THGB: A Comprehensive Benchmark for Text-attributed Heterogeneous Graphs

Lixin Zhou[1], **Zemin Liu**[1*], **Yuan Fang**[2], **Dan Niu**[3], **Jing Ying**[1*]
[1]College of Computer Science and Technology, Zhejiang University, Hangzhou, China
[2]School of Computing and Information Systems, Singapore Management University, Singapore
[3]School of Automation, Southeast University, Nanjing, Jiangsu, China
{lixinzhou, liu.zemin, csyingj}@zju.edu.cn, yfang@smu.edu.sg, danniu1@163.com

Abstract

Text-attributed heterogeneous graphs (TAHGs), characterized by nodes interconnected through diverse relationships and enriched with textual descriptions, are prevalent in numerous real-world applications. Recent advancements in integrating pre-trained language models (PLMs) and large language models (LLMs) with heterogeneous graph neural networks (HGNNs) have enhanced learning on TAHGs. However, the absence of standardized benchmark datasets tailored to TAHGs has impeded further progress. To bridge this gap, we propose the **T**ext-attributed **H**eterogeneous **G**raphs **B**enchmark (**THGB**), a comprehensive collection of heterogeneous graphs from diverse domains, with each node enriched by relevant text attributes. Alongside dataset construction, we conduct extensive benchmark experiments using various graph learning methods, including GNN, PLM-GNN, and LLM-GNN approaches, for node classification and link prediction tasks. We evaluated model performance across supervised, few-shot, and zero-shot learning scenarios to assess their ability to leverage limited and unseen data. Our experiments highlight THGB's potential to improve the integration of heterogeneous structural and textual information. By providing curated datasets, robust evaluation protocols, and baseline implementations, THGB introduces a standardized benchmark and solid groundwork for TAHGs research.

Code & Datasets — https://github.com/xxxinxxx/THGB

1 Introduction

Heterogeneous graphs (HGs) are ubiquitous in various real-world scenarios (Yang et al. 2020; Wang et al. 2022). Their ability to effectively capture the intricate relationships among diverse node types has led to extensive investigation and application across multiple domains. Early heterogeneous graph embedding approaches (Dong, Chawla, and Swami 2017; Fu, Lee, and Lei 2017) typically sample paths in the graph, employing pre-defined patterns (*e.g.*, meta-paths (Sun et al. 2011)) for representation learning. Recently, the emergence of heterogeneous graph neural networks (HGNNs), which utilize neighborhood aggregation for message passing, has garnered significant attention due to their promising performance on various downstream tasks. Furthermore, the rise of heterogeneous Transformer approaches (Hu et al. 2020; Mao et al. 2023), which leverage the Transformer architecture (Min et al. 2022) for their encoders, has marked a significant advancement in this critical area.

Background. Despite the rapid development of heterogeneous representation learning approaches, their performance remains constrained by the limited information preserved in their datasets (Yang et al. 2020; Lv et al. 2021), which typically contain only categorical or shallow features derived from raw attributes (Hu et al. 2020; Dong, Chawla, and Swami 2017). This distillation process can lead to the loss of important information or even distort the original semantic content. Textual attributes can provide detailed and nuanced characteristics of the objects involved, significantly enhancing model performance (Jin et al. 2024; Mao et al. 2024). However, as a more comprehensive source of information, they have been largely overlooked.

The rise of large language models (LLMs) has revolutionized numerous domains, driven by their unparalleled ability to process and comprehend textual information (Zhao et al. 2023; Jin et al. 2024). In graph learning, LLMs improve representation and inference capabilities by integrating graph structures with text insights to create richer representations that capture both the relational and semantic content (Jin et al. 2024; Li et al. 2023; Mao et al. 2024). The recently proposed benchmarks (Yan et al. 2023; Li et al. 2024b) offer a comprehensive analysis by incorporating various graph learning paradigms, demonstrating the remarkable success of LLMs in homogeneous graph learning (Wei et al. 2022; Jin et al. 2024; Liu et al. 2023b). However, their application to heterogeneous graphs remains underexplored.

Present work. In response to these limitations, we introduce the THGB, a comprehensive benchmark consisting of four multiscale TAHGs sourced from diverse domains, including online movies, citation networks, and patent applications. Beyond type information, each node in the THGB datasets is enriched with detailed textual descriptions. The construction of the dataset follows a rigorous process that includes the careful selection of data sources, the development of the heterogeneous graph structure, and the extraction of text and category information (detailed in Sec. 4). Unlike existing datasets (Yan et al. 2023; Lv et al. 2021) that ei-

*Corresponding authors.

