structural representations. Specifically, for partial heterophily, GCL-OT employs the RealSoftMax operator to identify important structure-text relations. For complete heterophily, GCL-OT employs a filter prompt to distinguish between alignable and non-alignable embeddings in the transport-based contrastive loss. For latent homophily, GCL-OT leverages OT assignments as extra supervision to capture hidden homophilic patterns. Theoretical analysis and extensive experiments in both homophilic and heterophilic settings can demonstrate the effectiveness and robustness of GCL-OT.

Acknowledgments

This paper is supported by the National Key R&D Program of China (Grant No. 2024YFC3308200).

References

Abu-El-Haija, S.; Perozzi, B.; Kapoor, A.; Alipourfard, N.; Lerman, K.; Harutyunyan, H.; Ver Steeg, G.; and Galstyan, A. 2019. Mixhop: Higher-order graph convolutional architectures via sparsified neighborhood mixing. In *Proc. of ICML*, 21–29.

Bo, D.; Wang, X.; Shi, C.; and Shen, H. 2021. Beyond low-frequency information in graph convolutional networks. In *Proc. of AAAI*, 3950–3957.

Brannon, W.; Kang, W.; Fulay, S.; Jiang, H.; Roy, B.; Roy, D.; and Kabbara, J. 2024. ConGraT: Self-Supervised Contrastive Pretraining for Joint Graph and Text Embeddings. In *Proc. of ACL Workshop on TextGraphs-17: Graph-based Methods for Natural Language Processing*, 19–39.

Chen, J.; Lei, R.; and Wei, Z. 2024. PolyGCL: GRAPH CONTRASTIVE LEARNING via Learnable Spectral Polynomial Filters. In *Proc. of ICLR*.

Chen, Y.; Guan, D.; Yuan, W.; and Zang, T. 2025. Beyond Homophily: Graph Contrastive Learning with Macro-Micro Message Passing. In *Proc. of AAAI*, 15948–15956.

Craven, M.; McCallum, A.; PiPasquo, D.; Mitchell, T.; and Freitag, D. 1998. Learning to extract symbolic knowledge from the World Wide Web. Technical report, Carnegie-mellon univ pittsburgh pa school of computer Science.

Deng, B.; Wang, T.; Fu, L.; Huang, S.; Chen, C.; and Zhang, T. 2025. THESAURUS: Contrastive Graph Clustering by Swapping Fused Gromov-Wasserstein Couplings. In *Proc. of AAAI*, 16199–16207.

Fang, Y.; Fan, D.; Zha, D.; and Tan, Q. 2024. GAugLLM: Improving Graph Contrastive Learning for Text-Attributed Graphs with Large Language Models. In *Proc. of KDD*, 747–758.

Gong, C.; Cheng, Y.; Li, X.; Shan, C.; Luo, S.; and Shi, C. 2024. Towards learning from graphs with heterophily: Progress and future. *arXiv preprint arXiv:2401.09769*.

Hamilton, W. L.; Ying, R.; and Leskovec, J. 2017. Inductive representation learning on large graphs. In *Proc. of NeurIPS*, 1025–1035.

He, X.; Bresson, X.; Laurent, T.; Perold, A.; LeCun, Y.; and Hooi, B. 2024. Harnessing Explanations: LLM-to-LM Interpreter for Enhanced Text-Attributed Graph Representation Learning. In *Proc. of ICLR*.

Hu, W.; Fey, M.; Zitnik, M.; Dong, Y.; Ren, H.; Liu, B.; Catasta, M.; and Leskovec, J. 2020. Open Graph Benchmark: Datasets for Machine Learning on Graphs. *arXiv preprint arXiv:2005.00687*.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *Proc. of ICLR*.

Leskovec, J.; Adamic, L. A.; and Huberman, B. A. 2007. The dynamics of viral marketing. *ACM TWEB*, 1(1): 5–es.

Li, S.; Wu, Y.; Shi, C.; and Fang, Y. 2025. HeTGB: A Comprehensive Benchmark for Heterophilic Text-Attributed Graphs. arXiv:2503.04822.

Liang, L.; Hu, X.; Xu, Z.; Song, Z.; and King, I. 2023. Predicting global label relationship matrix for graph neural networks under heterophily. In *Proc. of NeurIPS*, 10909–10921.

