MUG: Meta-path-aware Universal Heterogeneous Graph Pre-Training

Lianze Shan[1†]**, Jitao Zhao**[1†]**, Dongxiao He**[1*]**, Yongqi Huang**[1]**, Zhiyong Feng**[1]**, Weixiong Zhang**[2]

[1]College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Departments of Health Technology & Informatics; Data Science & Artificial Intelligence; and Computing,
The Hong Kong Polytechnic University, Kowloon, Hong Kong
{shanlz2119, zjtao, hedongxiao, yqhuang, zyfeng}@tju.edu.cn, weixiong.zhang@polyu.edu.hk

Abstract

Universal graph pre-training has emerged as a key paradigm in graph representation learning, offering a promising way to train encoders to learn transferable representations from unlabeled graphs and to effectively generalize across a wide range of downstream tasks. However, recent explorations in universal graph pre-training primarily focus on homogeneous graphs and it remains unexplored for heterogeneous graphs, which exhibit greater structural and semantic complexity. This heterogeneity makes it highly challenging to train a universal encoder for diverse heterogeneous graphs: (i) the diverse types with dataset-specific semantics hinder the construction of a unified representation space; (ii) the number and semantics of meta-paths vary across datasets, making encoding and aggregation patterns learned from one dataset difficult to apply to others. To address these challenges, we propose a novel Meta-path-aware Universal heterogeneous Graph pre-training (MUG) approach. Specifically, for challenge (i), MUG introduces a input unification module that integrates information from multiple node and relation types within each heterogeneous graph into a unified representation. This representation is then projected into a shared space by a dimension-aware encoder, enabling alignment across graphs with diverse schemas. Furthermore, for challenge (ii), MUG trains a shared encoder to capture consistent structural patterns across diverse meta-path views rather than relying on dataset-specific aggregation strategies, while a global objective encourages discriminability and reduces dataset-specific biases. Extensive experiments demonstrate the effectiveness of MUG on some real datasets.

Code — https://github.com/slz1024/MUG

Introduction

Universal Graph Pre-training (UGP), which aims to build strong, highly generalizable models applicable to a wide range of downstream applications, has attracted considerable interest lately (Qiu et al. 2020; Zhu et al. 2024). Unlike traditional pre-training approaches that are tailored to specific tasks or graphs (Zhuo et al. 2024a; He et al. 2023; Zhao et al. 2024a; Zhuo et al. 2024b), UGP aims to learn graph

---

representations that capture shared structural and semantic patterns, facilitating effective transfer of learned representations across tasks and datasets without extensive re-training. The transition from building individual models for specific tasks to designing reusable encoders has advanced the development of graph foundation models (Wang et al. 2024; Zhu et al. 2025; Huang et al. 2025).

Despite these advances, most existing UGP methods (Zhao et al. 2024b; Yu et al. 2024, 2025) focus on homogeneous graphs that involve a single node type and relatively simple, fixed relations. In fact, graphs in real world are typically heterogeneous (Wang et al. 2019), such as user–item interactions for recommendation (He et al. 2020), knowledge graphs(Li et al. 2021b), and social networks (Jia et al. 2025), which contain multiple node and edge types and complex semantics. This heterogeneity carries richer schema-specific signals and relational patterns that cannot be effectively modeled by homogeneous graph approaches. Outside the UGP paradigm, in the classical graph heterogeneous learning field, many non-universal Heterogeneous Graph Representation Learning methods (HGRLs) have been proposed to capture rich semantics in heterogeneous graphs (Wang et al. 2022). These HGRLs, whether supervised or self-supervised, have effectively modeled semantic complexity and demonstrated strong performance (Wang et al. 2019, 2021; Tian et al. 2023; Mo et al. 2024).

However, the pursuit of a universal modeling framework has not aligned well with the ubiquitous heterogeneous graphs. To the best of our knowledge, no method has been developed so far for universal heterogeneous graph pre-training. This significant gap is mainly attributed to two factors. The first is that existing UGP methods assume each dataset contains a fixed and consistent set of node and relation types, enabling unified input spaces and reusable encoders. Recent works, such as FUG (Zhao et al. 2024b) and SAMGPT (Yu et al. 2025), can achieve attribute and structure transfer across graphs, respectively. Nevertheless, both of them rely on each dataset having a fixed set of entity and relation types. In heterogeneous graphs, however, the entity and relation types within each graph vary significantly, making these methods ineffective. The second factor is the limited generalizability of traditional HGRLs. These methods are typically designed with dataset-specific assumptions, including type-specific encoders, relation-aware ag-

gregation schemes, and meta-path semantics tailored to particular graph schemas. As a result, they are tightly coupled to the training data and fail to generalize to unseen heterogeneous graphs

The above observation naturally raises a question: What are the key issues that must be addressed to enable effective universal pre-training on heterogeneous graphs? We identify two primary challenges. **First,** each heterogeneous graph inherently contains multiple types of nodes and relations, making it difficult to unify. Different heterogeneous graphs may contain completely different node types, relation types, and node attributes. For example, the ACM dataset (Zhao et al. 2020) involves papers, authors, and subjects, connected via paper–author and paper–subject relations, while the Freebase dataset (Li et al. 2021a) includes movies, actors, directors, and writers, with relations such as movie–actor, movie–director, and movie–writer. These differences are not only in the number of types but also in their semantics and relational patterns. As a result, it is difficult to design a universal model to directly encode multiple heterogeneous graphs without first addressing the mismatch in node types, relation types, and node attributes. **Second,** the information learned from heterogeneous graphs is hard to transfer. This stems from two core components of graph representation learning: encoding and aggregation. Existing HGRLs typically require the design of distinct encoding functions to encode different views of a node, and rely on task-specific or data-specific objective functions to learn aggregation function tailored to the given graph. As a result, the learned representations and aggregation patterns (such as attention weights or transformation matrices), tend to be tightly coupled with the characteristics of the training dataset, making it difficult to generalize or transfer effectively to unseen heterogeneous graphs.