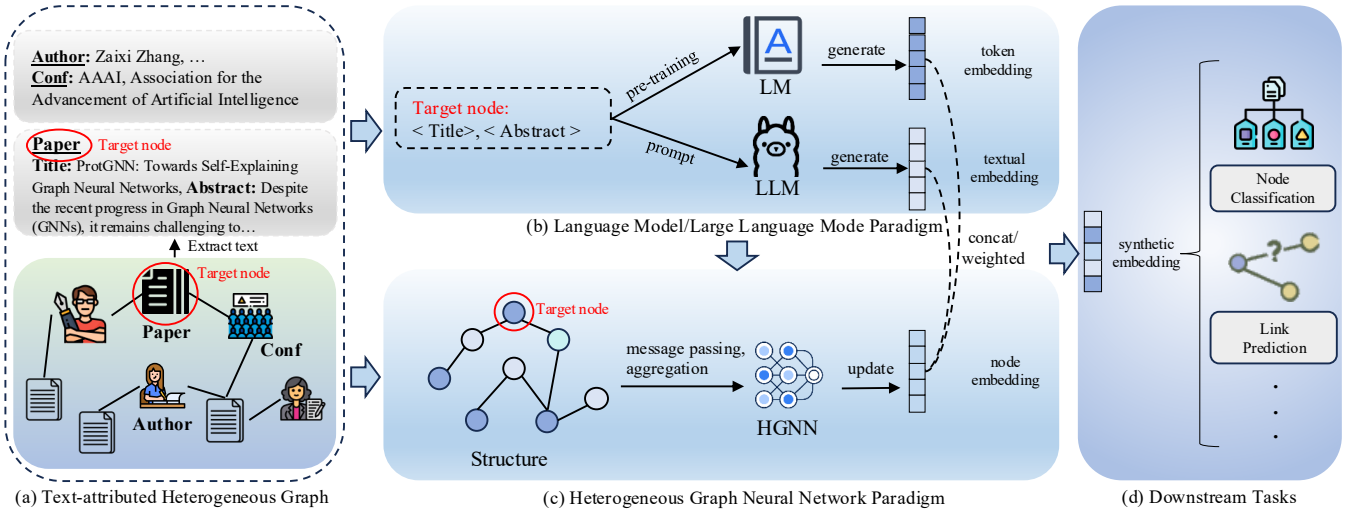

Figure 1: The pipeline for text-attributed heterogeneous graph representation learning comprises: (a) a graph with nodes and edges of multiple types, where specific nodes possess textual attributes; (b) a Language Model (LM/LLM) encodes the text into embeddings; (c) a Heterogeneous GNN encodes the graph structure, with outputs combined (e.g., via concatenation) into a synthetic embedding; (d) the final embedding is applied to downstream tasks like node classification and link prediction.

ther lack raw text or heterogeneity, THGB sets itself apart with its inclusion of rich textual attributes for different types of nodes and a multi-scale structure, ensuring a comprehensive and representative sample of real-world heterogeneous graph scenarios.

To establish a robust benchmark, we conduct extensive experiments on the proposed text-attributed heterogeneous graphs, focusing on node classification and link prediction tasks. We evaluated the THGB datasets using three main categories of representative graph learning approaches: ten GNN-based methods, three PLM-based methods, and three LLM-based models, as illustrated in Figure 1, emphasizing the importance of integrating textual and structural information in THGB. The comprehensive experimental analysis reveals the limitations and challenges of these methods, offering valuable insights for future research directions. Our contributions are summarized as follows:

- To the best of our knowledge, THGB is the first open benchmark specifically designed for text-attributed heterogeneous graphs. Datasets from four distinct domains and a unified structure provide a comprehensive foundation for evaluating model performance in this critical research area.
- We perform an in-depth evaluation of GNN-based, PLM-based, and LLM-based models for node classification and link prediction tasks, highlighting the main challenges, limitations, and insights in the learning of text-attributed heterogeneous graphs.
- We are releasing THGB, including datasets and baseline implementations, to support further research (*e.g.*, investigating LLM-based models on THGB) and to promote the development of effective models for text-attributed heterogeneous graphs.

2 Related Work

Heterogeneous graph representation learning. Heterogeneous graphs (HGs) have emerged as a powerful framework for modeling diverse and intricate relationships in real-world applications, including citation networks, recommendation systems, and knowledge graphs. Unlike homogeneous graphs, HGs incorporate multiple node and edge types, enabling more expressive representations of complex systems.

Early heterogeneous graph neural networks (HGNNs), such as RGCN (Schlichtkrull et al. 2018), extended graph convolutional networks (GCNs) to handle multi-relational data through relation-specific transformation matrices. HAN (Wang et al. 2019b) introduced hierarchical attention mechanisms to capture both node- and meta-path-level importance, enhancing representation learning in structurally diverse graphs. HGT (Hu et al. 2020) further advanced the field by leveraging type-specific attention mechanisms to dynamically model heterogeneous nodes and edges. Subsequent models addressed scalability and semantic representation. For instance, HetSANN (Hong et al. 2020) employed self-attention mechanisms for scalable processing of heterogeneous graphs, while MAGNN (Fu et al. 2020) utilized meta-path-based aggregation to capture high-order semantics. SimpleHGN (Ferrari Dacrema, Cremonesi, and Jannach 2019) simplified the design of heterogeneous attention mechanisms, achieving a balance between computational efficiency and strong performance. Despite these advancements, existing benchmarks predominantly rely on precomputed features such as bag-of-words (BoW) (Harris 1954) or TF-IDF (Salton and Buckley 1987), which often fail to capture the semantic richness of raw textual data. This limitation underscores the need for benchmarks that integrate rich, unstructured textual information to fully leverage the

potential of heterogeneous graph learning.

Existing benchmarks for TAGs. The integration of textual attributes into graph learning has led to the emergence of text-attributed graphs (TAGs) (Yan et al. 2023; Zhao et al. 2022; Yu et al. 2023). Benchmarks like TAGLAS (Feng et al. 2024) provide a wide range of TAG datasets, enabling standardized evaluation of models that integrate text and graph structures. However, TAGLAS primarily focuses on homogeneous graphs, lacking the diverse node and edge types essential for studying HGs. Similarly, TSGFM (Chen et al. 2024) explored text-space foundation models for graph learning, emphasizing zero- and few-shot learning scenarios. While it highlights the potential of LLMs to bridge textual and structural data, the datasets in TSGFM are restricted to single-node-type graphs with shallow text representations, limiting their applicability to heterogeneous scenarios. GLBench (Li et al. 2024b) evaluates large-scale datasets across various graph tasks, including text attributes, but does not address the challenges of heterogeneous graphs, such as managing diverse node types and raw textual data.

3 Preliminaries

Text-attributed heterogeneous graph. Text-attributed heterogeneous graphs (TAHGs) can be defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{U}, \mathcal{R}, \mathcal{T}, \mathbf{A}, \mathcal{X})$, where $\mathcal{V}$, $\mathcal{E}$, $\mathcal{U}$, and $\mathcal{R}$ represent the sets of nodes, edges, node types, and edge types, respectively. Each node $v \in \mathcal{V}$ has a node type $\phi(v) \in \mathcal{U}$, and each edge e_{v_i,v_j} has an edge type $\psi(e_{v_i,v_j}) \in \mathcal{R}$. $\mathcal{T}$ is the set of textual descriptions of nodes, and $\mathbf{A} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ represents the adjacency matrix of graph $\mathcal{G}$. The feature matrix set $\mathcal{X} = \{\mathbf{X}_u \in \mathbb{R}^{|\mathcal{V}_u|\times d_u} \mid u \in \mathcal{U}\}$, where $u \in \mathcal{U}$ refers to a specific node type, $\mathcal{V}_u \subseteq \mathcal{V}$ is the set of nodes of type u, and d_u is the feature dimension of nodes $\mathcal{V}_u$ with type u. Note that a TAHG should satisfy $|\mathcal{U}| + |\mathcal{R}| > 2$.