Lin, Y.; Zhang, J.; Huang, Z.; Liu, J.; Peng, X.; et al. 2023. Multi-granularity Correspondence Learning from Long-term Noisy Videos. In *Proc. of ICLR*.

Liu, J.; Yang, C.; Lu, Z.; Chen, J.; Li, Y.; Zhang, M.; Bai, T.; Fang, Y.; Sun, L.; Yu, P. S.; and Shi, C. 2025a. Graph Foundation Models: Concepts, Opportunities and Challenges. *IEEE TPAMI*, 1–23.

Liu, Y.; Giunchiglia, F.; Li, X.; Huang, L.; Feng, X.; and Guan, R. 2025b. Enhancing Unsupervised Graph Few-shot Learning via Set Functions and Optimal Transport. In *Proc. of KDD*, 871–882.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proc. of NeurIPS*.

Monge, G. 1781. Mémoire sur la théorie des déblais et des remblais. *Mémoires de l'Académie Royale des Sciences*, 666–704.

Pan, S.; Zheng, Y.; Liu, Y.; and Murugesan, S. 2024. Integrating Graphs With Large Language Models: Methods and Prospects. *IEEE Intelligent Systems*, 39(1): 64–68.

Pei, H.; Wei, B.; Chang, K. C.-C.; Lei, Y.; and Yang, B. 2020. Geom-GCN: Geometric Graph Convolutional Networks. In *Proc. of ICLR*.

Platonov, O.; Kuznedelev, D.; Diskin, M.; Babenko, A.; and Prokhorenkova, L. 2023. A critical look at the evaluation of GNNs under heterophily: Are we really making progress? In *Proc. of ICLR*.

SANGARE, A. S.; Dunou, N.; Giraldo, J. H.; and Malliaros, F. D. 2025. A Fused Gromov-Wasserstein Approach to Subgraph Contrastive Learning. *TMLR*.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *Proc. of NeurIPS Workshop on Energy Efficient Machine Learning and Cognitive Computing*.

Scetbon, M.; Cuturi, M.; and Peyré, G. 2021. Low-Rank Sinkhorn Factorization. In *Proc. of ICML*, 9344–9354.

Sen, P.; Namata, G.; Bilgic, M.; Getoor, L.; Galligher, B.; and Eliassi-Rad, T. 2008. Collective classification in network data. *AI magazine*, 29(3): 93–93.

Song, Y.; Zhou, C.; Wang, X.; and Lin, Z. 2023. Ordered GNN: Ordering Message Passing to Deal with Heterophily and Over-smoothing. In *Proc. of ICLR*.

Tang, J.; Yang, Y.; Wei, W.; Shi, L.; Su, L.; Cheng, S.; Yin, D.; and Huang, C. 2024. Graphgpt: Graph instruction tuning for large language models. In *Proc. of SIGIR*, 491–500.

Tsai, Y.-H. H.; Wu, Y.; Salakhutdinov, R.; and Morency, L.-P. 2021. Self-supervised Learning from a Multi-view Perspective. arXiv:2006.05576.

van den Oord, A.; Li, Y.; and Vinyals, O. 2019. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2018. Graph Attention Networks. In *Proc. of ICLR*.

Villani, C. 2009. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der Mathematischen Wissenschaften*. Springer Berlin Heidelberg. ISBN 978-3-540-71049-3, 978-3-540-71050-9.

Vincent-Cuaz, C.; Flamary, R.; Corneli, M.; Vayer, T.; and Courty, N. 2022. Template based Graph Neural Network with Optimal Transport Distances. In *Proc. of NeurIPS*, 11800–11814.

Wang, C.; Liu, Y.; Yang, Y.; and Li, W. 2024a. HeterGCL: graph contrastive learning framework on heterophilic graph. In *Proc. of IJCAI*, 2397–2405.

Wang, Q.; Pang, G.; Salehi, M.; Buntine, W.; and Leckie, C. 2023a. Cross-Domain Graph Anomaly Detection via Anomaly-Aware Contrastive Alignment. *Proc. of AAAI*, 4676–4684.

Wang, W.; and Cheng, D. 2025. Language Model-Enhanced Message Passing for Heterophilic Graph Learning. arXiv:2505.19762.

Wang, X.; Gao, H.; Wei, X.; Peng, L.; Li, R.; Liu, C.; Wu, S.; and Wong, H.-S. 2024b. Contrastive Graph Distribution Alignment for Partially View-Aligned Clustering. In *Proc. of MM*, 5240–5249.

