Modality-Independent Graph Neural Networks with Global Transformers for Multimodal Recommendation

Jun Hu[1], Bryan Hooi[1*], Bingsheng He[1], Yinwei Wei[2]

[1]School of Computing, National University of Singapore
[2]School of Software, Shandong University
jun.hu@nus.edu.sg, {bhooi, hebs}@comp.nus.edu.sg, weiyinwei@hotmail.com

Abstract

Multimodal recommendation systems can learn users' preferences from existing user-item interactions as well as the semantics of multimodal data associated with items. Many existing methods model this through a multimodal user-item graph, approaching multimodal recommendation as a graph learning task. Graph Neural Networks (GNNs) have shown promising performance in this domain. Prior research has capitalized on GNNs' capability to capture neighborhood information within certain receptive fields (typically denoted by the number of hops, K) to enrich user and item semantics. We observe that the optimal receptive fields for GNNs can vary across different modalities. In this paper, we propose GNNs with Modality-Independent Receptive Fields, which employ separate GNNs with independent receptive fields for different modalities to enhance performance. Our results indicate that the optimal K for certain modalities on specific datasets can be as low as 1 or 2, which may restrict the GNNs' capacity to capture global information. To address this, we introduce a Sampling-based Global Transformer, which utilizes uniform global sampling to effectively integrate global information for GNNs. We conduct comprehensive experiments that demonstrate the superiority of our approach over existing methods. Our code is publicly available at https://github.com/CrawlScript/MIG-GT.

Introduction

Recommendation systems predict user preferences by analyzing historical user-item interactions. Recently, deep learning has advanced the development of multimodal recommendation systems, which integrate rich multimodal data like texts and images alongside user-item interactions. Many existing studies (Zhou and Miao 2024; Sun et al. 2024) demonstrate that this utilization enables a richer, more comprehensive understanding of items, thereby enhancing the performance of recommendations. Multimodal recommendation systems have been widely used in applications such as e-commerce and micro-video platforms (Liu et al. 2023a; Shang et al. 2023; Cai et al. 2022; Liu et al. 2024).

In recent years, **GNN-based vertex representation learning** has emerged as a powerful technique in multimedia recommendation systems (Zhou et al. 2023a; Wu
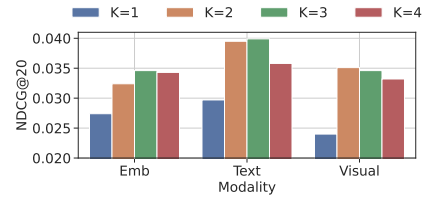
Figure 1: Performance of GNNs on Amazon Baby with features of different modalities at varying receptive fields (number of hops, K). "Emb" stands for learnable embeddings. The optimal K is modality-dependent: Emb and Text perform best at $K = 3$, while Visual performs best at $K = 2$.

et al. 2023b; Gao et al. 2023; Zhou et al. 2023b). These approaches use a graph to model the system, typically with graph vertices representing users and items, vertex features encapsulating multimodal data, and graph edges denoting user-item interactions. Building on this, recommendation is treated as a task of vertex representation learning. By employing GNNs for this purpose, the method effectively utilizes high-order interactions and multimodal data to derive low-dimensional embeddings of users and items. These embeddings are then used to compute similarities that reflect user preferences towards specific items, thereby enhancing the performance of the recommendation system.

To handle multimodal data, a typical and effective solution is to apply a separate GNN for each modality, and then pool the representations from these GNNs (Wei et al. 2019). At the feature level, it is obvious that input features of different modalities need to be encoded differently, e.g., using an image encoder for images. In this work, we show that the differences between modalities lies not just in how the features should be encoded, but also in *how each modality's information should be propagated over the graph*, i.e. the receptive field used for each modality. This issue is overlooked in the recent literature, which does not consider differences at the receptive field level across GNNs for different modalities.