Graph neural networks. GNNs follow the paradigm of neighborhood aggregation, where messages are recursively passed from neighboring nodes to target nodes. Formally, in the k -th layer, this process operates as

$$\mathbf{h}_v^k = \mathcal{M}\left(\mathbf{h}_v^{k-1}, \left\{\mathbf{h}_i^{k-1} : i \in \mathcal{N}_v\right\}; \theta^k\right), \tag{1}$$

where $\mathbf{h}_v^k \in \mathbb{R}^{d_k}$ is the d_k-dimensional embedding vector of node v in the k-th layer, and $\mathcal{M}(\cdot)$, parameterized by θ^k in the k-th layer, is the message passing function for neighborhood aggregation. In the input layer, $\mathbf{h}_v^0$ is given by the input node features, *i.e.*, $\mathbf{h}_v^0 \equiv \mathbf{x}_v$.

Meta-Paths in heterogeneous graphs. Meta-paths (Sun and Han 2012; Sun et al. 2011) have been widely adopted for mining and learning in heterogeneous graphs. A meta-path P is defined as a sequence of nodes and edges that follows a specific pattern of node and edge types, *i.e.*, $P \triangleq v_1 \xrightarrow{r_1} v_2 \xrightarrow{r_2} \cdots \xrightarrow{r_l} v_{l+1}$, where $r_i \in \mathcal{R}$ and $v_i \in \mathcal{V}$. It serves as a structural template to capture relationships between different types of nodes in a heterogeneous graph, enabling the extraction of higher-order semantics and facilitating more effective representation learning. For instance, in a citation network, a common meta-path is "author $\leftrightarrow$ paper $\leftrightarrow$ author", which connects papers through their shared authorship, revealing deeper relational information that goes beyond direct paper-to-author or paper-to-conference connections.

4 Dataset Details

In this section, we first provide motivation for THGB, and then present a detailed account of the dataset construction process and its statistical characteristics.

Motivation of THGB. In many existing studies (Wang et al. 2019c; Ferrari Dacrema, Cremonesi, and Jannach 2019; Yan et al. 2023), we observe that the commonly used heterogeneous graphs are essentially text-attributed. At the same time, many popular text-attributed graphs inherently exhibit heterogeneity. For example, the well-known heterogeneous graph dataset ACM (Ferrari Dacrema, Cremonesi, and Jannach 2019) and the widely used TAG dataset ogbn-arxiv-TA (Yan et al. 2023) are both derived from the citation network (McLaren and Bruner 2022), which is intrinsically tied to text attributes such as the titles and abstracts of papers, as well as information about authors and conferences. As shown in Table 1, however, ACM contains only numerical attributes derived from the corresponding elements of a bag-of-words representation of keywords. The ogbn-arxiv-TA contains only paper nodes, which results in the text information not being well organized and utilized effectively, thereby restricting advancements in heterogeneous graph representation learning.

Although these datasets are commonly used by HGNNs, they exhibit significant limitations when it comes to exploring representation learning on TAHGs. First, the majority of these datasets rely on simplistic bag-of-words models based on keywords to represent text attributes, which may lead to information loss. Second, these approaches may result in misrepresentation, as relying solely on predefined keywords to generate feature vectors can exclude important terms, potentially resulting in vectors with the opposite meaning. Finally, most of these datasets lack access to raw textual information. Given the widespread use of LLMs, this lack of text data significantly hampers the deeper exploration of LLMs in heterogeneous graph learning.

Dataset Construction. To address these challenges, we propose THGB, a comprehensive benchmark tailored for text-attributed heterogeneous graphs, which serves as a standardized evaluation framework for assessing the efficacy of representation learning techniques on TAHGs. We carefully select four real-world datasets from diverse domains, including online movies, citation networks, and patent applications. The graph structures were constructed from original publicly available sources, and relevant textual information was assigned to each node. To ensure fair comparisons, each dataset was processed into a standardized format. The reconstructed datasets may differ in the number and types of nodes and edges compared to prior heterogeneous graph research. In order to clearly distinguish between them, we added suffixes to the dataset names based on their textual content, *e.g.*, "-TS", "-TA", *etc.*, as shown in Table 1. We introduce the construction details of each dataset as follows.

	Dataset	Nodes	Edges	Classes	Domain	Node Types	Features	Heterogeneity	Raw Text
Previous Heterogeneous Graphs	**IMDB** (Lv et al. 2021)	21,420	86,642	5	Movie	4	bag-of-words	✓	✗
	ACM (Wang et al. 2019c)	10,942	547,872	4	Academic	4	bag-of-words	✓	✗
	DBLP (Wang et al. 2019c)	26,128	239,566	3	Academic	4	bag-of-words	✓	✗
	Amazon-ratings (Platonov et al. 2023)	24,492	93,050	5	E-commerce	5	FastText	✓	✗
Previous TAGs	**ogbn-arxiv-TA** (Yan et al. 2023)	169,343	1,166,243	40	Academic	1	PLMs	✗	✓
	CitationV8 (Yan et al. 2023)	1,106,759	6,120,897	-	Academic	1	PLMs	✗	✓
	Books-Children (Yan et al. 2023)	76,875	1,554,578	24	E-commerce	1	PLMs	✗	✓
	Ele-Computers (Yan et al. 2023)	87,229	721,081	10	E-commerce	1	PLMs	✗	✓
THGB	**IMDB-TS**	13,090	36,570	5	Movie	3	PLMs	✓	✓
	ACM-TA	53,204	143,280	10	Academic	3	PLMs	✓	✓
	DBLP-TA	233,620	853,098	10	Academic	3	PLMs	✓	✓
	Patent-TA	792,920	2,224,518	9	Patent	3	PLMs	✓	✓

Table 1: Statistics of datasets and comparison with existing datasets.

