
Towards Heterogeneous Continual Graph Learning via Meta-knowledge Distillation

Guiquan Sun

University of Connecticut
guiquan.sun@uconn.edu

Xikun Zhang *

Nanyang Technological University
xikun.zhang@ntu.edu.sg

Jingchao Ni

University of Houston
jni7@uh.edu

Dongjin Song *

University of Connecticut
dongjin.song@uconn.edu

Abstract

Machine learning on heterogeneous graphs has experienced rapid advancement in recent years, driven by the inherently heterogeneous nature of real-world data. However, existing studies typically assume the graphs to be static, while real-world graphs are continuously expanding. This dynamic nature requires models to adapt to new data while preserving existing knowledge. To this end, this work addresses the challenge of continual learning on heterogeneous graphs by introducing the Meta-learning based Knowledge Distillation framework (MKD), designed to mitigate catastrophic forgetting in evolving heterogeneous graph structures. MKD combines rapid task adaptation through meta-learning on limited samples with knowledge distillation to achieve an optimal balance between incorporating new information and maintaining existing knowledge. To improve the efficiency and effectiveness of sample selection, MKD incorporates a novel sampling strategy that selects a small number of target-type nodes based on node diversity and maintains fixed-size buffers for other types. The strategy retrieves first-order neighbors along metapaths and selects important neighbors based on their structural relevance, enabling the sampled subgraphs to retain key topological and semantic information. In addition, MKD introduces a semantic-level distillation module that aligns the attention distributions over different metapaths between teacher and student models, encouraging semantic consistency beyond the logit level. Comprehensive evaluations across three benchmark datasets validate MKD’s effectiveness in handling continual learning scenarios on expanding heterogeneous graphs.

1 Introduction

In recent years, heterogeneous graphs have emerged as a powerful structure for modeling complex real-world systems across diverse domains, including biological systems, social networks, and recommendation systems [1]. Unlike homogeneous graphs with uniform node and edge types, heterogeneous graphs capture intricate relationships among multiple node and edge types, enabling more nuanced and comprehensive representations for the target systems [2]. For example, in recommendation system research, the user-item networks contain user nodes, item nodes, and the purchase relationship. In a cellular system, the signalling network encodes the complex interaction among different types of entities including small molecules, proteins, and genes [3]. To extract meaningful insights from heterogeneous graphs, different Heterogeneous Graph Neural Networks

*Corresponding authors: Xikun Zhang and Dongjin Song

(HGNNs) have been developed. These approaches mainly fall into two categories: metapath-based HGNNs [4, 5, 6, 7, 8, 9, 10, 11, 12], which aggregate information over predefined meta-paths to mine the intricate relationship among the heterogeneous nodes and edges, and meta-path-free models [13, 14, 15, 16, 17, 18, 19] that automatically learn node interactions without preconfigured paths, offering enhanced adaptability but potentially sacrificing interpretability.

While HGNNs have demonstrated remarkable success across various tasks, they face fundamental challenges in dynamic environments in which the heterogeneous graphs are constantly evolving with significant data distribution shift. For example, in recommendation systems, the user-item graphs undergo continuous expansion through new products, users, and purchase patterns. In biological systems, the signalling networks also grow with the discovery of novel signalling pathways. In these scenarios, HGNNs are expected to continuously incorporate the incoming new knowledge while preserving the learnt patterns. However, naive incremental training approaches will inevitably trigger catastrophic forgetting over previously learnt knowledge, as optimization for new data distributions tends to overwrite previously encoded information in the model parameters. While this phenomenon has been extensively studied in continual learning research over independent data (e.g. images) [20, 21, 22, 23] and homogeneous graphs [24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34], addressing catastrophic forgetting in heterogeneous graph learning remains largely unexplored.

To bridge this critical gap, we present a systematic investigation of the Heterogeneous Continual Graph Learning (HCGL) problem and introduce Meta-learning based Knowledge Distillation (MKD), a novel framework specifically designed for HCGL. The proposed method consists of three core components: a Gradient-based Meta-learning Module (G-MM) for rapid adaptation, an Efficient Heterogeneous Subgraph Sampling (E-HSS) strategy for experience replay that leverages both node diversity and structural information, and a Heterogeneity-aware Knowledge Distillation (HKD) module that aligns cross-task knowledge at both the logit and semantic levels. First, unlike homogeneous graphs, the diverse node and edge types in heterogeneous graphs introduce extra challenges to knowledge transfer across different tasks. Therefore, for effective model adaptation over the continuously evolving heterogeneous graph structures, we design a Gradient-based Meta-learning Module (G-MM). G-MM learns an optimized initialization of model parameters by leveraging gradient-based meta-learning, which allows fast model adaptation to new tasks using only a small number of examples. Through task-specific adjustments, G-MM also avoids the task-wise interference and ensures performance robustness on learned tasks. Second, to preserve essential knowledge from previous tasks while avoiding excessive memory overhead, we propose an efficient Heterogeneous Subgraph Sampling (E-HSS) mechanism that integrates node diversity and structural information. E-HSS selects representative and important nodes from the target type and preserves the original topological structure guided by meta-paths. This strategy minimizes data redundancy by assigning independent buffers to different node types and computing node importance based on relation types and degree metrics, thereby achieving structural balance and semantic coverage across heterogeneous nodes and enhancing the model’s long-term memory retention. Finally, to align knowledge across tasks and maintain structural consistency in heterogeneous settings, we introduce a Heterogeneity-aware Knowledge Distillation (HKD) module. Beyond knowledge retained via experience replay, this module performs both logit-level and semantic-level alignment between the current and previous task models. By distilling prediction distributions and meta-path-based attention representations, HKD guides the student model to retain predictive information while incorporating multi-level semantic information, significantly improving continual learning performance on heterogeneous graphs. Our contribution can be summarized as follows:

- We systematically investigate the problem of Heterogeneous Continual Graph Learning (HCGL), which has been largely unexplored in previous research.
- We propose an efficient sampling strategy that preserves both node diversity and heterogeneous structural information for memory-efficient experience replay.
- We introduce a two-level heterogeneity-aware distillation module that aligns knowledge across tasks at both the logit and semantic levels.
- Extensive experiments on multiple benchmark heterogeneous graph datasets demonstrate that MKD significantly outperforms existing continual graph learning methods in terms of accuracy, efficiency, and memory utilization.

2 Related Work

Heterogeneous Graph Neural Networks. Heterogeneous graphs are characterized by their complex topological structures and rich semantic information, stemming from the presence of multiple node and edge types. Traditional Graph Neural Networks (GNNs) often struggle to handle such heterogeneity directly, leading to the development of specialized approaches known as Heterogeneous Graph Neural Networks (HGNNs). These models are designed to effectively capture the multi-typed relationships inherent in heterogeneous graphs. Existing HGNN methods can be broadly categorized into two groups: metapath-based and metapath-free approaches. Metapath-based methods [4, 5, 6, 7, 8, 9, 10, 11, 12] first aggregate features from neighbors sharing the same semantic context and then integrate information across different semantics. On the other hand, metapath-free methods [13, 14, 15, 16, 17, 18, 19] aggregate information from all types of neighbors within a local 1-hop neighborhood, akin to traditional GNNs, but incorporate additional mechanisms such as attention to encode semantic details (e.g., node and edge types) into the message-passing process. While HGNNs have achieved remarkable success in learning from static heterogeneous graphs, they typically rely on the assumption that the entire graph is accessible during training. This assumption limits their applicability in dynamic real-world scenarios, where graph structures evolve continuously over time, as seen in knowledge graphs and social networks. Consequently, Continual Learning (CL) emerges as a critical yet underexplored research direction for advancing HGNNs in such dynamic settings.

Continual Graph Learning. Continual Graph Learning (CGL) enables models to learn from evolving graph-structured data while mitigating catastrophic forgetting, where performance on earlier tasks deteriorates after learning new ones. Existing methods can be broadly categorized into three types: parameter-isolation, regularization, and memory-replay approaches. *Parameter-isolation methods* preserve important parameters by freezing them after training on a task, preventing interference from subsequent updates. Examples include HPNs [35] and the PI-GNN framework [27]. *Regularization methods* constrain updates to important parameters via penalty terms. Representative methods include EWC [21], MAS [20], and TWP [25], which considers local graph topology. Knowledge distillation is also commonly used to retain information from prior models [36, 37]. *Memory-replay methods* sample and store representative nodes or subgraphs for replay [38, 39]. However, these methods may suffer from memory explosion due to the growth of computational subgraphs. ER-GNN [26] addresses this by sampling node attributes only, while SSM [40] stores sparsified subgraphs. PDGNNS-TEM [41] maintains dynamic embeddings of subgraphs, and UGCL [42] unifies memory replay and distillation to support both node and graph classification tasks. TACO [43] replaces raw subgraph storage with graph coarsening, enabling compact memory usage while preserving key structural information. A recent work, RL-GNN [44], addresses graph-level continual learning by identifying invariant rationales across tasks to mitigate catastrophic forgetting caused by spurious correlations. Despite recent progress, CGL still faces challenges in reducing forgetting and computational cost. Moreover, continual learning on heterogeneous graphs remains underexplored. In this work, we address these challenges using a combination of meta-learning and knowledge distillation tailored to heterogeneous graph settings.