Define the **receptive field of GNNs** as its number of hops (K): existing studies default to setting the same K for GNNs across all modalities. However, we observe that the optimal receptive field for GNNs differs across modalities. We conduct experiments by applying GNNs with different K to each modality's features. When vertices lack features for a modality (e.g., user vertices usually do not have text and
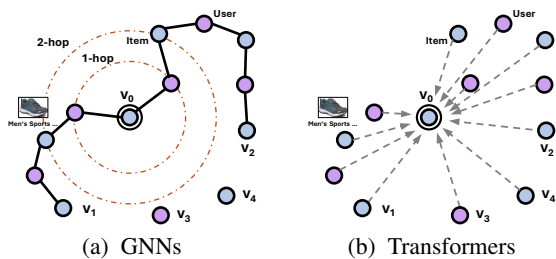
Figure 2: Examples of GNNs and Transformers.

visual features), we assign them zero vectors. Besides text and visual, the learnable embeddings are also treated as a modality. We conduct the experiment with a state-of-the-art (SOTA) model, MGDN (Hu et al. 2024) (which generalizes the widely used LightGCN (He et al. 2020)). Results on the Amazon Baby dataset (see Figure 1) show that the learnable embedding and text modalities perform best at $K = 3$, while the visual modality performs best at $K = 2$, highlighting the benefit of choosing different K for different modalities.

Another key consideration is the usage of global information. Our experiments (Experiments Section) indicate that the optimal K for certain modalities on specific datasets can be as low as 1 or 2, limiting GNNs' ability to capture global information from vertices across the whole graph. As shown in Figure 2a, when limiting the receptive fields of GNNs to 2 hops, the information of certain vertices, like v_1, v_2, v_3, and v_4 is missing for the target vertex v_0. Transformers (Vaswani et al. 2017) can potentially capture global information in graphs. Figure 2b shows that applying a Transformer to all vertices enriches the target vertex v_0 with information from all vertices. However, this is impractical, as Transformers typically require computing attention scores for every vertex pair, resulting in excessive time and space complexity. To address this, we introduce a Sampling-based Global Transformer, which utilizes uniform global sampling to effectively integrate global information for GNNs.

In this paper, we propose a framework named Modality-Independent Graph Neural Networks with Global Transformers (MIG-GT) for Multimodal Recommendation. It adopts modality-independent receptive fields to facilitate GNNs on multimodal graphs. Besides, to exploit the global information, we introduce a Sampling-based Global Transformer, which utilizes uniform global sampling to effectively integrate global information for GNNs.

We summarize our key contributions below:

- We propose GNNs with Modality-Independent Receptive Fields (MIRF), which apply separate GNNs with independent receptive fields (number of hops, K) on different modalities, to enhance performance. Our results indicate that the optimal K for certain modalities on specific datasets can be as low as 1, which limits the GNN's ability to capture global information.

- To better capture global information, we introduce a Sampling-based Global Transformer (SGT). This module leverages uniform global sampling to effectively incorporate global context into the learning process.

- We conduct comprehensive experiments that demonstrate the superiority of our method over baselines.

Related Work

Graph Neural Networks for Recommendation

Recent advances in GNNs have facilitated various social media research (Liu et al. 2023b; Zhu et al. 2023; Fang et al. 2023; Gao, Zhang, and Xu 2021; Han et al. 2021; Qian et al. 2023), with recommendation research being a representative case. GNN-based methods conceptualize user-item interactions as bipartite graphs, using GNNs to learn embeddings that reflect user preferences. GCMC (van den Berg, Kipf, and Welling 2017) employs Graph Convolutional Networks (GCNs) to build an autoencoder for recommendation. PinSage (Ying et al. 2018) uses GNNs with sampling for large datasets. NGCF (Wang et al. 2019) designs GNNs to enhance interactive signal capture between vertices and neighborhoods. LightGCN (He et al. 2020) optimizes GNNs by forgoing transformations and activation functions within GCN layers, streamlining message passing for effective embedding generation. UltraGCN (Mao et al. 2021) presents a paradigm shift by eschewing explicit GNN operations in favor of a constraint-based loss function. ApeGNN (Zhang et al. 2023) adaptively aggregates information based on local structures, capturing diverse patterns. MGDN (Hu et al. 2024) can generalize LightGCN and offer flexible controls to balance self and neighbor information.