- **IMDB-TS** is a movie network collected from an online movie website[1]. The nodes represent Movies, Actors, and Directors. Relationships are established based on actors acting in movies and movies directed by directors. **Movies** are treated as target nodes, with text descriptions extracted from their Titles and Storylines, while Actors and Directors are represented by their Names. The movie label is categorized into five classes: Action, Comedy, Drama, Romance, and Thriller.
- **ACM/DBLP-TA**[2] is a citation network that includes three types of nodes: Papers, Authors, and Conferences. Relationships are established based on authors writing papers and papers being published at conferences, with each paper linked to a single conference. **Papers** are treated as target nodes, and their text descriptions are derived from their Titles and Abstracts. Authors and conferences are represented by text descriptions composed of their Names or the respective conference Names. Papers are categorized into ten classes according to their research areas.
- **Patent-TA** is a novel patent network derived from the Harvard USPTO[3], comprising Patents, Examiners, and Inventors. Relationships are formed based on inventors filing patents and examiners reviewing them, introducing new node types that are not present in prior work. **Patents** are treated as target nodes, with text descriptions extracted from their Titles and Abstracts, while Examiners and Inventors are represented by their Names. Patents are classified into nine categories based on their research domains.

For more details on dataset construction–including node types, link types, and label definitions–please refer to Appendix A.1.

Statistical Information. Table 1 presents the statistics of previous heterogeneous graph datasets, existing text-attributed graphs, and our proposed THGB. Notably, the THGB datasets stand out for their incorporation of raw text attributes and heterogeneous structures, offering distinct advantages for graph learning. By incorporating original textual data, these datasets enable a deeper exploration of semantic relationships, thereby expanding the potential for improving model performance on tasks such as node classification and link prediction. Additionally, the heterogeneous nature of the datasets–spanning diverse domains such as online movie networks, citation networks, and patent data–supports a more comprehensive evaluation of models designed to handle various types of nodes and edges. This combination of features ensures that THGB serves as a robust benchmark for testing the capacity of models to integrate textual and structural information effectively.

[1]https://www.imdb.com

[2]https://www.aminer.cn/citation

[3]https://patentdataset.org

5 Formulation and Benchmarking Methods

In this section, we explore the formulation and benchmarking methods employed in our study, focusing on three key paradigms: GNN-based methods, PLM-based methods, and LLM-based methods. We provide more details of these approaches in Appendix A.2.

5.1 GNN-based Approaches

To validate the effectiveness of the THGB benchmark, we conduct comparisons with comprehensive SOTA baselines, encompassing both GNN-based methods and HGNN approaches.

Graph Neural Networks. Graph Neural Networks (GNNs) update node representations by iteratively exchanging messages with neighbors, capturing both local and global structural properties (Wu et al. 2020). To thoroughly evaluate the effectiveness of GNNs in the context of text-attributed heterogeneous graphs, we selected a set of representative models, including GCN (Kipf and Welling 2016), GAT (Veličković et al. 2017), GIN (Xu et al. 2018), and GraphSAGE (Hamilton, Ying, and Leskovec 2017). These models cover a wide range of graph learning techniques, allowing us to explore their strengths and limitations across different tasks.

Heterogeneous Graph Neural Networks. Heterogeneous Graph Neural Networks (HGNNs) are designed to handle graphs with multiple types of nodes and edges, capturing complex relationships and heterogeneous structures through specialized message-passing mechanisms (Fan 2022). The HGNNs can effectively model the interactions between diverse entities and their connections.

We selected representative models, *e.g.*, RGCN (Schlichtkrull et al. 2018), HAN (Wang et al. 2019c),

SimpleHGN (Ferrari Dacrema, Cremonesi, and Jannach 2019), HetGNN (Zhang et al. 2019), HGT (Hu et al. 2020), and MAGNN (Fu et al. 2020), which cover diverse approaches to modeling heterogeneity in graphs, providing a comprehensive evaluation across different tasks. Please see Appendix A.2.1 for more details.

5.2 PLM-based Approaches

PLM-based graph learning methods excel at integrating rich semantic information from large-scale textual data into node and edge representations, making them highly effective in tasks involving text-attributed graphs (Li et al. 2024a). Their strength lies in capturing deep semantic relationships, particularly in datasets rich with textual features (Wang et al. 2023; Han et al. 2021; He et al. 2019). However, PLMs can be resource-intensive and may struggle in scenarios where the text is sparse or when structural complexity dominates. While they offer clear advantages in semantically rich tasks, their performance can be limited by computational demands and the specific nature of the dataset. In our study, we selected representative PLM-based models, including G2P2 (Wen and Fang 2024), Heterformer (Jin et al. 2023), and THLM (Zou et al. 2023), to provide a comprehensive evaluation of their effectiveness. Please see Appendix A.2.2 for more details.

5.3 LLM-based Approaches

Large Language Models (LLMs) are advanced deep learning architectures that leverage vast textual data to generate human-like text and understand contextual nuances (Hadi et al. 2024; Kumar 2024). Their ability to capture complex semantic relationships makes them effective for various applications (Hadi et al. 2023). However, LLMs face challenges such as high computational costs and difficulties in fine-tuning for specific tasks, especially with limited or noisy data. We selected representative models, including TAPE (He et al. 2023), OFA (Liu et al. 2023a), and HiGPT (Tang et al. 2024), for a comprehensive analysis. Please see Appendix A.2.3 for more details.

6 Experiments

In this section, we conduct a comprehensive empirical evaluation of THGB across node classification and link prediction tasks, along with further analysis.

6.1 Implementation Settings

For heterogeneous graph learning models, we adopt the implementations of the DGL toolkit (Wang et al. 2019a). To incorporate textual information into GNN-based models, we utilize the off-the-shelf RoBERTa-Base (Liu et al. 2019) model to encode node texts for representation initialization, ensuring a controlled and reproducible comparison that isolates the contribution of TAHGs from potential gains of domain-adaptive pretraining.

We search for the learning rate within $\{1, 2, 5\} \times \{10^{-6}, 10^{-4}, 10^{-3}, 10^{-2}\}$ in all cases, and $\{0, 1, 2\} \times \{10^{-4}, 10^{-3}, 10^{-2}\}$ for the weight decay rate. For the GAT model, the number of heads is fixed at 8. In the GraphSAGE model, a GCN-based aggregator is used to propagate and aggregate neighbor information, normalizing and combining node and neighbor features during message passing. The large language model used in the methods discussed in this paper is Llama-3-8B-Instruct (Touvron et al. 2023). For the methods requiring meta-paths, the meta-paths used in benchmark datasets are shown in Appendix B. For the LM and LLM models, the open-source implementations are carried out using Hugging Face (Wolf 2019). More implementation details are provided in Appendix D. We run all models five times with different random seeds and report the average performance along with the standard deviation.