Meta Learning. Meta-learning aims to enable models to adapt rapidly to new tasks by leveraging shared knowledge. Existing approaches are generally categorized into optimization-based (e.g., MAML [45], Reptile [46]) and metric-based methods (e.g., Prototypical Networks [47], Matching Networks [48]). In continual learning, meta-learning has been applied to online [49] and few-shot scenarios [50]. Recently, its integration with graph learning has drawn increasing attention. For example, MetaCLGraph [28] incorporates meta-learning with experience replay for continual graph learning, and HG-Meta [51] applies it to few-shot graph classification on heterogeneous graphs. Motivated by these advances, we further explore its potential in mitigating forgetting and improving adaptability in heterogeneous continual learning.

3 Preliminary

In this section, we introduce fundamental concepts and formalize the problem of Heterogeneous Continual Graph Learning (HCGL). Our primary objective is to develop a continual learning framework on Dynamic Heterogeneous Graphs to mitigate the issue of catastrophic forgetting. In the HCGL setting, a Heterogeneous Graph Neural Network (HGNN) learns a sequence of tasks in a domain incremental setting [52] (See Appendix B.3 for the details of this setting), without access to the data

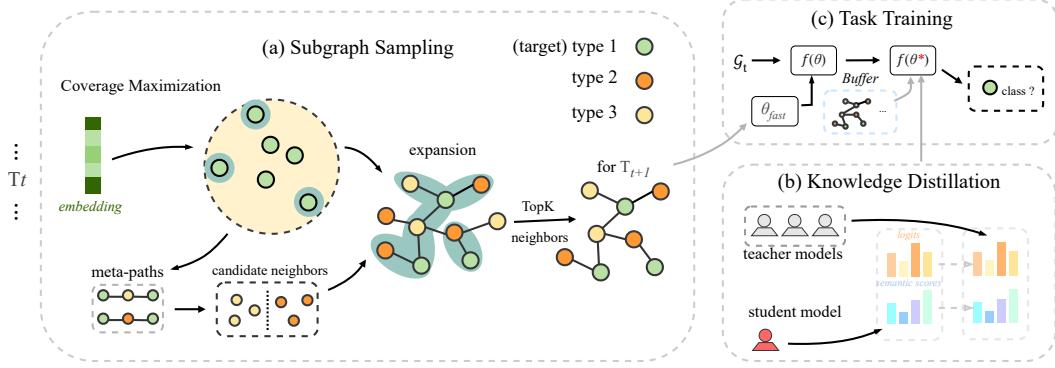


Figure 1: The overall framework of MKD. (a) Subgraph Sampling: Construct task-specific subgraphs by selecting diverse target-type nodes and expanding to related node types via relation-aware importance. (b) Knowledge Distillation: Align previous and current tasks via logit-level and semantic-level distillation. (c) Task Training: Use meta-learning for fast adaptation, jointly optimizing task loss, replay loss, and distillation loss.

from previous tasks. However, it is allowed to utilize a memory buffer with limited capacity to store representative information. The goal of the framework is to maximize the prediction accuracy across all tasks after training while minimizing the forgetting of previously acquired knowledge during the learning of new tasks.

In this work, we focus on node classification tasks and adopt a common continual learning setup in which the dataset is partitioned into a sequence of tasks based on node class labels. In this way, different splits have non-overlapping category labels.

Definition 1 (Heterogeneous Graph). A heterogeneous graph is defined as $G = (V, E)$, where V is the set of nodes, and E is the set of edges. Each node $v \in V$ is associated with a type given by the node-type mapping function $\phi : V \rightarrow \mathcal{A}$, and each edge $e \in E$ is associated with a type given by the edge-type mapping function $\psi : E \rightarrow \mathcal{R}$, where \mathcal{A} and \mathcal{R} denote the sets of node and edge types, respectively. The graph is considered heterogeneous if $|\mathcal{A}| + |\mathcal{R}| > 2$. Let V_τ denote the set of nodes of type $\tau \in \mathcal{T}$.

Definition 2 (Metapath). A metapath \mathcal{P} is a sequence of node and edge types in the form of $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \dots \xrightarrow{R_l} A_{l+1}$, which defines a composite relation $R_1 \circ R_2 \circ \dots \circ R_l$ between node types A_1 and A_{l+1} , where \circ denotes the composition operator. The metapath captures high-level semantic connections across different types of nodes through a specific sequence of edge types.

Definition 3 (Heterogeneous Continual Graph Learning). Let $\mathcal{G} = (V, E)$ be a dynamic heterogeneous graph, where V is the node set and E is the edge set. The feature set is type-specific, i.e., $\mathcal{F} = \{F_\tau \in \mathbb{R}^{|V_\tau| \times d_\tau} \mid \tau \in \mathcal{A}\}$. In the HCGL setting, the model learns a sequence of tasks $\{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_T\}$, without access to the data from previous tasks. Specifically, each task \mathcal{T}_t corresponds to a subgraph \mathcal{G}_t with disjoint category labels, and due to storage limitations, the model can only access the data available at the current task. For each subgraph \mathcal{G}_t , we divide it into a training set $\mathcal{G}_t^{\text{tr}} = (V_t^{\text{tr}}, E_t^{\text{tr}})$ and a testing set $\mathcal{G}_t^{\text{te}} = (V_t^{\text{te}}, E_t^{\text{te}})$ to train and evaluate the model f_θ .

4 Methodology

In this section, we present our Meta-learning based Knowledge Distillation (MKD) framework, providing a detailed analysis of its core components and demonstrating how they systematically address the identified challenges.

4.1 Fast Adaptation with Gradient-based Meta-Learning

Efficient adaptation to new data patterns is crucial for ensuring the practical usability of a model in an evolving heterogeneous graph. To achieve this, we introduce the Gradient-based Meta-learning Module (G-MM), which optimizes model parameter initialization to enable rapid adaptation to new tasks. This, in turn, helps preserve performance on the current task. Specifically, given the training node set V_t^{tr} of the current task \mathcal{T}_t , we select e samples from V_t^{tr} based on the Coverage Maximization (CM) strategy (introduced in Section 4.2), which identifies key nodes by maximizing node diversity. In practice, labeled data in heterogeneous graphs is often highly limited. By performing gradient descent on the selected small sample set \mathcal{E} , the model can quickly adapt to the current task, a process referred to as the inner update. Let θ denote the set of model parameters. For the current task \mathcal{T}_t , We feed the sampled node set \mathcal{E} into the model and compute the loss $\mathcal{L}_{\mathcal{E}}$, updating θ to θ_{fast} via gradient descent (Inner Update):

$$\theta_{fast} \leftarrow \theta - \alpha \nabla_{\theta} \mathcal{L}_{\mathcal{E}}(\theta), \quad (1)$$

where α is the step size for the inner update. Meta-learning with a small number of samples enables the model to prevent overfitting during experience replay while ensuring rapid adaptation to the current task, thereby maintaining robust overall performance.

4.2 Efficient Heterogeneous Subgraph Sampling

In addition to adapting to new data patterns, overcoming catastrophic forgetting is crucial, as previously observed patterns may reappear in practical scenarios. While many existing experience replay methods primarily consider homogeneous graphs or single-type nodes, they fail to capture the complex semantic and structural dependencies in heterogeneous graphs. To address this, we propose an efficient subgraph sampling strategy explicitly tailored for heterogeneous graphs, which jointly considers the diversity of nodes within the target node type and the structural connections with nodes of other types. Our method proceeds in two stages: (1) selecting representative target-type nodes by maximizing diversity in the feature or embedding space, and (2) expanding to other types of nodes through relation-type-aware importance estimation, ensuring that both semantic diversity and heterogeneous structural patterns are preserved.

Step 1: Target Node Selection via Coverage Maximization. Prior research has shown that different data points contribute unequally to model training: some significantly enhance performance, while others offer limited benefit [53]. Motivated by this, we identify key nodes from the current task by maximizing node diversity using the Coverage Maximization (CM) strategy. This approach allows us to retain essential historical information with only a small number of nodes. Formally, given the training node set V_t^{tr} of task \mathcal{T}_t , CM selects a subset by maximizing the coverage of the attribute/embedding space:

$$\mathcal{N}(v_i) = \{v_j \mid \text{dist}(v_i, v_j) < d, \mathcal{Y}(v_i) \neq \mathcal{Y}(v_j)\}, \quad (2)$$

where $\mathcal{Y}(v_i)$ denotes the label of node v_i , and $\mathcal{N}(v_i)$ represents the set of nodes from different classes that are within a distance d of v_i . Given the memory buffer \mathcal{B}_{τ_t} for the target node type τ_t , we select e nodes per class with the smallest $|\mathcal{N}(v_i)|$ as representative experiences.