To address these challenges, we propose a novel **M**eta-path-aware **U**niversal heterogeneous **G**raph pre-training approach (MUG). To address the first challenge, we propose a input unification module, which aims to encode heterogeneous types and attributes into a unified input representation and builds a shared representation space for different heterogeneous graphs. Specifically, we generate a new embedding for each node by encoding its contextual structural information. This enables the model to capture high-order semantic relationships across heterogeneous graphs, such as co-authorship in ACM and co-acting in Freebase. We then get a unified input representation by concatenating this contextual embedding with the original node attributes, and align the representation space by introducing a dimension-aware encoder. For the second challenge, we start from a key observation: in homogeneous graphs, transferable encoding and aggregation strategies have been successfully developed based on homophily assumption. To explore whether a similar property exists in heterogeneous graphs, we conduct a motivation experiment as shown in Figure 1. It can be observed that the average homophophily ratio across meta-paths is comparable to that of the homogeneous graphs. This suggests that a shared encoder can be used across meta-path views. And we can design an objective function , which is based on the homophily assumption, to guide the encoder

to capture the common connectivity patterns of neighboring nodes under different views. Furthermore, we introduce a global scattering objective function during aggregation. It encourages globally discriminative embeddings to guide aggregation instead of relying on dataset-specific prior knowledge, reducing overfitting to training graphs.
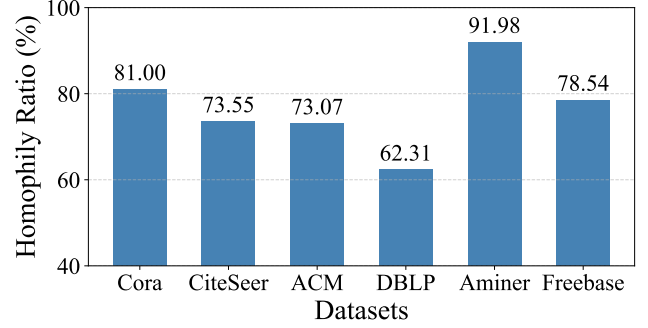

Figure 1: Homophily ratios on six datasets. Cora and CiteSeer (Yang, Cohen, and Salakhudinov 2016) are homogeneous graphs, while ACM (Zhao et al. 2020), DBLP (Fu et al. 2020), AMiner (Hu, Fang, and Shi 2019) and Freebase (Li et al. 2021a) are heterogeneous. For heterogeneous graphs, we compute the homophily ratio for each meta-path-based adjacency matrix and report the average.

Our contributions can be summarized as follows:

- To the best of our knowledge, this is the first work to explore LLM-free universal pre-training approach on heterogeneous graphs. It enables a model pre-trained on one heterogeneous graph and then to be generalized to unseen datasets without any retraining, greatly extending the applicability of graph foundation models to heterogeneous application domains.

- We propose MUG, a novel universal heterogeneous graph pre-training approach. MUG constructs a unified representation space and learns transferable information through universal encoding and aggregation .

- We conduct extensive experiments on four datasets under cross-domain and few-shot settings. MUG consistently outperforms all baseline methods, demonstrating its strong generalization ability and effectiveness across diverse scenarios.

Related Work

Heterogeneous Graph Representation Learning. This paradigm aims to encode the rich structural and semantic heterogeneity of graphs into low-dimensional embeddings that are suitable for downstream tasks. Early works in HGRLs focus on supervised or semi-supervised heterogeneous graph neural networks, which encode nodes and relations through meta-path-based or type-specific parameters. For example, HAN (Wang et al. 2019) employs a hierarchical attention mechanism to model the importance of different neighbors and their corresponding semantic relations.

MAGNN (Fu et al. 2020) captures semantics by jointly modeling node attributes and intermediate nodes along meta-paths, followed by aggregation across multiple meta-paths. HGT (Hu et al. 2020) extends the transformer architecture to heterogeneous graphs by utilizing type-specific parameters and attention, thereby enabling effective large-scale modeling. Despite their effectiveness, these methods rely heavily on labeled data, which limits their application to problems with data with a limited number of labels. To overcome this limitation, heterogeneous graph self-supervised learning has attracted increasing interest. It exploits the intrinsic structural and semantic information of heterogeneous graphs and designs pretext tasks to train heterogeneous graph neural network encoders without any labels. For example, HeCo (Wang et al. 2021) constructs dual views based on the network schema and meta-paths and performs cross-view contrastive learning to capture both local and high-order semantics. HGCL (Huo et al. 2025) further improves contrastive learning by introducing attribute and topology views, aligning them through a reciprocal contrastive mechanism to improve robustness. HGMAE (Tian et al. 2023) employs a masked autoencoder paradigm, training the encoder to reconstruct structural, attribute, and positional information. However, these methods trained on one dataset struggle to generalize to unseen datasets, which hinders the development of graph foundation models.

Universal Graph Pre-training. This paradigm has emerged as a promising direction to learn transferable representations across diverse graphs without relying on task-specific or data-specific supervision (Sun et al. 2024). Existing methods typically unify input spaces through structural encodings (Qiu et al. 2020; Yu et al. 2025), attribute projection (Zhao et al. 2024b; Yu et al. 2024), or attribute textualization (Liu et al. 2023). While these methods reduce distribution shifts and improve generalization, they are primarily designed for homogeneous graphs with uniform node and edge types and thus struggle to capture the rich semantics and relational diversity inherent in heterogeneous graphs. **Notably**, HiGPT (Tang et al. 2024) investigates cross-domain transferable representation learning for heterogeneous graphs by incorporating Large Language Models (LLM). However, its reliance on the textualization of node attributes poses challenges for generalizing to broader heterogeneous graphs with sparse or non-textual attributes. Our work instead focuses on developing a LLM-free universal pre-training approach for heterogeneous graphs, enabling broader applicability across diverse real-world scenarios.

Preliminary

We now introduce the notations and preliminaries adopted in the paper. We use script letters (e.g., $\mathcal{V}, \mathcal{E}$) to represent sets, bold uppercase letters (e.g., $\mathbf{H}$) to indicate matrices, bold lowercase letters (e.g., $\mathbf{h}_i$) to represent vectors (typically rows or columns from a matrix), and plain lowercase letters (e.g., h) to denote scalar values. All graphs, node attributes, meta-paths, and model outputs adhere to this unified notation, unless otherwise specified.