6.2 Node Classification

We evaluated model performance on node classification tasks using our THGB benchmark across three settings: few-shot, supervised, and zero-shot classification. The evaluation metrics used are ***Micro-F1***, ***Macro-F1***, and ***Accuracy***. Due to space limitations, we present the experimental results and analysis for the few-shot setting in the main text, while the results for the supervised and zero-shot settings are provided in Appendix C.7 and Appendix C.8, respectively. For few-shot learning, each dataset is randomly partitioned into training, validation, and test sets in a 1:1:8 ratio. The detailed results are shown in Table 2. We analyze the performance of each method group in the following section.

Analysis of GNN-based methods. Among GNNs, models like GAT and GraphSAGE excel due to their attention mechanisms and graph architecture design. While effective to some extent, these methods are often tailored to specific homogeneous graphs and struggle to handle situations where nodes and edges vary in type. In contrast, HGNN models, such as SimpleHGN and HGT, employ meta-path attention mechanisms and transformers to capture intricate dependencies, generally achieving higher F1 scores across four datasets. The superior performance of HGNN methods can be attributed to their ability to leverage heterogeneity, enabling more effective graph learning compared to traditional approaches that do not account for diverse node and edge types. This highlights the necessity and advantages of incorporating relational information in heterogeneous graph learning.

Analysis of PLM-based methods. In our benchmark, PLM-based methods generally outperform GNN-based approaches due to their effective integration of textual and structural information. Heterformer surpasses G2P2, which is limited to homogeneous graphs and struggles to integrate semantic information across diverse node types. However, compared to THLM, Heterformer has limitations. THLM captures higher-order information, enriching the model's understanding, while Heterformer lacks key abstract content, leading to reduced text richness and slightly weaker performance.

Analysis of LLM-based methods. Among the available results, LLM-based models, such as OFA and HiGPT—the latter specifically tailored for heterogeneous graphs—demonstrate even greater performance gains. These

Methods	IMDB-TS		ACM-TA		DBLP-TA		Patent-TA	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
GCN	0.6390±0.0074	0.5863±0.0125	0.8123±0.0115	0.6579±0.0267	0.7623±0.0091	0.5699±0.0192	0.5757±0.0277	0.3273±0.0227
GAT	0.6562±0.0079	0.6051±0.0067	0.8473±0.0130	0.6747±0.0314	0.7764±0.0312	0.6133±0.0582	0.6466±0.0053	0.3978±0.0036
GIN	0.6079±0.0205	0.5205±0.0320	0.8254±0.0550	0.6128±0.0864	0.7503±0.0277	0.6025±0.0522	0.6110±0.0580	0.3605±0.0489
GraphSAGE	0.6488±0.0122	0.5976±0.0191	0.8357±0.0322	0.7894±0.0399	0.7600±0.0511	0.5901±0.0896	0.6338±0.0151	0.3755±0.0185
RGCN	0.6581±0.0143	0.6096±0.0281	0.8419±0.0241	0.7458±0.0334	0.7853±0.0695	0.6751±0.0667	0.6424±0.0127	0.3823±0.0159
HAN	0.6642±0.0048	0.6209±0.0073	0.8489±0.0894	0.7951±0.0588	0.7854±0.0100	0.7193±0.0141	0.6629±0.0280	0.4163±0.0369
SimpleHGN	0.6877±0.0106	0.6568±0.0223	0.8592±0.0194	0.8191±0.0343	0.7989±0.0135	0.6845±0.0259	0.6847±0.0133	0.4198±0.0233
HetGNN	0.4800±0.0030	0.4516±0.0027	0.5575±0.0274	0.5389±0.0193	–	–	–	–
HGT	0.6788±0.0136	0.6548±0.0170	0.8566±0.0557	0.7927±0.0365	0.7869±0.0495	0.5355±0.0815	0.6553±0.0154	0.3862±0.0247
MAGNN	0.6799±0.0085	0.6387±0.0124	0.8378±0.0675	0.8091±0.0427	0.7635±0.0341	0.5634±0.0365	–	–
G2P2	–	–	0.8585±0.0047	0.8216±0.0064	0.8209±0.0195	0.7018±0.0115	–	–
Heterformer	–	–	0.8604±0.0181	0.8253±0.0101	0.8332±0.0158	0.7161±0.0172	–	–
THLM	0.7092±0.0039	0.6739±0.0074	0.8643±0.0045	0.8262±0.0028	0.8416±0.0105	0.7259±0.0116	**0.7080**±0.0157	**0.4337**±0.0173
TAPE	0.7501±0.0071	0.7423±0.0063	0.8919±0.0118	0.8272±0.0138	0.8783±0.0114	0.8343±0.0143	–	–
OFA	–	–	0.8993±0.0063	0.8336±0.0038	**0.8823**±0.0089	**0.8391**±0.0052	–	–
HiGPT	**0.7843**±0.0133	**0.7523**±0.0149	**0.9144**±0.0048	**0.8523**±0.0053	–	–	–	–

Table 2: Few-shot node classification comparison. Vacant positions ("-") indicate either that the models ran out of memory when applied to large graphs or that they are not applicable to multi-label data, including IMDB-TS and Patent-TA. The results are averaged across five runs, with the best and second-best results are highlighted in **bold** and underline.

models excel at capturing complex language patterns and semantic relationships, enabling them to process longer and more intricate textual descriptions more effectively than both GNNs and PLMs. Additionally, in most cases, LLM-based models not only achieve higher F1 scores but also exhibit generally lower standard deviations in performance metrics, indicating greater consistency and reliability across diverse datasets.

Overall, the results reveal a clear performance hierarchy: as models advance from GNNs to PLMs and then to LLMs, their ability to leverage textual information improves, leading to enhanced node classification performance. This trend emphasizes the importance of incorporating rich textual attributes and advanced language modeling techniques.