Step 2: Multi-hop Neighbor Expansion. Given target nodes V_{τ_t} , we collect candidate nodes \mathcal{C}_r for each relation type $r \in \mathcal{R}$ through:

$$\mathcal{C}_r = \bigcup_{v \in V_{\tau_t}} \{u \mid (v, u) \in E_r \vee (u, v) \in E_r\}, \quad (3)$$

where E_r denotes edges of relation type r . We define node importance $\pi(v)$ as the sum of relation-specific degrees:

$$\pi(v) = \sum_{r \in \mathcal{R}} \deg_r(v), \quad (4)$$

where $\deg_r(v)$ measures the connectivity strength of node v under relation type r . For each node type τ with buffer \mathcal{B}_{τ} , we select top- $|\mathcal{B}_{\tau}|$ nodes by $\pi(v)$:

$$V_{\tau}^* = \text{topk}_{v \in \mathcal{C}_{\tau}} (\pi(v), |\mathcal{B}_{\tau}|), \quad (5)$$

With sampled target type node set V_{τ_t} and the selected heterogeneous neighbor nodes set $\{V_\tau\}_{\tau \in \mathcal{A}/\tau_t}$, we construct the heterogeneous subgraph $\mathcal{G}_{sub} = (V', E')$ through:

$$V' = V_{\tau_t} \cup \bigcup_{\tau \neq \tau_t} V_\tau, \quad E' = \{(u, v) \in E \mid u, v \in V'\}, \quad (6)$$

For all $v \in V_{\tau_t}$, we retain their original features and labels, and maintain the structure of the type-specific projection. When learning a new task, we perform the experience replay by incorporating an auxiliary loss computed over the memory buffer \mathcal{B} into the current task loss (i.e., the cross-entropy loss between the given labels \mathcal{Y} and the predicted labels $\hat{\mathcal{Y}}$)

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{\mathcal{T}_t}(\mathcal{G}_t, \theta) + \lambda_{er} \mathcal{L}_{er}(\mathcal{G}_{sub}, \theta) \\ &= - \sum_{v_i \in \mathcal{V}_t^{tr}} \mathcal{Y}(v_i) \log \hat{\mathcal{Y}}(v_i) - \lambda_{er} \sum_{v_j \in \mathcal{B}} \mathcal{Y}(v_j) \log \hat{\mathcal{Y}}(v_j), \end{aligned} \quad (7)$$

where $\hat{\mathcal{Y}}(v_i)$ is the predicted label of node v_i , λ_{er} modulates the contribution of buffered data in the overall loss.

4.3 Heterogeneity-aware Knowledge Distillation for Task Alignment

Existing knowledge distillation methods are often developed for homogeneous scenarios and fail to account for the multi-relational semantics and cross-type dependencies in heterogeneous graphs. In contrast, our Heterogeneity-aware Knowledge Distillation (HKD) module is explicitly tailored for heterogeneous graph continual learning. Instead of relying solely on logit-based distillation, we propose a two-level alignment strategy that leverages both prediction and semantic signals to capture graph heterogeneity. The logit-level distillation transfers soft label distributions to retain discriminative knowledge, while the semantic-level alignment matches meta-path-aware attention scores, enabling the student model to preserve high-order structural patterns unique to heterogeneous graphs.

Logit-level Distillation Inspired by previous distillation-based approaches [54], we transfer the soft knowledge from the teacher model (previous task classifier) to the student model (current task classifier), enabling the student model to learn both new knowledge from the current task and retained knowledge from past tasks. Mimicking teacher's prediction results enables the student model to learn the secondary information from previous tasks that cannot be expressed by the experience replay data alone. Soft knowledge from teacher model is formulated as the predicted probability of the labels in the current task data:

$$P^T(z_i, t) = \text{Softmax}(f_T(z_i), t) = \frac{\exp[f_T(z_i)/t]}{\sum_j \exp[f_T(z_j)/t]}, \quad (8)$$

where z_i is the embedding of node v_i in \mathcal{V}_i^{tr} , $f_T(z_i)$ is the score logit of z_i obtained from teacher model, and t is the temperature index to soften the peaky softmax distribution [55]. Thus, the knowledge distillation loss of the teacher model and the student model is defined as follows:

$$\mathcal{L}_{\text{logit}} = \text{Mean}(t^2 \sum_l^N \sum_{v_i \in \mathcal{V}_i^{tr}} P_l^T(z_i, t) \log \frac{P_l^T(z_i, t)}{P^S(z_i, t)}), \quad (9)$$

where N is the number of teacher models, P^T and P^S are the predicted distributions of teacher model and student model respectively.

Semantic-level Distillation To preserve metapath-induced semantic patterns, we propose an attention-based structural distillation method. Given a predefined metapath set $\mathcal{P} = \{P_1, \dots, P_M\}$, let $\alpha_{P_m}^{(T)}(v_i)$ and $\alpha_{P_m}^{(S)}(v_i)$ denote the attention coefficients for metapath P_m computed by the teacher and student models, respectively. These attention coefficients reflect the importance of different semantic contexts encoded by heterogeneous structural patterns. To align the semantic-level knowledge between the teacher and the student, we define the semantic alignment loss as:

$$\mathcal{L}_{\text{sem}} = \sum_{m=1}^M \|\alpha_{P_m}^{(T)} - \alpha_{P_m}^{(S)}\|_2 \quad (10)$$

where $\alpha_{P_m} \in \mathbb{R}^{|V^{tr}|}$ is the normalized attention vector over all nodes for metapath P_m . We provide a detailed discussion in Appendix A.2 on how this module can be extended to arbitrary HGNNs beyond metapath-based architectures.

This loss guides the student model to capture relational importance and structural dependencies aligned with the teacher model, thereby preserving semantic information that is critical in heterogeneous graphs. Afterwards, we combine the logit level loss and the semantic level loss:

$$\mathcal{L}_{kd} = \lambda_{\text{logit}} \mathcal{L}_{\text{logit}} + \lambda_{\text{sem}} \mathcal{L}_{\text{sem}}, \quad (11)$$

where λ_{logit} and λ_{sem} are both trade-off weights for balancing the losses. The final objective of node representation learning is to minimize the joint loss including current task loss $\mathcal{L}_{\mathcal{T}_t}$, experience replay loss \mathcal{L}_{er} , and the KD loss \mathcal{L}_{kd} :

$$\mathcal{L}_{joint} = \mathcal{L}_{\mathcal{T}_t} + \lambda_{er} \mathcal{L}_{er} + \lambda_{kd} \mathcal{L}_{kd}, \quad (12)$$

where λ_{kd} is a trade-off weight for balancing the losses. Taking the parameters θ_{fast} obtained from Equation 1 as initialization, the student model for the current task is further updated as follows:

$$\theta^* = \theta - \beta \nabla_{\theta} \mathcal{L}_{joint}(\theta), \quad (13)$$

where β is learning rate. By minimizing \mathcal{L}_{joint} , parameters θ of HGNN model are optimized for downstream node classification task. The detailed framework of MKD is provided in Appendix 1.

5 Experiments

We conducted experiments on four datasets: DBLP [14], IMDB [14], Freebase [14] and Yelp [56] to evaluate the performance of MKD. We adopt three widely used HGNN backbones in our experiments: HAN [6], MAGNN [4], and HGT [13]. See Appendix B.1 for the details of the datasets and B.4 for experiment setup.

Baselines To evaluate the effectiveness of our proposed method, we select strong baselines from the Continual Graph Learning Benchmark (CGLB) [24], including Elastic Weight Consolidation (EWC) [21], Memory Aware Synapses (MAS) [20], Experience Replay GNN (ER-GNN) [26], and Topology-aware Weight Preserving (TWP) [25]. EWC and MAS were previously not designed for graphs, so we train a HGNN on new tasks but ignore the graph structure when applying continual learning strategies. Additionally, we include two experience-replay-based baselines: MetaCLGraph [28] and FTF-ER [57]. Furthermore, we use Finetune method (without any continual learning techniques) as a lower bound, and Joint Train (which allows access to previous data during training) as an approximate upper bound [58].

Metrics In our experiments, *Average Performance* (AP) and *Average Forgetting* (AF) [23], are used to measure the performance on test sets. AP and AF are defined as $AP = \frac{1}{T} \sum_{t=1}^T a_{T,t}$, $AF = \frac{1}{T-1} \sum_{t=1}^{T-1} (a_{T,t} - a_{t,t})$, where T is the total number of tasks and $a_{i,j}$ is the accuracy of the model on the test set of task j after it is trained on task i .

Performance Evaluation We first systematically evaluate the extent of catastrophic forgetting in three HGNNs. From Table 1, we can observe that all HGNNs exhibit catastrophic forgetting on previous tasks. For example, on the DBLP dataset, the average forgetting (AF) for HAN and MAGNN is over 26% and 34%, respectively; on the IMDB dataset, the AF for all HGNNs is over 26%, 17%, and 21%, respectively.