Definition 3.1. Heterogeneous Graph. A heterogeneous graph is defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{X}, \mathcal{A}, \mathcal{R})$, where $\mathcal{V}$ is the set

of nodes, $\mathcal{E}$ is the set of edges, $\mathcal{X}$ is the set of node attribute matrices, $\mathcal{A}$ and $\mathcal{R}$ are finite sets of node types and edge (relation) types respectively. Every node $v \in \mathcal{V}$ is associated with a type of assignment via a mapping $\phi_v : \mathcal{V} \to \mathcal{A}$, and every edge $e \in \mathcal{E}$ is associated with a relation type $\phi_e : \mathcal{E} \to \mathcal{R}$. The graph is considered *heterogeneous* when $|\mathcal{A}| + |\mathcal{R}| > 2$, indicating multiple node or edge types. A dataset-specific graph $\mathcal{G}^{(i)}$ follows this unified formulation but can be instantiated with different types of sets $\mathcal{A}^{(i)}$ and $\mathcal{R}^{(i)}$.

Definition 3.2. Pretraining and Evaluation Datasets. We denote a heterogeneous graph pre-training dataset as $\mathcal{D}_{\text{train}} = \{\mathcal{G}^{(1)}, \mathcal{G}^{(2)}, \ldots, \mathcal{G}^{(t)}\}$, where each $\mathcal{G}^{(i)}$ represents an individual heterogeneous graph instance. A universal encoder with parameters θ $f_\theta : \mathcal{V} \to \mathbb{R}^k$ is trained on $\mathcal{D}_{\text{train}}$, where k is the embedding dimension. During downstream evaluation, the learned encoder is transferred without any parameter updates to a set of unseen heterogeneous graphs $\mathcal{D}_{\text{test}} = \{\mathcal{G}'^{(1)}, \mathcal{G}'^{(2)}, \ldots, \mathcal{G}'^{(t')}\}$. The encoder f_θ is fixed, and its output representations are used directly for downstream tasks such as node classification.

Definition 3.3. Meta-path. A meta-path $\mathcal{P}$ is defined as a sequence of node and relation types in the form of $\mathcal{P} : A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \cdots \xrightarrow{R_l} A_{l+1}$, where $A_i \in \mathcal{A}$ and $R_i \in \mathcal{R}$. Each meta-path represents a composite semantic relation connecting nodes of type A_1 and A_{l+1} through a sequence of relations. For instance, a meta-path Author-Paper-Venue reflects a two-hop semantic dependency from authors to venues. In our setting, meta-paths serve as structural and semantic templates that guide the modeling of contextual information across graphs. Although different datasets may define distinct meta-paths, many of them reflect transferable structural or semantic patterns across heterogeneous graphs. Our model is designed to extract such patterns and apply them across heterogeneous graphs.

Method

In this section, we present MUG, a novel universal pre-training approach designed for heterogeneous graphs. As shown in Figure 2, MUG integrates two core components: a module that unifies variety of heterogeneous input by encoding context structural information and aligning diverse heterogeneous representation spaces; and a module that enables transferable heterogeneous information including encoding and aggregation. We discuss each component in the following subsections.

Heterogeneous Input Unification

Contextual structural encoding. As defined in preliminary, nodes and edges in each heterogeneous graph may be associated with multiple types, leading to diversity in node types (e.g., papers, authors, conferences), relation types (e.g., writes, published-in), and node attributes. A typical heterogeneous graph neural network encodes these rich semantics by defining type-specific transformation matrices and encoders (Wang et al. 2021; Tian et al. 2023). For a node v with type ϕ_i, a type-specific transformation matrix $\mathbf{W}_{\phi_i}^v$ is applied to project its original attributes $\mathbf{x}_v$ into a
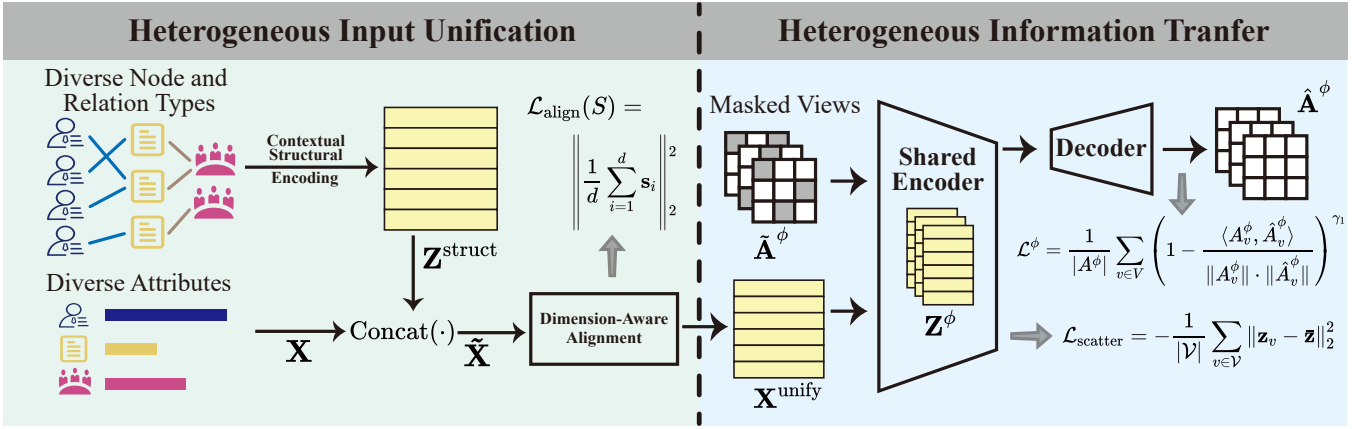
Figure 2: Overview of the MUG. Given a heterogeneous graph with diverse node types, relation types, and attributes, we first get embedding $\mathbf{Z}^{\text{struct}}$ by contextual structural encoding. The embedding is concatenated with the original node attributes $\mathbf{X}$ to obtain unified representation $\tilde{\mathbf{X}}$, which contains diverse type and attribute information. Then, a dimension-aware encoder is applied to align representation spaces and produce the unified input $\mathbf{X}^{\text{unify}}$. Finally, a shared GNN encoder is used to encode each masked adjacency matrix $\tilde{\mathbf{A}}^\phi$ with the unified input $\mathbf{X}^{\text{unify}}$, producing $\mathbf{Z}^\phi$. The model is optimized by three objectives: a dimension alignment loss $\mathcal{L}_{\text{align}}$, a meta-path masked reconstruction loss $\mathcal{L}^\phi$ and a global scattering regularization loss $\mathcal{L}_{\text{scatter}}$.