6.3 Link Prediction

We perform link prediction on two benchmark datasets, focusing on key relationships: movie-actor in IMDB-TS and paper-author in ACM-TA. We evaluated various GNN-based methods, as well as typical HGNNs such as RGCN, HetGNN, HGT, and SimpleHGN, along with representative PLM-based and LLM-based methods, including THLM and OFA. Using ***ROC-AUC*** and ***MRR*** as evaluation metrics, we formulate link prediction as a binary classification problem in THGB. The edges are split into training, validation, and test sets in a 5:1:4 ratio. The results, presented in Table 3, show that PLM-based and LLM-based methods significantly outperform GNN-based approaches in ROC-AUC scores. Additionally, GIN and HetGNN achieve strong performance under the MRR metric. In the following sections, we analyze the reasons behind these findings.

First, GIN and HetGNN achieved the highest MRR scores on the IMDB-TS and ACM-TA datasets, respectively, due to their ability to capture local node distinctions, which enhance node representation in link prediction. Second, advanced models like HGT and RGCN struggle to outperform simpler models like GCN and GAT, suggesting that added complexity doesn't always yield better results. Third, LLM-based methods generally outperform traditional GNNs in ROC-AUC scores by leveraging richer semantic representations from textual node content. While GNNs effectively leverage structural dependencies, their limited capacity to fully exploit textual information constrains their overall performance. Notably, LLM-based methods were not evaluated using MRR due to differences in data output interfaces compared to GNN-based methods, making fair comparisons challenging in the available open-source implementations, as "N.A." is shown in Table 3. This limitation will be addressed in future work to ensure a unified evaluation framework.

Methods	IMDB-TS		ACM-TA	
	ROC-AUC	MRR	ROC-AUC	MRR
GCN	0.7763±0.0061	0.8914±0.0086	0.7190±0.0233	0.8844±0.0036
GAT	0.7166±0.0112	0.8763±0.0042	0.7018±0.0969	0.8868±0.0318
GIN	0.7849±0.0090	**0.9065**±0.0129	0.8086±0.0517	0.9214±0.0223
GraphSAGE	0.7365±0.0069	0.8773±0.0038	0.5900±0.0851	0.8524±0.0207
RGCN	0.6757±0.0022	0.8343±0.0020	0.4767±0.1308	0.7043±0.0964
HetGNN	0.7392±0.0000	0.8751±0.0004	0.9353±0.0014	**0.9694**±0.0019
HGT	0.5373±0.0020	0.7523±0.0022	0.4502±0.0438	0.6934±0.0223
SimpleHGN	0.8053±0.0421	0.8974±0.0313	0.7627±0.1280	0.9017±0.0662
THLM	**0.8895**±0.0062	N.A.	0.9196±0.0163	N.A.
OFA	–	N.A.	**0.9696**±0.0078	N.A.

Table 3: Link prediction. "-" indicates that the models are not applicable for multi-label data. The results are averaged across five runs, with the best result highlighted in **bold**.

6.4 Further Analysis

To evaluate the effectiveness of the proposed benchmark, we conduct a comprehensive analysis encompassing an ablation

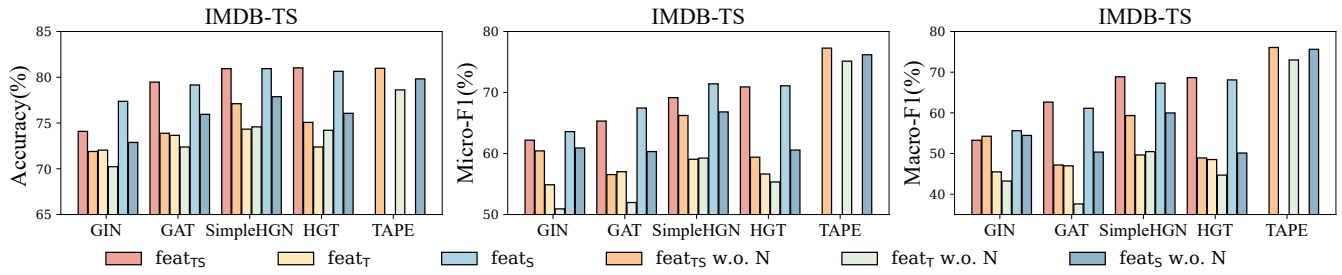

Figure 2: Ablation experiments of different text-attributed components

study, a scalability study, a convergence study, the impact of different LLMs and features, time and memory consumption, a case study, and a discussion on mitigating potential information leakage, *etc.* Due to space limitations, we present the first two analyzes in the main paper and provide the remaining results in Appendix C.

Ablation Study. To evaluate the contribution of each text-attributed component, we perform an ablation study by comparing several feature variants in few-shot node classification (using IMDB-TS as an example): (1) feat_{TS}: embeddings generated from the target node's title and storylines, integrated with other non-target node types (all features combined); (2) feat_{TS} w.o. N: using only the target node's title and storyline features; (3) feat_{T}: embeddings generated solely from the target node's title, combined with features from other non-target node types'; (4) feat_{T} w.o. N: using only the target node's title features; (5) feat_{S}: embeddings from the target node's storylines combined with non-target node types' features; and (6) feat_{S} w.o. N: using only the target node's storyline features.

The results of the ablation study, shown in Figure 2, lead to the following observations. First, excluding features from non-target node types significantly degrades performance, highlighting the importance of incorporating all available features. Second, using only titles or storylines causes a performance drop, especially when only title information is used, indicating the impact of information richness. Third, integrating textual information from other node types improves performance, demonstrating the benefits of additional contextual data. Although the TAPE model focuses only on target node features, our findings suggest that the lack of information negatively affects performance.

Additionally, we observe that PLM-based and LLM-based methods (such as THLM and TAPE) generally converge more quickly than GNN-based approaches. This demonstrates that effectively leveraging textual information can accelerate model training and further underscores the importance of text in THGB.

Scalability Study. On the larger DBLP-TA dataset, we sampled five subgraphs ranging from 10k to 60k nodes. In Figure 3, we report the training time (per epoch) for GAT, SimpleHGN, G2P2, and OFA across these subgraphs in node classification. Both training times increase linearly with the number of nodes, demonstrating the scalability of these models to large graphs in real-world scenarios. Notably, G2P2 and OFA demonstrate relatively slower inference times compared to GAT and SimpleHGN, likely due to the additional complexity introduced by their language model components. This suggests a trade-off between computational efficiency and the richer node representations achieved through LLM-based methods.