We evaluate the performance of MKD and other baselines on three datasets and three backbone HGNN models. We report only the results of HAN on the Yelp dataset in Table 4, as the other two models run out of memory on this dataset. Additionally, we adapt both 2-way (i.e., two node classes per task) and 3-way settings for the Yelp dataset. We report the average values and the standard deviations over 5 runs. It shows the proposed MKD achieves superior performance across all datasets and HGNN backbones. It is noteworthy that, in some cases, FTF-ER, MetaCLGraph, and ER-GNN perform better than MKD in terms of AF, despite their lower AP scores compared to MKD. This is because these methods rely on retraining nodes from previously learned tasks. Once enough nodes are sampled, these models can largely mitigate catastrophic forgetting. However, this also sacrifices performance when learning new tasks. Although these methods attempt to address the new-old task trade-off through different strategies (e.g., ER-GNN reweights the loss between new and old tasks), they still cannot fully avoid the performance drop. We further discuss this in Appendix C.3.

Table 1: Performance comparison with different HGNN backbones on three benchmark datasets. The symbol \uparrow (\downarrow) means higher (lower) is better. The best results are highlighted in **bold**, while the second best results are underlined. "OOM" means that the model runs out of memory on large graphs.

Base Models	Methods	DBLP		IMDB		Freebase	
		AP / % \uparrow	AF / % \downarrow	AP / % \uparrow	AF / % \downarrow	AP / % \uparrow	AF / % \downarrow
HAN	Finetune	82.8 \pm 4.6	26.6 \pm 9.1	68.8 \pm 2.2	26.0 \pm 2.8	53.8 \pm 4.7	25.7 \pm 8.8
	EWC	85.4 \pm 5.5	21.5 \pm 10.6	69.6 \pm 1.2	25.8 \pm 2.4	58.3 \pm 2.4	18.4 \pm 4.2
	MAS	91.1 \pm 0.8	<u>8.3 \pm 1.5</u>	72.0 \pm 2.3	20.5 \pm 3.9	59.0 \pm 1.0	14.8 \pm 0.9
	TWP	90.5 \pm 1.3	10.1 \pm 2.6	74.8 \pm 2.7	14.5 \pm 4.4	60.9 \pm 4.3	10.8 \pm 6.9
	ER-GNN	89.7 \pm 2.9	13.1 \pm 6.1	74.9 \pm 2.3	12.7 \pm 4.2	56.5 \pm 1.3	19.6 \pm 0.9
	FTF-ER	90.8 \pm 1.4	9.8 \pm 2.4	72.0 \pm 4.2	19.7 \pm 6.4	58.1 \pm 0.7	14.6 \pm 0.6
	MetaCLGraph	89.9 \pm 0.9	11.4 \pm 2.3	74.8 \pm 3.2	14.8 \pm 5.7	59.4 \pm 2.5	13.0 \pm 4.1
	Ours	93.6 \pm 0.8	4.3 \pm 1.6	<u>78.4 \pm 1.7</u>	<u>7.3 \pm 3.0</u>	60.9 \pm 1.5	<u>11.1 \pm 1.6</u>
	Joint Train	95.2 \pm 0.6	1.7 \pm 1.0	80.4 \pm 0.7	3.9 \pm 0.8	65.6 \pm 0.9	2.8 \pm 2.0
HGT	Finetune	86.2 \pm 1.5	12.7 \pm 2.2	75.1 \pm 7.5	17.0 \pm 9.4	67.8 \pm 1.8	16.5 \pm 2.5
	EWC	89.1 \pm 1.2	10.7 \pm 1.8	77.3 \pm 0.4	12.5 \pm 0.2	69.6 \pm 3.0	14.7 \pm 2.9
	MAS	89.8 \pm 2.5	9.0 \pm 3.0	77.5 \pm 1.2	11.8 \pm 2.6	70.4 \pm 3.1	14.8 \pm 2.0
	TWP	89.1 \pm 2.1	10.3 \pm 1.1	77.7 \pm 0.6	11.1 \pm 1.6	69.6 \pm 2.0	15.1 \pm 0.9
	ER-GNN	90.7 \pm 1.5	<u>3.5 \pm 0.4</u>	77.3 \pm 2.2	<u>9.0 \pm 3.6</u>	69.5 \pm 1.4	13.6 \pm 1.7
	MetaCLGraph	90.0 \pm 0.7	10.9 \pm 1.4	76.1 \pm 0.3	<u>12.2 \pm 1.8</u>	69.1 \pm 2.2	16.3 \pm 2.5
	FTF-ER	89.4 \pm 2.2	5.5 \pm 1.9	76.1 \pm 1.4	8.9 \pm 3.5	69.7 \pm 2.8	13.0 \pm 2.5
	Ours	92.8 \pm 0.8	3.3 \pm 1.2	78.4 \pm 0.4	6.6 \pm 1.7	70.6 \pm 1.2	9.1 \pm 2.7
	Joint Train	93.9 \pm 0.6	1.7 \pm 1.0	80.0 \pm 0.7	2.0 \pm 0.9	72.5 \pm 0.6	8.5 \pm 1.2
MAGNN	Finetune	79.1 \pm 5.1	34.8 \pm 10.1	71.3 \pm 0.6	21.7 \pm 3.0	OOD	OOD
	EWC	91.0 \pm 2.3	10.9 \pm 4.5	73.4 \pm 1.6	16.7 \pm 2.5	OOD	OOD
	MAS	92.1 \pm 2.1	7.8 \pm 5.3	74.3 \pm 2.0	14.2 \pm 3.7	OOD	OOD
	TWP	91.6 \pm 2.1	7.4 \pm 5.0	74.0 \pm 0.8	13.7 \pm 3.3	OOD	OOD
	ER-GNN	90.0 \pm 2.1	12.4 \pm 4.3	75.3 \pm 0.9	<u>10.3 \pm 1.5</u>	OOD	OOD
	FTF-ER	92.8 \pm 0.8	1.9 \pm 2.3	73.0 \pm 0.6	<u>15.8 \pm 1.4</u>	OOD	OOD
	MetaCLGraph	92.3 \pm 0.4	<u>2.1 \pm 0.7</u>	75.6 \pm 1.0	11.8 \pm 1.3	OOD	OOD
	Ours	93.2 \pm 0.5	3.1 \pm 0.8	76.8 \pm 1.1	7.5 \pm 1.4	OOD	OOD
	Joint Train	94.1 \pm 0.1	1.9 \pm 1.0	77.4 \pm 0.7	4.7 \pm 1.5	OOD	OOD

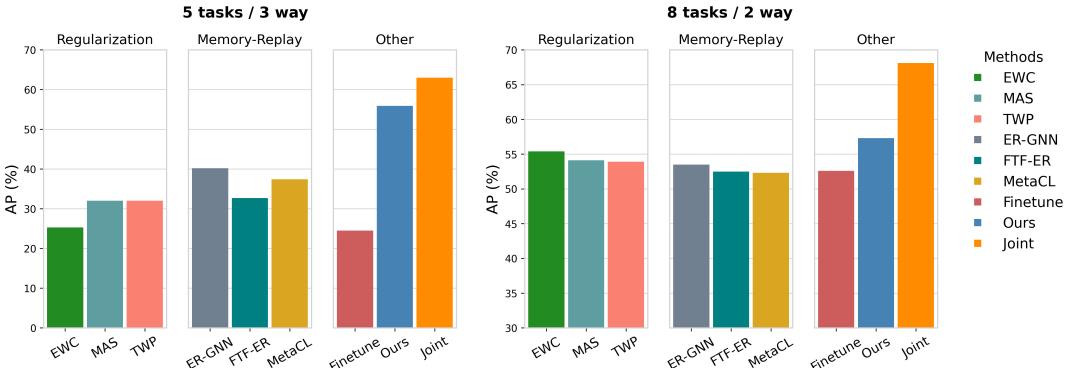


Figure 2: Performance comparison on the Yelp dataset under two settings using HAN as the backbone model.

Ablation Study To evaluate the effectiveness of individual modules in MKD, we conducted comprehensive ablation experiments using HAN as the backbone network across four benchmark datasets. The results are presented in Table 2.

We first examined the impact of each component by sequentially removing Experience Replay (ER), Heterogeneity-aware Knowledge Distillation (HKD), and the Gradient-based Meta-learning Module (GMM) while keeping other components intact. The results demonstrate that removing any module leads to performance degradation in both AP and AF, indicating each component's essential role in our framework. Notably, the removal of GMM shows relatively smaller performance impact (0.3-3.4% decrease in AP compared to 1.5-11.0% for ER/HKD), which aligns with its primary function of facilitating rapid adaptation to new tasks rather than long-term knowledge retention - the latter being mainly handled by the synergistic effect of ER and HKD. Particularly, the absence of HKD causes

Table 2: Ablation results of MKD using HAN as the backbone. We remove Experience Replay (ER), Heterogeneity-aware Knowledge Distillation (HKD), and the Gradient-based Meta-learning Module (GMM) individually. And “w/o E-HSS” denotes replacing the proposed *Efficient Heterogeneous Subgraph Sampling* with Coverage Maximization applied to all node types.