shared space:

$$\mathbf{h}_i = \sigma(\mathbf{W}_{\phi_i}\mathbf{x}_v + \mathbf{b}_{\phi_i}), \tag{1}$$

where the number and semantics of the parameters set $\{\mathbf{W}_{\phi_i}, \mathbf{b}_{\phi_i}\}$ are tightly coupled with the node type ϕ_i, making them inherently dependent on the schema of the given heterogeneous graph. Similarly, each relation type r is associated with a dedicated encoder $\text{Enc}_r(\cdot)$:

$$\mathbf{z}_i^{(r)} = \text{Enc}_r\left(\left\{h_j \mid j \in \mathcal{N}_i^{(r)}\right\}\right). \tag{2}$$

However, node and relation types may vary significantly in both number and semantics across datasets, making the transformation matrices and encoders defined on one specific dataset difficult to transfer to another.

To enable unified input across heterogeneous graphs, we construct contextual structural embeddings independent of specific node and relation types. Instead of encoding type identities directly, we embed each node through the semantic structures formed by its interactions with other node types, captured via meta-paths. Formally, given a heterogeneous graph $\mathcal{G}$ and a set of meta-paths $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_L\}$, we define the structural embedding of node v as:

$$\mathbf{z}_v^{\text{struct}} = \mathcal{F}_{\text{context}}(v, \mathcal{G}, \mathcal{P}), \tag{3}$$

where $\mathcal{F}_{\text{context}}(\cdot)$ denotes a general structural embedding function that can be instantiated by methods capable of capturing high-order semantic dependencies across different node and relation types. Inspired by (Dong, Chawla, and Swami 2017), we instantiate $\mathcal{F}_{\text{context}}$ using a meta-path guided context encoder, which samples node sequences along meta-paths and optimizes node embeddings to predict their structural context. Specifically, for each meta-path $\mathcal{P}_\ell \in \mathcal{P}$, we perform meta-path guided random walks to

sample structural contexts for each node v:

$$\mathcal{C}_v^{(\ell)} = \left\{(v_0 = v, v_1, \ldots, v_K) \left| \begin{array}{c} (v_i, v_{i+1}) \in \mathcal{E} \\ (t_i, r_{i+1}, t_{i+1}) \in \mathcal{P}_\ell \end{array} \right. \right\}, \tag{4}$$

where $i = 0, 1, \ldots, K-1$, and $\mathcal{C}_v^{(\ell)}$ is the set of node sequences starting from v and following the meta-path $\mathcal{P}_\ell$, K is the walk length, t_i and r_{i+1} denote the node and relation types along the path, respectively. To learn unified structural embeddings across heterogeneous node and relation types, we employ skip-gram with negative sampling (Dong, Chawla, and Swami 2017), which brings together nodes that co-occur in meta-path contexts into a shared representation space. And the objective can be defined as:

$$\mathcal{L}_{\text{struct}} = -\frac{1}{L}\sum_{\ell=1}^{L} \mathbb{E}_{(v,u)\sim\mathcal{C}_v^{(\ell)}} \left[\log\sigma\left(\mathbf{z}_v^{\text{struct}\top}\mathbf{z}_u^{\text{struct}}\right)\right]$$
$$- \mathbb{E}_{(v,\tilde{u})\sim P_n} \left[\log\left(1 - \sigma\left(\mathbf{z}_v^{\text{struct}\top}\mathbf{z}_{\tilde{u}}^{\text{struct}}\right)\right)\right], \tag{5}$$

where P_n is a negative sampling distribution (e.g., uniform over all nodes), and all embeddings $\mathbf{z}_v^{\text{struct}}$ are learned jointly across all meta-paths. This loss encourages the embeddings to capture heterogeneous, high-order structural contexts in the graph. After obtaining the trained context structural embedding $\mathbf{z}_v^{\text{struct}}$, we freeze its parameters and concatenate the embedding with the original node attributes to unify the input of heterogeneous graph:

$$\tilde{\mathbf{x}}_v = \text{concat}(\mathbf{x}_v, \mathbf{z}_v^{\text{struct}}), \tag{6}$$

where $\mathbf{x}_v$ denotes the original attributes of node v, and $\text{concat}(\cdot)$ denotes column-wise concatenation that appends the structural embedding as additional dimensions. This contextual structural encoding strategy does not require

type-specific transformation matrices or encoders for nodes and relations. And we can encode both node and relation types into embedding $\mathbf{z}_v^{\text{struct}}$.

Dimension-aware alignment. While the contextual structural encoding provides unified node representation $\tilde{\mathbf{x}}_v$ for a single heterogeneous graph, its representation space can vary significantly across different graphs, which poses challenges for universal processing. To address this, we introduce a dimension-aware alignment module, inspired by the semantic basis learning strategy of (Zhao et al. 2024b). We treat each attribute dimension as an independent semantic unit and learn a basis vector that reflects its latent global meaning. For each input graph, we randomly sample n_s nodes, and for the i-th attribute dimension, we aggregate its values as a column vector $\tilde{\mathbf{X}}_{:,i}^s \in \mathbb{R}^{n_s \times 1}$. We then apply a simple MultiLayer Perceptron (MLP) to encode it into a semantic basis vector $\mathbf{s}_i \in \mathbb{R}^k$:

$$\mathbf{s}_i = \text{MLP}(\tilde{\mathbf{X}}_{:,i}^s) = (\tilde{\mathbf{X}}_{:,i}^s)^{\text{T}} \mathbf{W} + \mathbf{b}, \qquad (7)$$

where $\mathbf{W} \in \mathbb{R}^{n_s \times k}$ are learnable parameters, $\mathbf{s}_i \in \mathbb{R}^{1 \times k}$ is the basis vector for dimension i. Each node attribute $\tilde{\mathbf{x}}_v$ is then projected into a unified space $\mathbb{R}^k$ by aggregating the basis vectors weighted by the dimension values:

$$\mathbf{x}_v^{\text{unify}} = \sum_{i=1}^{d} \tilde{\mathbf{x}}_v[i] \cdot \mathbf{s}_i. \qquad (8)$$

We train the dimension encoder with a mean-centering loss, which encourages the mean of all basis vectors to stay close to the origin:

$$\mathcal{L}_{\text{align}}(S) = \left\| \frac{1}{d} \sum_{i=1}^{d} \mathbf{s}_i \right\|_2^2. \qquad (9)$$

In this way ,we can prevent the learned basis vectors from local bias and ensures that all attribute dimensions are aligned in a consistent way.