Wang, Y.; Zhao, Y.; Wang, D. Z.; and Li, L. 2023b. GALOPA: Graph Transport Learning with Optimal Plan Alignment. In *Proc. of NeurIPS*, 9117–9130.

Wen, Z.; and Fang, Y. 2023. Augmenting low-resource text classification with graph-grounded pre-training and prompting. In *Proc. of SIGIR*, 506–516.

Wu, Y.; Li, S.; Fang, Y.; and Shi, C. 2025. Exploring the Potential of Large Language Models for Heterophilic Graphs. In *Proc. of NAACL*.

Xiao, T.; Chen, Z.; Guo, Z.; Zhuang, Z.; and Wang, S. 2022. Decoupled self-supervised learning for graphs. In *Proc. of NeurIPS*, 620–634.

Xie, S.; and Giraldo, J. H. 2024. Variational Graph Contrastive Learning. In *Proc. of NeurIPS Workshop on Self-Supervised Learning– Theory and Practice*.

Xu, H.; Liu, J.; Luo, D.; and Carin, L. 2023. Representing Graphs via Gromov-Wasserstein Factorization. *IEEE TPAMI*, 45(1): 999–1016.

Xu, K.; Li, C.; Tian, Y.; Sonobe, T.; Kawarabayashi, K.-i.; and Jegelka, S. 2018. Representation learning on graphs with jumping knowledge networks. In *Proc. of ICML*, 5453–5462.

Yan, H.; Li, C.; Long, R.; Yan, C.; Zhao, J.; Zhuang, W.; Yin, J.; Zhang, P.; Han, W.; Sun, H.; Deng, W.; Zhang, Q.; Sun, L.; Xie, X.; and Wang, S. 2023. A Comprehensive Study on Text-attributed Graphs: Benchmarking and Rethinking. In *Proc. of NeurIPS*, 17238–17264.

Yang, J.; Liu, Z.; Xiao, S.; Li, C.; Lian, D.; Agrawal, S.; S, A.; Sun, G.; and Xie, X. 2021. GraphFormers: GNN-nested Transformers for Representation Learning on Textual Graph. In *Proc. of NeurIPS*.

Zhang, P.; Li, C.; Kang, L.; Huang, F.; Wang, S.; Xie, X.; and Kim, S. 2024a. High-Frequency-aware Hierarchical Contrastive Selective Coding for Representation Learning on Text Attributed Graphs. In *Proc. of WWW*, 4316–4327.

Zhang, Q.; Zhang, L.; Song, R.; Cong, R.; Liu, Y.; and Zhang, W. 2024b. Learning Common Semantics via Optimal Transport for Contrastive Multi-View Clustering. *IEEE TIP*, 33: 4501–4515.

Zhang, Y.; Jin, R.; and Zhou, Z.-H. 2010. Understanding bag-of-words model: a statistical framework. *International journal of machine learning and cybernetics*, 1(1): 43–52.

Zhao, J.; Qu, M.; Li, C.; Yan, H.; Liu, Q.; Li, R.; Xie, X.; and Tang, J. 2023. Learning on Large-scale Text-attributed Graphs via Variational Inference. In *Proc. of ICLR*.

Zheng, X.; Wang, Y.; Liu, Y.; Li, M.; Zhang, M.; Jin, D.; Yu, P. S.; and Pan, S. 2022. Graph neural networks for graphs with heterophily: A survey. *arXiv preprint arXiv:2202.07082*.

Zhu, J.; Rossi, R. A.; Rao, A.; Mai, T.; Lipka, N.; Ahmed, N. K.; and Koutra, D. 2021. Graph neural networks with heterophily. In *Proc. of AAAI*, 11168–11176.

Zhu, J.; Yan, Y.; Zhao, L.; Heimann, M.; Akoglu, L.; and Koutra, D. 2020. Beyond homophily in graph neural networks: Current limitations and effective designs. In *Proc. of NeurIPS*, 7793–7804.

Zhu, Y.; Guo, J.; Wu, F.; and Tang, S. 2022. RoSA: A Robust Self-Aligned Framework for Node-Node Graph Contrastive Learning. In *Proc. of IJCAI*, 3795–3801.

Zhu, Y.; Wang, Y.; Shi, H.; and Tang, S. 2024. Efficient Tuning and Inference for Large Language Models on Textual Graphs. In *Proc. of IJCAI*, 5734–5742.