Methods	DBLP		IMDB		Freebase		Yelp	
	AP / %↑	AF / %↓	AP / %↑	AF / %↓	AP / %↑	AF / %↓	AP / %↑	AF / %↓
MKD	93.6 ± 0.8	4.3 ± 1.6	78.4 ± 1.7	7.3 ± 3.0	60.9 ± 1.5	11.1 ± 1.6	55.9 ± 0.6	7.5 ± 1.2
w/o ER	89.7 ± 0.6	12.0 ± 0.7	75.1 ± 0.9	12.8 ± 1.7	58.6 ± 1.2	16.3 ± 2.2	53.9 ± 2.4	6.8 ± 2.4
w/o HKD	89.6 ± 1.4	12.1 ± 2.6	72.2 ± 1.5	19.3 ± 2.3	59.4 ± 1.1	14.7 ± 1.8	44.9 ± 14.3	19.3 ± 17.0
w/o G-MM	92.6 ± 1.7	5.2 ± 0.9	75.0 ± 2.7	13.3 ± 4.5	59.7 ± 0.9	12.0 ± 2.2	55.6 ± 1.9	8.0 ± 2.0
w/o E-HSS	92.4 ± 1.4	6.0 ± 3.1	75.4 ± 2.2	12.1 ± 5.0	56.2 ± 0.4	18.0 ± 0.5	45.7 ± 14.0	23.7 ± 15.6

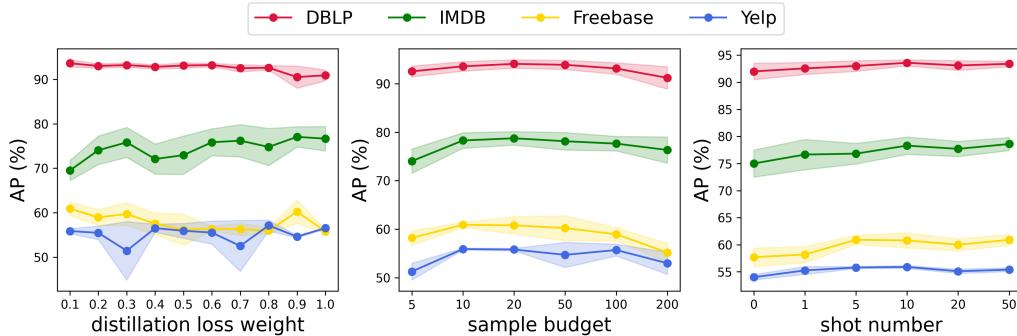


Figure 3: Sensitivity analysis of MKD with respect to key hyperparameters. The shaded areas represent variances. From left to right: (1) distillation loss weight, (2) experience replay sampling budget, and (3) shot number used in meta-learning. Due to space constraints, we only report the Average Performance (AP) here.

the most significant performance deterioration (3.6-12.0% increase in FR), highlighting its crucial role as the core mechanism against catastrophic forgetting.

Furthermore, we replaced our proposed Efficient Heterogeneous Subgraph Sampling (E-HSS) with the basic Coverage Maximization (CM) strategy. As shown in the last row of Table 2, E-HSS demonstrates substantial performance advantages, especially on edge-dense datasets like Yelp. These results confirm that E-HSS significantly outperforms conventional sampling methods in preserving the original topological structure of heterogeneous graphs. We further investigate the impact of each component in the HKD module in Appendix C.1.

Hyperparameters Analysis We analyze the impact of key hyperparameters on model performance, including the loss weight (λ_{kd}) in knowledge distillation (KD), the sampling budget for experience replay, and the shot number (i.e., the number of sampled nodes used in meta-learning), as shown in Figure 3. We further discuss in Appendix C.2 the impact of the trade-off weight used to balance the losses of two submodules in the knowledge distillation component. The experimental findings are as follows:

Distillation loss Weight (λ_{kd}): On IMDB, higher weights better support knowledge transfer, while lower values ($\lambda_{kd} < 0.2$) lead to increased forgetting. On Freebase, smaller weights ($0.1 \sim 0.3$) help mitigate forgetting, but larger weights ($\lambda_{kd} > 0.3$) hinder adaptability. For Yelp, the 3-way setting introduces large task shifts, requiring higher weights for stability and effective forgetting mitigation.

Sample Budget: The number of sampled nodes influences the stability-plasticity trade-off. Too few nodes lead to insufficient information, degrading performance, while excessive sampling limits the model’s adaptability to new tasks. As shown in Figure 3, on datasets like Yelp, a large sampling budget (e.g., >20) significantly impairs learning on new tasks due to the accumulation of historical information. **Shot Number:** The number of sampled nodes impacts the stability-plasticity trade-off. Meta-learning enables rapid adaptation to new tasks. As shown in the results, model performance steadily improves up to 10-shot, but declines at 20-shot, likely due to overfitting to the current task’s data patterns. The optimal shot number is 10, balancing historical information retention and computational efficiency.

6 Conclusion

Our proposed method integrates meta-learning, heterogeneity-aware knowledge distillation, and experience replay, specifically designed to address challenges in heterogeneous continual graph learning. By jointly considering the diversity within target node types and the structural relations across different node types, our efficient sampling strategy preserves critical topological and semantic information with limited data. The gradient-based meta-learning module enables rapid adaptation to new tasks, while the knowledge distillation module aligns knowledge across tasks at both prediction and semantic levels, effectively balancing model stability and plasticity. Extensive experiments on multiple real-world datasets validate the effectiveness of the proposed method on the Heterogeneous Continual Graph Learning (HCGL) problem.

7 Acknowledgment

This project is supported by the National Science Foundation under CAREER Award IIS-2338878 and a generous research gift from Morgan Stanley.

References

- [1] Chuan Shi, Binbin Hu, Wayne Xin Zhao, Philip S Yu, and Yanchi Liu. Heterogeneous information network embedding for recommendation. 2018.
- [2] Yizhou Sun and Jiawei Han. *Mining heterogeneous information networks: principles and methodologies*. Morgan & Claypool Publishers, 2012.
- [3] Marinka Zitnik, Monica Agrawal, and Jure Leskovec. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):i457–i466, 2018.
- [4] Xinyu Fu, Jian Zhang, Ziqiao Meng, and Irwin King. Magnn: Metapath aggregated graph neural network for heterogeneous graph embedding. In *Proceedings of the web conference 2020*, pages 2331–2341, 2020.
- [5] Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla. Heterogeneous graph neural network. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 793–803, 2019.
- [6] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. Heterogeneous graph attention network. In *The world wide web conference*, pages 2022–2032, 2019.
- [7] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 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*, pages 593–607. Springer, 2018.
- [8] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 135–144, 2017.
- [9] Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S Yu, and Tianyi Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment*, 4(11):992–1003, 2011.
- [10] Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J Kim. Graph transformer networks. *Advances in neural information processing systems*, 32, 2019.
- [11] Xiaocheng Yang, Mingyu Yan, Shirui Pan, Xiaochun Ye, and Dongrui Fan. Simple and efficient heterogeneous graph neural network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 10816–10824, 2023.

- [12] Guanghui Zhu, Zhennan Zhu, Hongyang Chen, Chunfeng Yuan, and Yihua Huang. Hagnn: Hybrid aggregation for heterogeneous graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.
- [13] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. Heterogeneous graph transformer. In *Proceedings of the web conference 2020*, pages 2704–2710, 2020.
- [14] Qingsong Lv, Ming Ding, Qiang Liu, Yuxiang Chen, Wenzheng Feng, Siming He, Chang Zhou, Jianguo Jiang, Yuxiao Dong, and Jie Tang. 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*, pages 1150–1160, 2021.
- [15] Shichao Zhu, Chuan Zhou, Shirui Pan, Xingquan Zhu, and Bin Wang. Relation structure-aware heterogeneous graph neural network. In *2019 IEEE international conference on data mining (ICDM)*, pages 1534–1539. IEEE, 2019.
- [16] Huiting Hong, Hantao Guo, Yucheng Lin, Xiaoqing Yang, Zang Li, and Jieping Ye. An attention-based graph neural network for heterogeneous structural learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 4132–4139, 2020.
- [17] Rui Zhang, Arthur Zimek, and Peter Schneider-Kamp. A simple meta-path-free framework for heterogeneous network embedding. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2600–2609, 2022.
- [18] Chaofan Fu, Guanjie Zheng, Chao Huang, Yanwei Yu, and Junyu Dong. Multiplex heterogeneous graph neural network with behavior pattern modeling. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 482–494, 2023.
- [19] Cuiying Huo, Dongxiao He, Yawen Li, Di Jin, Jianwu Dang, Witold Pedrycz, Lingfei Wu, and Weixiong Zhang. Heterogeneous graph neural networks using self-supervised reciprocally contrastive learning. *ACM Transactions on Intelligent Systems and Technology*, 16(1):1–21, 2025.
- [20] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.
- [21] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.
- [22] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.
- [23] David Lopez-Paz and Marc’Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.
- [24] Xikun Zhang, Dongjin Song, and Dacheng Tao. Cglb: Benchmark tasks for continual graph learning. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.
- [25] Huihui Liu, Yiding Yang, and Xinchao Wang. Overcoming catastrophic forgetting in graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 8653–8661, 2021.
- [26] Fan Zhou and Chengtai Cao. Overcoming catastrophic forgetting in graph neural networks with experience replay. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4714–4722, 2021.
- [27] Peiyan Zhang, Yuchen Yan, Chaozhuo Li, Senzhang Wang, Xing Xie, Guojie Song, and Sunghun Kim. Continual learning on dynamic graphs via parameter isolation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 601–611, 2023.