Heterogeneous Information Transfer

Building on the unified input, it remains crucial to overcome the limited transferability, which is mainly reflected in two aspects: encoding and aggregation. First, in the encoding stage, existing methods typically require to design a distinct encoding function for the adjacency matrix generated by each meta-path, which ties the learned representations to a particular set of meta-paths and makes it difficult to transfer. Second, in the aggregation stage, these approaches further depend on learning meta-path-specific attention weights to combine the representations from different encoding functions, which further limits transferability.

Universal encoding. To enable universal encoding for heterogeneous graphs, we investigate the nature of different meta-path views and find that most exhibit a high degree of homophily (i.e., nodes of the similar semantics tend to connect). The high degree of homophily makes it possible to use a single shared encoder $\text{GNN}_{\text{shared}}$ to encode the topological views generated by different meta-paths. Based on this, we need to design a objective function, which can guide

the $\text{GNN}_{\text{shared}}$ to capture the common connectivity patterns of neighboring nodes under different meta-path views. Inspired by the success of masked auto-encoder methods in effectively learning structural information (Hou et al. 2022; Tian et al. 2023), we apply a mask-and-reconstruct strategy for each meta-path view. Specifically, for each meta-path based adjacency matrix $\mathbf{A}^\phi$, we construct a binary mask $\mathbf{M}_A^\phi \sim \text{Bernoulli}(p_e)$, where p_e is the edge masking rate. The embedding of each views can be encoded as follows:

$$\mathbf{Z}^\phi = \text{GNN}_{\text{shared}}(\tilde{\mathbf{A}}^\phi, \mathbf{X}^{\text{unify}}), \tilde{\mathbf{A}}^\phi = \mathbf{M}_A^\phi \odot \mathbf{A}^\phi. \qquad (10)$$

We then apply a GNN decoder to obtain the decoded node embeddings $\hat{\mathbf{Z}}^\phi$ and get the reconstructed adjacency matrix by $\hat{\mathbf{A}}^\phi = \sigma(\hat{\mathbf{Z}}^\phi \hat{\mathbf{Z}}^{\phi^T})$. Then, a scaled cosine loss is used to train the model:

$$\mathcal{L}^\phi = \frac{1}{|A^\phi|} \sum_{v \in V} \left(1 - \frac{\langle A_v^\phi, \hat{A}_v^\phi \rangle}{\|A_v^\phi\| \cdot \|\hat{A}_v^\phi\|} \right)^{\gamma_1}, \qquad (11)$$

where γ_1 is a scaling parameter.

Universal aggregation. After obtaining the encoding loss $\mathcal{L}^\phi$ from different meta-paths, we need to design a universal aggregation mechanism. In conventional heterogeneous graph pre-training, a semantic-level attention vector $\mathbf{q}$ is introduced to learn attention weights β^ϕ for each meta-path:

$$c^\phi = \mathbf{q}^\top \cdot \tanh(\mathbf{W} \cdot \mathbf{Z}^\phi + \mathbf{b}), \beta^\phi = \frac{\exp(c^\phi)}{\sum_{\phi \in \Phi} \exp(c^\phi)}, \qquad (12)$$

where $\mathbf{W}$ is the weight matrix, $\mathbf{b}$ is the bias vector, c^ϕ is the contribution of meta-path ϕ. However, such aggregation strategies are often tightly coupled to the training dataset and fail to generalize to new graphs or tasks. This is because the training objective leverages dataset-specific priors to shape the embedding distribution in the representation space, such as pulling semantically similar nodes closer or pushing dissimilar ones apart. As a result, the model tends to learn aggregation weights that are tailored to the semantic patterns of specific datasets, thereby limiting generalizability. To address this issue, we introduce an additional global scattering regularization loss inspired by (He et al. 2024):

$$\mathcal{L}_{\text{scatter}} = -\frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \|\mathbf{z}_v - \bar{\mathbf{z}}\|_2^2, \qquad (13)$$

where $\bar{\mathbf{z}} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbf{z}_v$. Rather than guiding aggregation via dataset-specific signals, it encourages node embeddings to be pushed away from the global mean in the representation space, thereby increasing their discriminability. This regularization alleviates the reliance on specific aggregation functions and promotes more transferable node representations across heterogeneous graphs. And we compute the final loss by the following equation:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{align}} + \lambda_2 \sum_{\phi \in \Phi} \beta^\phi \mathcal{L}^\phi + \lambda_3 \mathcal{L}_{\text{scatter}}, \qquad (14)$$

where $\lambda_1, \lambda_2, \lambda_3$ are hyper-parmters to adjust the weight of loss. We optimize all trainable parameters through this loss.