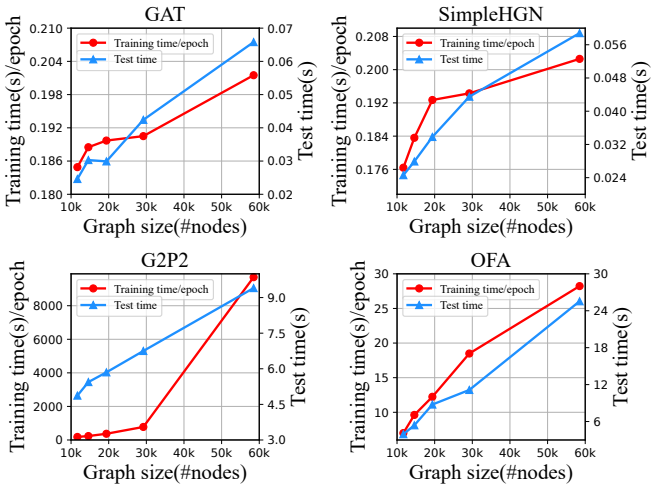

Figure 3: Scalability experiments with different graph sizes

7 Conclusions

In this study, we propose THGB, the first open benchmark specifically designed for text-attributed heterogeneous graph learning tasks. We introduce four diverse datasets sourced from distinct domains, establishing a comprehensive and robust foundation for evaluating model performance and generalization capabilities. Extensive experiments conducted with state-of-the-art heterogeneous graph learning methods demonstrate that integrating rich textual information substantially boosts the performance of (H)GNN models. These findings underscore the significant potential of leveraging textual descriptions to enhance representation learning and predictive accuracy in heterogeneous graph structures. Future work will focus on enriching THGB by addressing textual imbalance across node types and increasing structural heterogeneity to improve generalizability.

Acknowledgements

This paper is supported by the National Natural Science Foundation of China (62374031) and by NSFC-Jiangsu Province (BK20240173).

References

Chen, Z.; Mao, H.; Liu, J.; Song, Y.; Li, B.; Jin, W.; Fatemi, B.; Tsitsulin, A.; Perozzi, B.; Liu, H.; et al. 2024. Text-space Graph Foundation Models: Comprehensive Benchmarks and New Insights. *arXiv preprint arXiv:2406.10727*.

Dong, Y.; Chawla, N. V.; and Swami, A. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 135–144.

Fan, Y. 2022. *Heterogeneous Graph Representation Learning: Techniques and Applications*. Ph.D. thesis, Case Western Reserve University.

Feng, J.; Liu, H.; Kong, L.; Zhu, M.; Chen, Y.; and Zhang, M. 2024. TAGLAS: An atlas of text-attributed graph datasets in the era of large graph and language models. *arXiv preprint arXiv:2406.14683*.

Ferrari Dacrema, M.; Cremonesi, P.; and Jannach, D. 2019. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM conference on recommender systems*, 101–109.

Fu, T.-y.; Lee, W.-C.; and Lei, Z. 2017. Hin2vec: Explore meta-paths in heterogeneous information networks for representation learning. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 1797–1806.

Fu, X.; Zhang, J.; Meng, Z.; and King, I. 2020. Magnn: Metapath aggregated graph neural network for heterogeneous graph embedding. In *Proceedings of the web conference 2020*, 2331–2341.

Hadi, M. U.; Al Tashi, Q.; Shah, A.; Qureshi, R.; Muneer, A.; Irfan, M.; Zafar, A.; Shaikh, M. B.; Akhtar, N.; Wu, J.; et al. 2024. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*.

Hadi, M. U.; Qureshi, R.; Shah, A.; Irfan, M.; Zafar, A.; Shaikh, M. B.; Akhtar, N.; Wu, J.; Mirjalili, S.; et al. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*.

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Han, X.; Zhang, Z.; Ding, N.; Gu, Y.; Liu, X.; Huo, Y.; Qiu, J.; Yao, Y.; Zhang, A.; Zhang, L.; et al. 2021. Pre-trained models: Past, present and future. *AI Open*, 2: 225–250.

Harris, Z. S. 1954. Distributional structure.

He, B.; Zhou, D.; Xiao, J.; Liu, Q.; Yuan, N. J.; Xu, T.; et al. 2019. Integrating graph contextualized knowledge into pre-trained language models. *arXiv preprint arXiv:1912.00147*.

He, X.; Bresson, X.; Laurent, T.; Perold, A.; LeCun, Y.; and Hooi, B. 2023. Harnessing explanations: Llm-to-lm interpreter for enhanced text-attributed graph representation learning. *arXiv preprint arXiv:2305.19523*.

Hong, H.; Guo, H.; Lin, Y.; Yang, X.; Li, Z.; and Ye, J. 2020. An attention-based graph neural network for heterogeneous structural learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 4132–4139.

Hu, Z.; Dong, Y.; Wang, K.; and Sun, Y. 2020. Heterogeneous graph transformer. In *Proceedings of the web conference 2020*, 2704–2710.

Jin, B.; Liu, G.; Han, C.; Jiang, M.; Ji, H.; and Han, J. 2024. Large language models on graphs: A comprehensive survey. *IEEE Transactions on Knowledge and Data Engineering*.

Jin, B.; Zhang, Y.; Zhu, Q.; and Han, J. 2023. Heterformer: Transformer-based deep node representation learning on heterogeneous text-rich networks. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, 1020–1031.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kumar, P. 2024. Large language models (LLMs): survey, technical frameworks, and future challenges. *Artificial Intelligence Review*, 57(10): 260.

Li, X.; Wu, Z.; Wu, J.; Cui, H.; Jia, J.; Li, R.-H.; and Wang, G. 2024a. Graph Learning in the Era of LLMs: A Survey from the Perspective of Data, Models, and Tasks. *arXiv preprint arXiv:2412.12456*.

Li, Y.; Li, Z.; Wang, P.; Li, J.; Sun, X.; Cheng, H.; and Yu, J. X. 2023. A survey of graph meets large language model: Progress and future directions. *arXiv preprint arXiv:2311.12399*.