- [28] Altay Unal, Abdullah Akgül, Melih Kandemir, and Gozde Unal. Meta continual learning on graphs with experience replay. *Transactions on Machine Learning Research*, 2023.
- [29] Junwei Su, Difan Zou, Zijun Zhang, and Chuan Wu. Towards robust graph incremental learning on evolving graphs. In *International Conference on Machine Learning*, pages 32728–32748. PMLR, 2023.
- [30] Yixin Ren, Li Ke, Dong Li, Hui Xue, Zhao Li, and Shuigeng Zhou. Incremental graph classification by class prototype construction and augmentation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 2136–2145, 2023.
- [31] Chaoxi Niu, Guansong Pang, Ling Chen, and Bing Liu. Replay-and-forget-free graph class-incremental learning: A task profiling and prompting approach. *arXiv preprint arXiv:2410.10341*, 2024.
- [32] Dong Li, Aijia Zhang, Junqi Gao, and Biqing Qi. An efficient memory module for graph few-shot class-incremental learning. *arXiv preprint arXiv:2411.06659*, 2024.
- [33] Arnab Kumar Mondal, Jay Nandy, Manohar Kaul, and Mahesh Chandran. Stochastic experience-replay for graph continual learning. In *The Third Learning on Graphs Conference*, 2024.
- [34] Jiajun Liu, Wenjun Ke, Peng Wang, Ziyu Shang, Jinhua Gao, Guozheng Li, Ke Ji, and Yanhe Liu. Towards continual knowledge graph embedding via incremental distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8759–8768, 2024.
- [35] Xikun Zhang, Dongjin Song, and Dacheng Tao. Hierarchical prototype networks for continual graph representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4622–4636, 2022.
- [36] Yishi Xu, Yingxue Zhang, Wei Guo, Huifeng Guo, Ruiming Tang, and Mark Coates. Graphsail: Graph structure aware incremental learning for recommender systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2861–2868. ACM, October 2020.
- [37] Songlin Dong, Xiaopeng Hong, Xiaoyu Tao, Xinyuan Chang, Xing Wei, and Yihong Gong. Few-shot class-incremental learning via relation knowledge distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:1255–1263, 05 2021.
- [38] Xikun Zhang, Dongjin Song, and Dacheng Tao. Ricci curvature-based graph sparsification for continual graph representation learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
- [39] Kian Ahrabian, Yishi Xu, Yingxue Zhang, Jiapeng Wu, Yuening Wang, and Mark Coates. Structure aware experience replay for incremental learning in graph-based recommender systems. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2832–2836, 2021.
- [40] Xikun Zhang, Dongjin Song, and Dacheng Tao. Sparsified subgraph memory for continual graph representation learning. In *2022 IEEE International Conference on Data Mining (ICDM)*, pages 1335–1340. IEEE, 2022.
- [41] Xikun Zhang, Dongjin Song, Yixin Chen, and Dacheng Tao. Topology-aware embedding memory for continual learning on expanding networks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4326–4337, 2024.
- [42] Thanh Duc Hoang, Do Viet Tung, Duy-Hung Nguyen, Bao-Sinh Nguyen, Huy Hoang Nguyen, and Hung Le. Universal graph continual learning. *arXiv preprint arXiv:2308.13982*, 2023.
- [43] Xiaoxue Han, Zhuo Feng, and Yue Ning. A topology-aware graph coarsening framework for continual graph learning. *Advances in Neural Information Processing Systems*, 37:132491–132523, 2024.

- [44] Lei Song, Jiaxing Li, Qinghua Si, Shihan Guan, and Youyong Kong. Exploring rationale learning for continual graph learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 20540–20548, 2025.
- [45] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.
- [46] Alex Nichol and John Schulman. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2(3):4, 2018.
- [47] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.
- [48] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.
- [49] Gunshi Gupta, Karmesh Yadav, and Liam Paull. Look-ahead meta learning for continual learning. *Advances in Neural Information Processing Systems*, 33:11588–11598, 2020.
- [50] Khurram Javed and Martha White. Meta-learning representations for continual learning. *Advances in neural information processing systems*, 32, 2019.
- [51] Qiannan Zhang, Xiaodong Wu, Qiang Yang, Chuxu Zhang, and Xiangliang Zhang. Hg-meta: Graph meta-learning over heterogeneous graphs. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pages 397–405. SIAM, 2022.
- [52] Xikun Zhang, Dongjin Song, and Dacheng Tao. Continual learning on graphs: Challenges, solutions, and opportunities. *arXiv preprint arXiv:2402.11565*, 2024.
- [53] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018.
- [54] Yijun Tian, Shichao Pei, Xiangliang Zhang, Chuxu Zhang, and Nitesh V Chawla. Knowledge distillation on graphs: A survey. *arXiv preprint arXiv:2302.00219*, 2023.
- [55] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- [56] Carl Yang, Yuxin Xiao, Yu Zhang, Yizhou Sun, and Jiawei Han. Heterogeneous network representation learning: A unified framework with survey and benchmark. *IEEE Transactions on Knowledge and Data Engineering*, 34(10):4854–4873, 2020.
- [57] Jinhui Pang, Changqing Lin, Xiaoshuai Hao, Rong Yin, Zixuan Wang, Zhihui Zhang, Jinglin He, and Huang Tai Sheng. Ftf-er: Feature-topology fusion-based experience replay method for continual graph learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8336–8344, 2024.
- [58] Rich Caruana. Multitask learning. *Machine learning*, 28:41–75, 1997.

A Supplement Methodology

A.1 Detailed Algorithm

The detailed framework of the proposed method MKD for HCGL is described in Algorithm 1.

Algorithm 1 Framework of our MKD method

Require: Continual tasks $\mathcal{T} = \{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_T\}$; Buffer set $\{\mathcal{B}_\tau\}_{\tau \in \mathcal{A}}$; Learning rate α for the inner update; Learning rate β ; Number of target type nodes in each class added to \mathcal{B} : e .

Ensure: Model f_θ which can mitigate catastrophic forgetting of preceding tasks.

- 1: Initialize model parameters θ
- 2: **while** Continual Tasks \mathcal{T} remains **do**
- 3: Obtain current training set V_t^{tr}
- 4: Obtain teacher models $\{f_T\}$ and Buffer set $\{\mathcal{B}_\tau\}_{\tau \in \mathcal{A}}$
- 5: Select meta-learning examples $\mathcal{E} = Select(\mathcal{V}_i^{tr}, e)$
- 6: **for** $epoch = 1$ **to** E **do**
- 7: Update the parameters θ using \mathcal{E} via Equation 1
- 8: Replay the experience via Equation 7
- 9: Distill the soft knowledge and semantic information from teacher models via Equation 9 and 10
- 10: Update the parameters of student model by optimizing Equation 12
- 11: **end for**
- 12: Select target type nodes via 2 and other types nodes via 5
- 13: Add all sampled nodes to Buffer
- 14: Add current task model to teacher models $\{f_T\}$
- 15: $\mathcal{T} = \mathcal{T} \setminus \{\mathcal{T}_i\}$
- 16: **end while**
- 17: **return** Model f_θ

A.2 Extension to General HGNNs

Previously, we have discussed how to align semantic-level information via knowledge distillation in meta-path-guided attention networks. In this section, we further demonstrate that the proposed method can be easily extended to HGNNs without attention mechanisms (such as RGCN), enabling the preservation of semantic information from previous tasks.

For each pair of target-type nodes $v_i, v_j \in V_{\tau_t}$, we define the attention score e_{ij} based on their hidden representations from the penultimate layer of the network $h_i^{(L-1)}$ and $h_j^{(L-1)}$, using the weight matrix $W^{(L)}$ from the last layer:

$$e_{ij} = \left(h_i^{(L-1)} W^{(L)} \right)^\top \tanh \left(h_j^{(L-1)} W^{(L)} \right), \quad (14)$$

We then apply softmax normalization to the attention scores for each node v_i to obtain an attention vector:

$$\boldsymbol{\alpha}_i = \text{softmax}(e_{i1}, e_{i2}, \dots, e_{i|V_{\tau_t}|}) \quad (15)$$

We compute attention vectors for both the teacher model $\boldsymbol{\alpha}_i^{(T)}$ and the student model $\boldsymbol{\alpha}_i^{(S)}$, and define the semantic alignment loss as the L_2 distance between the two:

$$\mathcal{L}_{\text{sem}} = \sum_{i \in V_{\tau_t}} \left\| \boldsymbol{\alpha}_i^{(T)} - \boldsymbol{\alpha}_i^{(S)} \right\|_2 \quad (16)$$

In addition to the node similarity-based attention defined above, for edge-type-aware attention models such as HGT, attention is computed for each relation type $r \in \mathcal{R}$. Let $\boldsymbol{\beta}_r^{(T)}$ and $\boldsymbol{\beta}_r^{(S)}$ denote the attention vectors for relation r from the teacher and student models, respectively. The semantic alignment loss for such edge-type-based attention mechanisms is formulated as:

$$\mathcal{L}_{\text{sem}} = \sum_{r \in \mathcal{R}} \left\| \boldsymbol{\beta}_r^{(T)} - \boldsymbol{\beta}_r^{(S)} \right\|_2 \quad (17)$$

This formulation ensures that the student model preserves relation-aware structural semantics by aligning its attention distribution with that of the teacher model on each edge type.