Train	Method	ACM		DBLP		AMiner		Freebase	
		Ma-F1	Mi-F1	Ma-F1	Mi-F1	Ma-F1	Mi-F1	Ma-F1	Mi-F1
ACM	HeCo	80.22±2.45	79.71±3.23	76.76±0.44	77.97±0.41	24.48±1.31	51.18±6.21	31.22±1.33	40.67±0.76
	HGMAE	84.22±0.52	84.01±0.50	87.17±0.19	88.23±0.23	29.08±0.83	41.91±7.06	32.59±1.12	42.95±1.31
	HERO	84.37±0.29	84.12±0.30	84.60±0.61	85.80±0.54	44.08±1.38	50.14±1.31	33.69±1.77	43.32±0.43
	MUG	**85.52±0.79**	**84.90±1.13**	**91.69±0.13**	**92.38±0.30**	**76.35±0.04**	**87.02±0.15**	**46.05±0.52**	**49.78±1.29**
DBLP	HeCo	82.93±0.80	83.43±0.70	90.11±0.33	90.73±0.31	26.89±1.09	35.23±5.45	35.58±0.94	39.53±2.11
	HGMAE	83.99±0.47	83.44±0.82	89.97±0.30	90.89±0.27	33.95±0.37	45.42±4.30	33.81±0.97	41.16±2.00
	HERO	84.92±0.46	84.68±0.48	87.47±0.36	88.57±0.34	51.66±1.30	61.04±2.93	32.63±2.62	42.99±0.65
	MUG	**85.81±0.50**	**85.06±0.54**	**90.67±0.28**	**91.40±0.33**	**73.44±1.05**	**84.96±0.42**	**50.48±0.97**	**55.22±0.89**
AMiner	HeCo	78.01±3.06	77.70±3.06	80.79±0.38	82.20±0.39	24.69±0.48	44.39±8.62	35.93±1.33	43.54±1.50
	HGMAE	83.67±0.51	83.48±0.49	87.66±0.64	88.82±0.56	27.83±1.08	48.71±6.31	35.92±0.60	41.52±1.81
	HERO	84.12±0.23	83.89±0.17	88.34±0.47	89.27±0.44	54.37±1.33	63.21±1.45	33.39±2.17	42.46±0.66
	MUG	**85.34±0.14**	**84.94±0.10**	**91.81±0.15**	**92.82±0.16**	**75.08±0.24**	**85.56±0.33**	**47.61±0.54**	**55.28±0.78**
Freebase	HeCo	77.03±0.89	76.30±1.44	82.37±0.47	83.26±0.47	29.82±0.82	34.51±4.34	42.34±1.68	47.92±2.32
	HGMAE	84.78±0.54	84.90±0.51	83.97±0.94	84.65±1.25	24.16±0.94	47.26±6.44	33.17±0.84	41.07±1.75
	HERO	84.65±0.36	84.35±0.37	84.29±0.57	85.84±0.62	48.26±1.13	58.19±1.17	31.25±1.73	42.14±0.98
	MUG	**85.21±1.26**	**85.22±1.05**	**91.79±0.28**	**92.24±0.17**	**78.10±1.35**	**87.94±0.56**	**52.33±0.26**	**57.50±2.41**

Table 1: Performance on cross-domain node classification with standard deviations. The best results are highlighted in bold.

Experiment

To comprehensively evaluate the effectiveness of the proposed MUG, we conducted extensive experiments. In the following subsections, we first introduce the experimental setup. Subsequently, we give the performance analysis on cross-domain and few-shot node classification tasks and model analysis to understand the contribution of each component of the model.

Experimental Setup

Datasets. We follow the prior work (Wang et al. 2021) and evaluate the performance of MUG on four widely-used heterogeneous datasets: ACM (Zhao et al. 2020), DBLP (Fu et al. 2020), AMiner (Hu, Fang, and Shi 2019) and Freebase (Li et al. 2021a). Each dataset contains a target node type (in bold), and our downstream node classification task is conducted specifically on nodes of this type.

- ACM contains three types of nodes: **papers**, authors, and subjects. It includes two types of relations (paper–author and paper–subject), and two meta-paths: PAP and PSP.
- DBLP contains four types of nodes: **authors**, papers, conferences, and terms. It includes three types of relations (author–paper, paper–conference, and paper–term), and three meta-paths: APA, APCPA, and APTPA.
- AMiner contains three types of nodes: **papers**, authors, and references. It includes two types of relations (paper–author and paper–reference), and two meta-paths: PAP and PRP.
- Freebase contains four types of nodes: **movies**, actors, directors, and writers. It includes three types of relations (movie–actor, movie–director, and movie–writer), and three meta-paths: MAM, MDM, and MWM.

Baselines. As there is no existing work that focuses on universal pre-training for heterogeneous graphs, we compare MUG with several well-established self-supervised learning methods for heterogeneous graph representation. Specifically, we include HeCo (Wang et al. 2021), a contrastive learning method that jointly captures local and high-order semantics from schema and meta-path views; HG-MAE, a generative method that trains the model by reconstructing masked node attributes, meta-paths, and positional attributes; and HERO (Mo et al. 2024), a recent method that jointly captures homophily and heterogeneity without relying on pre-defined meta-paths.

Implementation Details. To comprehensively evaluate the performance of MUG, we conducted experiments in two general application scenarios, i.e., cross-domain node classification and cross-domain few-shot node classification. For the cross-domain node classification, we train the model on one dataset and evaluate it on all datasets. During the downstream tasks, all model parameters are frozen. The training, validation, and test splits follow previous work (Wang et al. 2021), with 60 nodes per class for training and 1,000 nodes each for validation and testing. To further evaluate performance with extremely limited labels, we conduct few-shot experiments with 1-shot, 3-shot, and 5-shot settings, where each training class label in ths test dataset is provided with only 1, 3, or 5, respectively.

Performance Analysis

Cross-Domain node classification. In this scenario, we follow the evaluation protocol in (Wang et al. 2021), reporting the test performance when validation set achieves the best result. All experiments are repeated 50 times, and the mean and standard deviation are reported. We adopt both Macro-F1 and Micro-F1 scores as evaluation metrics. As shown in Table 1, MUG consistently outperforms all baseline models across all datasets, demonstrating its superior cross-domain transfer capability and robust generalization. For all three baseline methods, we apply SVD(Stewart 1993) to unify the input space across datasets. However, this approach inevitably leads to the loss of rich semantic information present in heterogeneous data. In contrast, our