Li, Y.; Wang, P.; Zhu, X.; Chen, A.; Jiang, H.; Cai, D.; Chan, V. W. K.; and Li, J. 2024b. Glbench: A comprehensive benchmark for graph with large language models. *arXiv preprint arXiv:2407.07457*.

Liu, H.; Feng, J.; Kong, L.; Liang, N.; Tao, D.; Chen, Y.; and Zhang, M. 2023a. One for all: Towards training one graph model for all classification tasks. *The Twelfth International Conference on Learning Representations*.

Liu, J.; Yang, C.; Lu, Z.; Chen, J.; Li, Y.; Zhang, M.; Bai, T.; Fang, Y.; Sun, L.; Yu, P. S.; et al. 2023b. Towards graph foundation models: A survey and beyond. *arXiv preprint arXiv:2310.11829*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lv, Q.; Ding, M.; Liu, Q.; Chen, Y.; Feng, W.; He, S.; Zhou, C.; Jiang, J.; Dong, Y.; and Tang, J. 2021. Are we really making much progress? revisiting, benchmarking and refining heterogeneous graph neural networks. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 1150–1160.

Mao, Q.; Liu, Z.; Liu, C.; Li, Z.; and Sun, J. 2024. Advancing Graph Representation Learning with Large Language Models: A Comprehensive Survey of Techniques. *arXiv preprint arXiv:2402.05952*.

Mao, Q.; Liu, Z.; Liu, C.; and Sun, J. 2023. Hinormer: Representation learning on heterogeneous information networks with graph transformer. In *Proceedings of the ACM Web Conference 2023*, 599–610.

McLaren, C. D.; and Bruner, M. W. 2022. Citation network analysis. *International Review of Sport and Exercise Psychology*, 15(1): 179–198.

Min, E.; Chen, R.; Bian, Y.; Xu, T.; Zhao, K.; Huang, W.; Zhao, P.; Huang, J.; Ananiadou, S.; and Rong, Y. 2022. Transformer for graphs: An overview from architecture perspective. *arXiv preprint arXiv:2202.08455*.

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? *arXiv preprint arXiv:2302.11640*.

Salton, G.; and Buckley, C. 1987. Term weighting approaches in automatic text retrieval. Technical report, Cornell University.

Schlichtkrull, M.; Kipf, T. N.; Bloem, P.; Van Den Berg, R.; Titov, I.; and Welling, M. 2018. Modeling relational data with graph convolutional networks. In *The semantic web: 15th international conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, proceedings 15*, 593–607. Springer.

Sun, Y.; and Han, J. 2012. *Mining heterogeneous information networks: principles and methodologies*. Morgan & Claypool Publishers.

Sun, Y.; Han, J.; Yan, X.; Yu, P. S.; and Wu, T. 2011. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment*, 4(11): 992–1003.

Tang, J.; Yang, Y.; Wei, W.; Shi, L.; Xia, L.; Yin, D.; and Huang, C. 2024. Higpt: Heterogeneous graph language model. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2842–2853.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Wang, H.; Li, J.; Wu, H.; Hovy, E.; and Sun, Y. 2023. Pretrained language models and their applications. *Engineering*, 25: 51–65.

Wang, M.; Zheng, D.; Ye, Z.; Gan, Q.; Li, M.; Song, X.; Zhou, J.; Ma, C.; Yu, L.; Gai, Y.; et al. 2019a. Deep graph library: A graph-centric, highly-performant package for graph neural networks. *arXiv preprint arXiv:1909.01315*.

Wang, X.; Bo, D.; Shi, C.; Fan, S.; Ye, Y.; and Philip, S. Y. 2022. A survey on heterogeneous graph embedding: methods, techniques, applications and sources. *IEEE Transactions on Big Data*, 9(2): 415–436.

Wang, X.; Ji, H.; Shi, C.; Wang, B.; Ye, Y.; Cui, P.; and Yu, P. S. 2019b. Heterogeneous graph attention network. In *The world wide web conference*, 2022–2032.

Wang, X.; Ji, H.; Shi, C.; Wang, B.; Ye, Y.; Cui, P.; and Yu, P. S. 2019c. Heterogeneous graph attention network. In *The world wide web conference*, 2022–2032.

Wei, S.; Wang, J.; Zhao, Y.; Chen, X.; Li, Q.; Zhuang, F.; Liu, J.; Ren, F.; and Kou, G. 2022. Graph learning and its advancements on large language models: A holistic survey. *arXiv preprint arXiv:2212.08966*.

Wen, Z.; and Fang, Y. 2024. Prompt tuning on graph-augmented low-resource text classification. *IEEE Transactions on Knowledge and Data Engineering*.

Wolf, T. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; and Philip, S. Y. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1): 4–24.

Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2018. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*.

Yan, H.; Li, C.; Long, R.; Yan, C.; Zhao, J.; Zhuang, W.; Yin, J.; Zhang, P.; Han, W.; Sun, H.; et al. 2023. A comprehensive study on text-attributed graphs: Benchmarking and rethinking. *Advances in Neural Information Processing Systems*, 36: 17238–17264.

Yang, C.; Xiao, Y.; Zhang, Y.; Sun, Y.; and Han, J. 2020. Heterogeneous network representation learning: A unified framework with survey and benchmark. *IEEE Transactions on Knowledge and Data Engineering*, 34(10): 4854–4873.

Yu, J.; Ren, Y.; Gong, C.; Tan, J.; Li, X.; and Zhang, X. 2023. Empower text-attributed graphs learning with large language models (llms). *arXiv preprint arXiv:2310.09872*.

Zhang, C.; Song, D.; Huang, C.; Swami, A.; and Chawla, N. V. 2019. Heterogeneous graph neural network. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 793–803.

Zhao, J.; Qu, M.; Li, C.; Yan, H.; Liu, Q.; Li, R.; Xie, X.; and Tang, J. 2022. Learning on large-scale text-attributed graphs via variational inference. *arXiv preprint arXiv:2210.14709*.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Zou, T.; Yu, L.; Huang, Y.; Sun, L.; and Du, B. 2023. Pretraining language models with text-attributed heterogeneous graphs. *arXiv preprint arXiv:2310.12580*.