B Supplemental experiment setups

B.1 Details of dataset

DBLP is a bibliographic dataset in the field of computer science. A widely adopted subset is used, containing four research areas, where nodes represent authors, papers, terms, and venues. IMDB is a dataset derived from a movie information platform. A subset including classes such as Action, Comedy, Drama, Romance, and Thriller is selected for use. Freebase is a large-scale knowledge graph. A subgraph comprising eight genres of entities and approximately one million edges is sampled following procedures outlined in prior studies. Yelp contains a network constructed from businesses, users, locations, and reviews. Although node features are not available, many business nodes are annotated with one or more labels from sixteen predefined categories.

Table 3: Details of datasets and continual learning tasks setting

<i>Node Classification</i>	#Nodes	#Node Types	#Edges	#Edge Types	#Classes
DBLP	26,128	4	239,566	6	4
IMDB	21,420	4	86,642	6	5
Freebase	180,098	8	1,057,688	36	7
Yelp	82,465	4	32,548,358	7	16

B.2 Baseline Introduction

We introduce the baseline models that we choose to compare in the following section:

Elastic weight consolidation (EWC) [21] is a technique that regularizes the loss function, such that the model is encouraged to only modify the weights that are less important for the previous tasks. This is achieved by penalizing changes to weights that have large importance for the previously learned tasks, thereby mitigating catastrophic forgetting when learning new tasks.

Memory aware synapses (MAS) [20] is a regularization-based method that measures the importance of parameters based on the sensitivity of predictions to these parameters. When learning a new task, the model adjusts the weights according to the network’s activations to update the important parameters relevant to the new task data.

Topology-aware weight preserving(TWP) [25] is a method introduced for graph continual learning that incorporates weight preservation mechanisms based on graph topology. By considering the local structure of the graph, TWP aims to overcome catastrophic forgetting. After computing the loss, the importance scores of the model weights are calculated according to the topology of the given graph, and the loss is regularized accordingly. This method leverages the properties of the graph.

ER-GNN [26] stores samples from previous tasks as experiences and replays them when learning new tasks. After learning task T_i , sample nodes are saved to the buffer using a selection function. When learning the next task T_{i+1} , separate graphs are constructed for each learned task $\{T_t\}_{t=1,\dots,i}$. Then, the GNN is trained using these graphs. The overall loss is calculated by regularizing the current task’s loss with the loss from the separately constructed graphs through experience replay.

MetaCLGraph [28] combines experience replay and meta-learning. In this method, the initial parameters of the model are calculated using the current task data. When learning a new task, stored samples from previous tasks are merged with the current task data to form a new graph, which is then used to update the model parameters.

FTF-ER [57] combines feature and global topological information by normalizing and weightedly integrating two types of node importance scores to evaluate node importance. Specifically, it introduces the Hodge Potential Score (HPS) module to capture global topological information. When learning a new task, a subgraph is induced using all the experience nodes stored in the buffer, the overall loss is calculated as the sum of the loss for the current task and the loss of the subgraph.

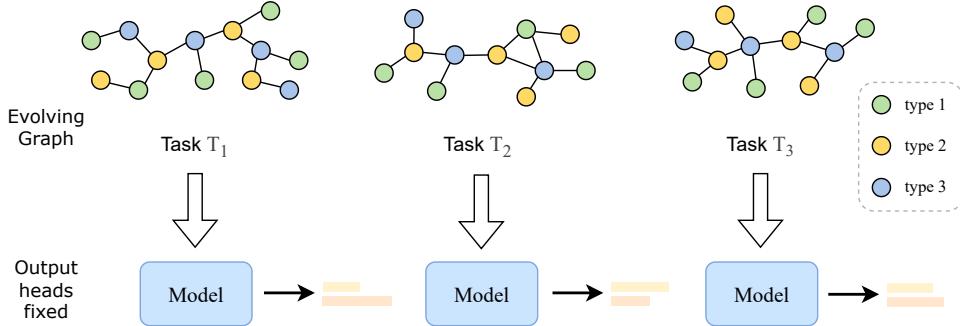


Figure 4: Illustration of the domain incremental setting.

B.3 Experimental Setting

Domain-Incremental Learning (Domain-IL) refers to a continual learning scenario where the task objective remains the same across tasks, but the data distribution (domain) shifts over time. This implies that the semantic meaning of the model’s output stays fixed. For example, in knowledge graph completion, each task may involve different sets of entities and relations, but the prediction goal—completing triplets—remains unchanged. Similarly, temporal data splits can form Domain-IL settings, where data from different time periods vary in distribution, yet the learning objective stays consistent.

B.4 Implementation Detail

To evaluate the effectiveness and generalizability of our method, we conduct experiments on three HGNNs backbones: HAN [6], MAGNN [4], and HGT [13]. Adam optimizer is used and the initial learning rate is set to 0.005 for all datasets. For HAN, each task is trained for 200 epochs with an early stopping patience of 100. For MAGNN, each task is trained for 100 epochs, with the early stopping patience set to 5 on DBLP and 10 on IMDB. For HGT, all tasks are trained for 300 epochs with a patience of 30. For datasets trained with mini-batches, the batch size is set to 8. We adopt coverage maximization as the selection function for all experience-replay-based continual learning methods. The buffer size is uniformly set to 50 for the target node type in experience replay baselines (e.g., ER-GNN, MetaCLGraph, and FTF-ER), and 20 under mini-batch settings. For non-target node types, the buffer size is set to 200. The hyperparameters for regularization-based baselines (e.g., EWC, MAS, and TWP) are set following the TWP official repository and the benchmark study [24].

Our meta-learning module is configured under a 10-shot setting (2-way or 3-way depending on the dataset). For our method’s hyperparameters, the knowledge distillation loss weight λ_{kd} is selected between 0.1 and 1.0 for different datasets, and the logit-level distillation weight λ_{logit} is selected between 0.6 and 1.5. We generally set the experience replay loss weight λ_{er} and the semantic-level distillation weight λ_{sem} to 1.0 and 10.0, respectively. The temperature in knowledge distillation is globally set to 1.0. All experiments are repeated five times with random seeds on Nvidia RTX A4000 and RTX 4090D GPUs, and we report the mean and standard deviation across all methods and datasets.

C Additional Results and Analysis

C.1 Additional Ablation Study

We further analyze the contribution of the two submodules in the HKD module—logit-level distillation and semantic-level distillation—to overall model performance, and assess the effectiveness of our proposed design. Figure 5 illustrates the impact of different distillation strategies—node-level only (“with node”), semantic-level only (“with sem”), and the combination of both (“with both”)—on the model’s Average Performance (AP) and Average Forgetting (AF) across four datasets (DBLP, IMDB, Freebase, Yelp). The results show that the combined strategy (“with both”) consistently achieves the highest AP and lowest AF across all datasets, demonstrating that logit-level and semantic-level

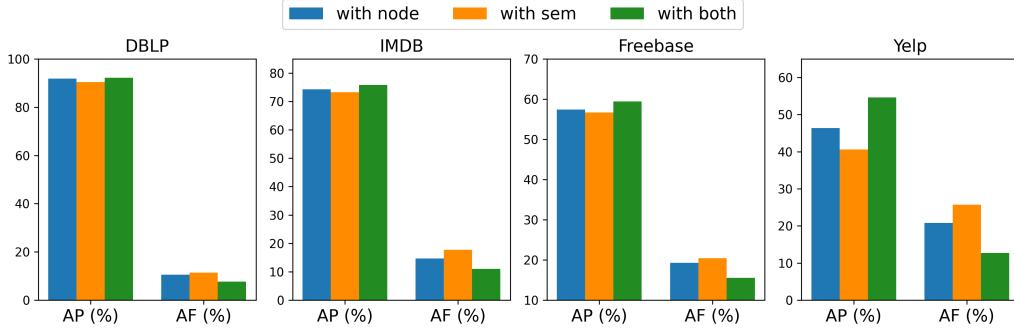


Figure 5: Average performance (AP) and average forgetting (AF) of the MKD method under three settings: with logit-level distillation, with semantic-level distillation, and with both.

distillation complement each other. Semantic-level distillation alone (“with sem”) exhibits weaker forgetting suppression on Freebase and Yelp, suggesting it may struggle to independently capture task-level semantic changes. On complex heterogeneous graphs such as Yelp, the synergy of both distillation levels yields especially significant performance improvements, indicating that addressing both structural and semantic forgetting is crucial.