Shot-num	Method	ACM		DBLP		Aminer		Freebase	
		Ma-F1	Mi-F1	Ma-F1	Mi-F1	Ma-F1	Mi-F1	Ma-F1	Mi-F1
1-shot	HeCo	59.36±12.32	63.41±9.52	45.74±10.32	49.94±9.53	23.85±2.61	39.69±10.43	30.70±3.08	39.58±3.68
	HGMAE	73.17±5.04	75.48±4.96	61.46±5.68	65.94±5.54	20.65±4.46	36.88±7.33	30.65±3.87	38.20±4.37
	HERO	51.39±1.99	57.37±1.31	40.49±4.12	41.87±4.88	44.18±0.78	50.14±1.31	32.20±2.21	42.56±0.97
	MUG	**79.49±4.26**	**79.54±3.93**	**84.24±5.64**	**85.76±4.95**	**49.12±6.16**	**72.10±4.68**	**33.24±2.63**	**42.20±1.96**
3-shot	HeCo	73.53±8.71	74.77±7.29	64.49±6.69	66.09±6.50	24.93±1.92	40.56±8.13	33.15±2.66	40.12±2.39
	HGMAE	79.42±7.82	79.58±7.53	71.39±8.10	73.88±7.61	23.73±3.80	41.71±8.45	32.47±2.75	40.74±3.10
	HERO	63.78±2.25	63.74±2.23	61.24±0.67	61.50±0.94	47.38±0.67	53.17±1.26	34.17±1.32	42.63±1.49
	MUG	**84.39±2.48**	**83.74±2.87**	**90.56±0.66**	**91.62±0.49**	**66.80±2.85**	**81.92±2.64**	**35.01±1.14**	**44.48±1.98**
5-shot	HeCo	77.58±4.92	78.08±4.19	68.75±4.53	70.05±4.71	25.47±2.12	40.86±7.60	33.50±2.08	40.55±1.94
	HGMAE	81.68±4.72	81.26±5.32	79.03±6.92	80.39±7.00	24.84±3.10	41.96±9.75	33.45±2.20	40.46±3.46
	HERO	74.08±1.14	73.76±1.14	63.31±0.80	63.42±0.84	44.18±0.78	50.14±1.31	32.20±2.21	42.56±0.97
	MUG	**83.83±1.10**	**82.64±2.16**	**90.76±0.37**	**91.80±0.38**	**68.30±1.56**	**83.76±1.60**	**39.36±4.08**	**45.82±1.93**

Table 2: Performance on few-shot node classification with standard deviations. The best results are highlighted in bold.

contextual structure encoder explicitly encodes both node types and relation types, and utilizes a dimension encoder to achieve alignment in the representation space. This design enables MUG to preserve and leverage the diverse semantics inherent in heterogeneous graphs, supporting more effective transfer across domains. Additionally, we observe that the performance gap between MUG and the baseline methods is particularly pronounced on the AMiner dataset. This is mainly because node attributes in AMiner are represented as one-hot vectors, making them especially susceptible to information loss under dimension reduction. This observation further highlights the importance of our approach in unifying heterogeneous inputs and maintaining the integrity of original semantic information.
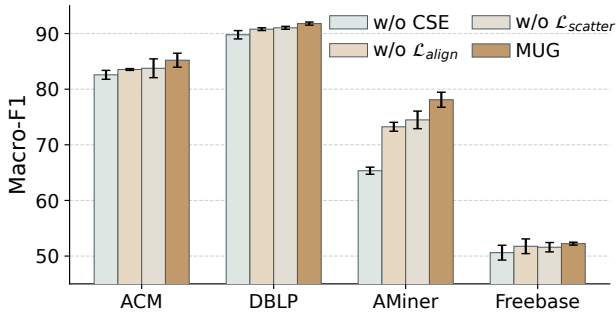

Figure 3: Ablation results for cross-domain node classification, trained on Freebase and evaluated on four datasets.

Cross-Domain few-shot node classification. For the cross-domain few-shot node classification experiments, we follow the same evaluation protocol as described above. In this setting, we trained MUG on the ACM dataset and directly evaluated it on the all datasets using only 1, 3, or 5 labeled train samples per class. As shown in Table 2, MUG achieves the best performance among all methods under all few-shot settings. Notably, in the 5-shot scenario, the performance of MUG approaches the results reported in Table 1 for the cross-domain node classification setting. This result demonstrates the strong generalization and transfer ca-

pability of our method, even when the amount of available labeled data is limited. The consistently high performance of MUG in few-shot scenarios highlights the effectiveness of its unified representation space and its ability to capture transferable heterogeneous information during pre-training.

Model Analysis

Figure 3 presents the results of ablation study under cross-domain node classification, where the model is trained on the Freebase dataset and evaluated on all four datasets. Here, CSE refers to the Contextual Structural Encoding component. Removing CSE consistently leads to the largest performance degradation, especially on cross-domain datasets such as AMiner, indicating that contextual structural encoding is crucial for capturing universal heterogeneous semantics beyond node attributes. Moreover, removing the alignment loss $\mathcal{L}_{align}$ or the scattering loss $\mathcal{L}_{scatter}$ also results in noticeable performance drops. This suggests that both losses play key roles in improving cross-domain generalization by mitigating domain-specific biases and promoting more transferable representations. Overall, the full model (MUG) achieves the best performance across all datasets.

Conclusion

In this work, we conduct an initial exploration into extending universal graph pre-training to heterogeneous graphs. We identify two key challenges: (i) the difficulty of unifying input due to the complex heterogeneity (i.e., node types, relation types and node attributes); and (ii) the limited transferability of heterogeneous information due to dataset-specific encoding and aggregation functions. To address these challenges, we propose MUG, a novel approach designed to learn transferable representations across diverse heterogeneous graphs. By leveraging a meta-path–based contextual encoding and a dimension-aware encoder, MUG can construct a unified representation space. Furthermore, MUG introduces a universal encoding and aggregation functions, facilitating effective information transfer while avoiding overfitting to specific datasets. Extensive experiments on cross-domain and few-shot settings demonstrate the strong performance of MUG.

Acknowledgments

References

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.

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.

He, D.; Shan, L.; Zhao, J.; Zhang, H.; Wang, Z.; and Zhang, W. 2024. Exploitation of a latent mechanism in graph contrastive learning: Representation scattering. *Advances in Neural Information Processing Systems*, 37: 115351–115376.

He, D.; Zhao, J.; Guo, R.; Feng, Z.; Jin, D.; Huang, Y.; Wang, Z.; and Zhang, W. 2023. Contrastive Learning Meets Homophily: Two Birds with One Stone. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 12775–12789. PMLR.

He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; and Wang, M. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In Huang, J. X.; Chang, Y.; Cheng, X.; Kamps, J.; Murdock, V.; Wen, J.; and Liu, Y., eds., *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, 639–648. ACM.

Hou, Z.; Liu, X.; Cen, Y.; Dong, Y.; Yang, H.; Wang, C.; and Tang, J. 2022. GraphMAE: Self-Supervised Masked Graph Autoencoders. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, 594–604. ACM.

Hu, B.; Fang, Y.; and Shi, C. 2019. Adversarial learning on heterogeneous information networks. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 120–129.

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

Huang, Y.; Zhao, J.; He, D.; Wang, X.; Li, Y.; Huang, Y.; Jin, D.; and Feng, Z. 2025. One Prompt Fits All: Universal Graph Adaptation for Pretrained Models. *arXiv preprint arXiv:2509.22416*.

Huo, C.; He, D.; Li, Y.; Jin, D.; Dang, J.; Pedrycz, W.; Wu, L.; and Zhang, W. 2025. Heterogeneous graph neural networks using self-supervised reciprocally contrastive learning. *ACM Transactions on Intelligent Systems and Technology*, 16(1): 1–21.

Jia, D.; Romić, I.; Shi, L.; Su, Q.; Liu, C.; Liu, J.; Holme, P.; Li, X.; and Wang, Z. 2025. Social networking agency and prosociality are inextricably linked in economic games. *Nature Human Behaviour*, 1–12.

Li, X.; Ding, D.; Kao, B.; Sun, Y.; and Mamoulis, N. 2021a. Leveraging meta-path contexts for classification in heterogeneous information networks. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, 912–923. IEEE.

Li, Z.; Liu, H.; Zhang, Z.; Liu, T.; and Xiong, N. N. 2021b. Learning knowledge graph embedding with heterogeneous relation attention networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(8): 3961–3973.

Liu, H.; Feng, J.; Kong, L.; Liang, N.; Tao, D.; Chen, Y.; and Zhang, M. 2023. One for all: Towards training one graph model for all classification tasks. *arXiv preprint arXiv:2310.00149*.

Mo, Y.; Nie, F.; Hu, P.; Shen, H. T.; Zhang, Z.; Wang, X.; and Zhu, X. 2024. Self-supervised heterogeneous graph learning: a homophily and heterogeneity view. In *The Twelfth International Conference on Learning Representations*.

Qiu, J.; Chen, Q.; Dong, Y.; Zhang, J.; Yang, H.; Ding, M.; Wang, K.; and Tang, J. 2020. Gcc: Graph contrastive coding for graph neural network pre-training. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 1150–1160.

Stewart, G. W. 1993. On the early history of the singular value decomposition. *SIAM review*, 35(4): 551–566.

Sun, X.; Cheng, H.; Li, J.; Liu, B.; and Guan, J. 2024. All in One: Multi-task Prompting for Graph Neural Networks (Extended Abstract). In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, 8460–8465. ijcai.org.

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.

Tian, Y.; Dong, K.; Zhang, C.; Zhang, C.; and Chawla, N. V. 2023. Heterogeneous graph masked autoencoders. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 9997–10005.

Wang, X.; Bo, D.; Shi, C.; Fan, S.; Ye, Y.; and Yu, P. S. 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. 2019. Heterogeneous graph attention network. In *The world wide web conference*, 2022–2032.

Wang, X.; Liu, N.; Han, H.; and Shi, C. 2021. Self-supervised heterogeneous graph neural network with co-contrastive learning. In *Proceedings of the 27th ACM*

SIGKDD conference on knowledge discovery & data mining, 1726–1736.

Wang, Z.; Zhang, Z.; Chawla, N.; Zhang, C.; and Ye, Y. 2024. Gft: Graph foundation model with transferable tree vocabulary. *Advances in Neural Information Processing Systems*, 37: 107403–107443.

Yang, Z.; Cohen, W.; and Salakhudinov, R. 2016. Revisiting semi-supervised learning with graph embeddings. In *International conference on machine learning*, 40–48. PMLR.

Yu, X.; Gong, Z.; Zhou, C.; Fang, Y.; and Zhang, H. 2025. SAMGPT: Text-free graph foundation model for multi-domain pre-training and cross-domain adaptation. In *Proceedings of the ACM on Web Conference 2025*, 1142–1153.

Yu, X.; Zhou, C.; Fang, Y.; and Zhang, X. 2024. Text-free multi-domain graph pre-training: Toward graph foundation models. *arXiv preprint arXiv:2405.13934*.

Zhao, J.; He, D.; Ge, M.; Huang, Y.; Shan, L.; Qin, Y.; and Feng, Z. 2024a. GA-GGD: Improving semantic discriminability in graph contrastive learning via Generative Adversarial Network. *Information Fusion*, 110: 102465.

Zhao, J.; Jin, D.; Ge, M.; Shan, L.; Wang, X.; He, D.; and Feng, Z. 2024b. FUG: Feature-universal graph contrastive pre-training for graphs with diverse node features. *Advances in Neural Information Processing Systems*, 37: 4003–4034.

Zhao, J.; Wang, X.; Shi, C.; Liu, Z.; and Ye, Y. 2020. Network schema preserving heterogeneous information network embedding. In *International joint conference on artificial intelligence (IJCAI)*.

Zhu, Y.; Shi, H.; Wang, X.; Liu, Y.; Wang, Y.; Peng, B.; Hong, C.; and Tang, S. 2025. Graphclip: Enhancing transferability in graph foundation models for text-attributed graphs. In *Proceedings of the ACM on Web Conference 2025*, 2183–2197.

Zhu, Y.; Wang, Y.; Shi, H.; Zhang, Z.; Jiao, D.; and Tang, S. 2024. Graphcontrol: Adding conditional control to universal graph pre-trained models for graph domain transfer learning. In *Proceedings of the ACM Web Conference 2024*, 539–550.

Zhuo, J.; Lu, Y.; Ning, H.; Fu, K.; Niu, B.; He, D.; Wang, C.; Guo, Y.; Wang, Z.; Cao, X.; and Yang, L. 2024a. Unified Graph Augmentations for Generalized Contrastive Learning on Graphs. In *NeurIPS*.

Zhuo, J.; Qin, F.; Cui, C.; Fu, K.; Niu, B.; Wang, M.; Guo, Y.; Wang, C.; Wang, Z.; Cao, X.; and Yang, L. 2024b. Improving Graph Contrastive Learning via Adaptive Positive Sampling. In *CVPR*, 23179–23187.