C.2 Additional Hyperparameters Analysis

We further analyze the impact of the weight λ_{logit} used to balance the logit-level distillation loss in the Heterogeneity-aware Knowledge Distillation (HKD) module. Since λ_{sem} is typically set to 10 in our experiments, we do not discuss it here. Figure 6 shows the changes in AP (left) and AF

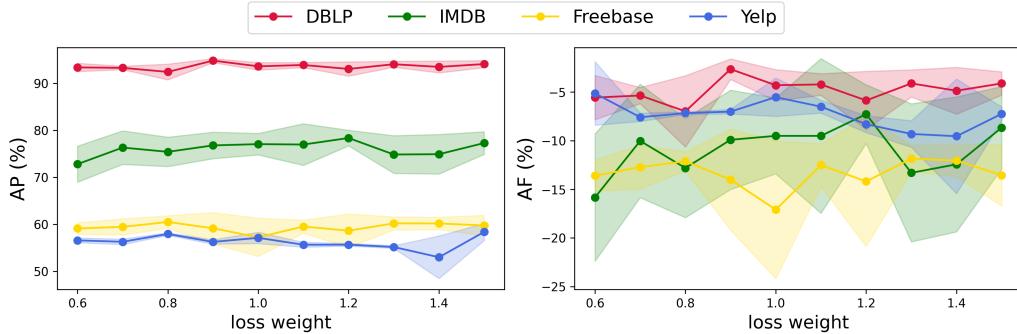


Figure 6: Sensitivity analysis of the logit-level distillation weight λ_{logit} in the HKD module. Left: Average Performance (AP); Right: Average Forgetting (AF).

(right) across datasets under varying logit-level distillation loss weights (x-axis). For most datasets, AP remains relatively stable, indicating robustness to this hyperparameter. IMDB and Yelp exhibit more fluctuation, likely due to significant distribution shifts across tasks, making them more sensitive to weight changes. The AF curves reveal that moderate weights (0.8 to 1.2) generally result in lower forgetting, while too high or too low weights cause an imbalance between new and old task focus. Yelp shows the smallest AF variation, suggesting that the HKD mechanism performs steadily in complex heterogeneous settings, aiding the model’s generalization ability.

C.3 Forgetting-aware Gap

AP (Average Performance) measures the mean test accuracy across all previously learned tasks, while AF (Average Forgetting) quantifies the performance drop on a specific task after learning subsequent ones. However, we observe that in order to retain knowledge from earlier tasks, the model often compromises its performance on the final task. To assess this trade-off, we introduce a new metric called Forgetting-aware Gap (FaG), defined as the difference between the test accuracy obtained by

Table 4: Forgetting-aware Gap (FaG) comparison on four datasets with HAN as the backbone model (averaged over 5 runs). Lower is better.

Method	DBLP	IMDB	Freebase	Yelp
Joint Train	0.69	2.47	11.59	11.58
EWC	2.15	2.78	10.17	1.62
MAS	0.40	0.97	9.53	2.09
TWP	1.03	2.27	10.44	5.58
ER-GNN	0.22	2.62	8.44	8.46
MetaCLGraph	1.36	2.87	13.68	13.61
FTF-ER	0.83	3.43	10.59	7.19
MKD (Ours)	0.63	1.31	9.90	6.88

training the model solely on the final task and the test accuracy on the same task after completing all tasks under the continual learning framework:

$$\text{FaG} = a_T^{\text{FT}} - a_T^{\text{CL}} \quad (18)$$

where a_T denotes the test accuracy on the final task \mathcal{T}_T , with “FT” referring to the finetuning baseline (i.e., training only on task \mathcal{T}_T), and “CL” referring to the accuracy obtained after training with a continual learning method. This gap reflects the performance degradation caused by the inherent plasticity-stability trade-off in continual learning.

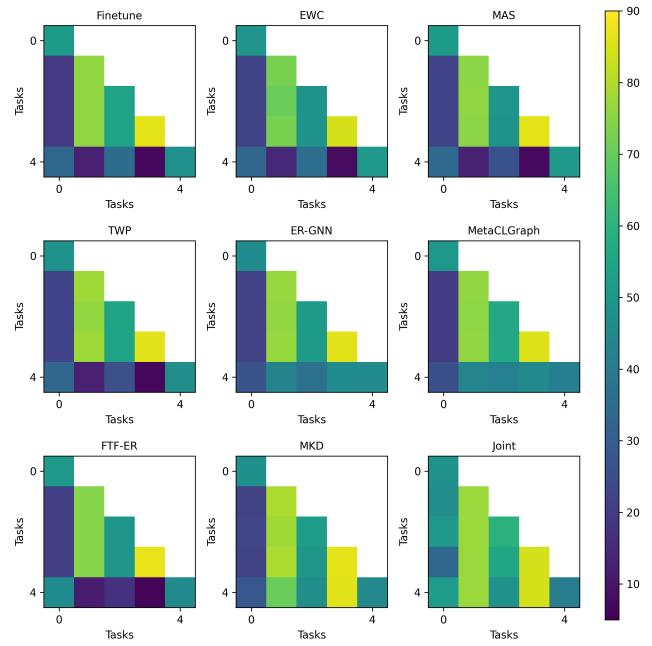
Results Analysis: The experimental results indicate that the Joint Train method shows consistently large Final-task Performance Gaps (FaG) across all benchmark datasets, suggesting that directly replaying all data severely impairs adaptability to new tasks. MetaCLGraph and FTF-ER also suffer high FaG values on Freebase and Yelp (13.68 and 10.59), reflecting performance degradation under distribution shifts—a manifestation of the stability-plasticity dilemma. ER-GNN achieves the lowest FaG (0.22) on DBLP, showing good adaptability, but its FaG sharply rises to 8.46 on Yelp, indicating that simple replay mechanisms struggle with complex heterogeneous structures. In contrast, our MKD method consistently maintains lower FaG values. Notably, it achieves 1.31 on IMDB and 6.88 on Yelp, outperforming experience-replay and regularization-based methods like TWP (2.27 on IMDB). These results validate that MKD, through its meta-learning and heterogeneity-aware distillation design, effectively mitigates final-task degradation while preserving task adaptability. Overall, FaG serves as a useful metric to quantify the trade-off between knowledge retention and adaptation, further supporting the effectiveness of MKD in heterogeneous continual learning.

D Limitations

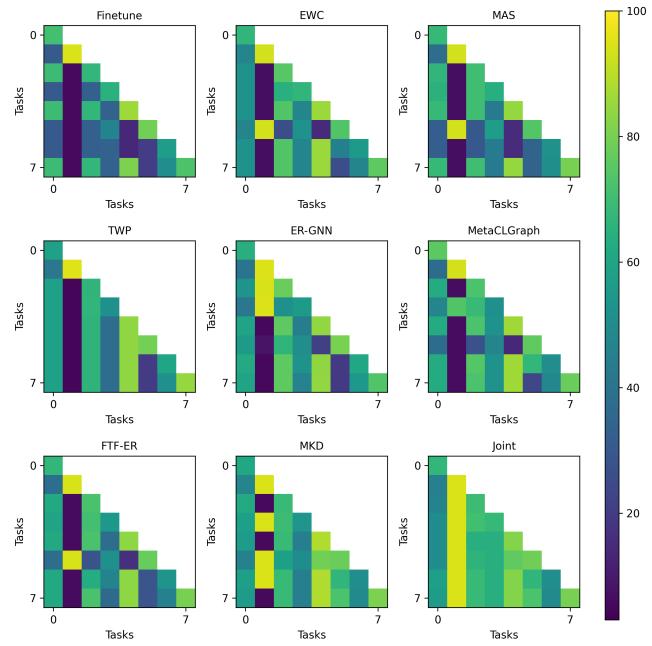
This work is confined to domain-incremental setting and primarily addresses the node classification task. In future work, we plan to extend our approach to other settings (e.g., class-incremental) and broader graph learning tasks, such as link prediction in recommendation systems.

E Broader Impacts

Heterogeneous graphs are widely used to model complex relational systems in critical domains such as biomedical networks, social platforms, and recommender systems. Our work on Heterogeneous Continual Graph Learning (HCGL) contributes to improving the robustness and adaptability of graph models under dynamic environments. This can benefit downstream applications such as real-time recommendation, dynamic knowledge graph completion, and biomedical discovery, where information constantly evolves. By enabling efficient knowledge retention and adaptation across tasks, our framework promotes safer and more reliable deployment of graph-based models in real-world systems.



(a) Yelp (3-way)



(b) Yelp (2-way)

Figure 7: Visualization: Accuracy matrices on the Yelp dataset under both 2-way (a) and 3-way (b) settings.