Graph Transformers

Transformers can capture global information but suffer from quadratic complexity w.r.t. vertex count, making them inefficient for large graphs in recommendation tasks. Recent works, such as SGFormer (Wu et al. 2023a) and Polynormer (Deng, Yue, and Zhang 2024), address this by removing softmax normalization, reducing complexity to linear. Both methods combine Graph Transformer outputs with GNN models, differing in their fusion strategies.

Multimodal Recommendation

Initial advancements (He and McAuley 2016b; Liu, Wu, and Wang 2017) enhanced the Bayesian Personalized Ranking (BPR) (Rendle et al. 2009) by combining learnable embeddings and visual features. VECF(Chen et al. 2019) uses VGG (Simonyan and Zisserman 2015) for image pre-processing and region-specific attention for item visual features. The advent of GNNs has facilitated multimodal user/item representation learning. MMGCN (Wei et al. 2019) builds modality-aware graphs and applies separate GNNs to learn modality-specific features, which are then aggregated. GRCN (Wei et al. 2020) progresses this concept by refining user-item graph structures, sieving out misleading connections. DualGNN (Wang et al. 2023) introduces a user co-occurrence graph with a feature preference module to capture multimodal item feature dynamics.

Previous methods typically use user-item interaction graphs, capturing item relationships implicitly, while some explicitly model item relationships. LATTICE (Zhang et al. 2021) performs modality-aware structure learning, obtaining item-item structures separately for each modality and then combining them. FREEDOM (Zhou and Shen 2023)

simplifies the process by freezing the item-item graph structure and denoising the user-item interaction graph.

Most existing approaches focus on denoising or explicitly modeling item-item relations. Different from them, this paper addresses the task from the perspective of modality-independent receptive fields and global information, showing the necessity of these elements for multimodal recommendation. Even without the denoising and explicit modeling of item-item relations, our model already outperforms or matches the performance of SOTA models.

Preliminary

Problem Definition

Let U and V denote the sets of users and items, with sizes $|U|$ and $|V|$, respectively. We denote the i-th user as u_i and the j-th item as v_j. Each item in V is associated with multimodal data, including text and an image. A user-item interaction matrix $B \in \mathbb{R}^{|U| \times |V|}$ represents observed interactions, where B_{ij} is set to 1 or 0 to denote whether the interaction between u_i and v_j has been observed. The task is to predict unobserved user preferences over items. Our model learns a d-dimensional vector ($d \ll |U| + |V|$) as the representation for each user/item vertex and uses the dot product between user and item representations to reflect preference scores, with higher scores indicating higher preferences.

Multimodal User-Item Graph

In this paper, we adopt the method of modeling users and items within a single homogeneous graph. Specifically, the users and items are modeled as vertices of a user-item graph $\mathcal{G}$, consisting of $|N| = |U| + |V|$ vertices, where the first $|U|$ vertices represent users and the subsequent $|V|$ vertices represent items. The observed user-item interactions are modeled as edges in the graph, thus, the adjacency matrix of the graph $A \in \{0,1\}^{|N| \times |N|}$ is defined as follows:

$$A = \begin{pmatrix} 0 & B \\ B' & 0 \end{pmatrix} \tag{1}$$

where B' denotes the transpose of B. For multimodal data, each item vertex has a text feature vector and a visual feature vector extracted via pre-trained models. Each user vertex is assigned a d-dimensional learnable embedding.

Method

We present our framework, Modality-Independent Graph Neural Networks with Global Transformers (MIG-GT).

Overall Framework

Figure 3 provides an overview of our framework. The upper-left part shows the input graph with users and items as vertices and user-item interactions as edges. Each item is associated multimodal data, including text and an image. Note that our method is applied directly to the original user-item interaction graph without constructing new graphs.

Our proposed framework mainly consists of two components: (1) **Modality-Independent Receptive Fields (MIRF)** applies separate GNNs with independent receptive fields for data of different modalities in the graph. For each

GNN, first an MLP is used to encode vertex features of the corresponding modality into d-dimensional feature vectors, which is a common operation in existing works. (Wei et al. 2019) Then we perform message propagation with the encoded features, and different from existing work, we propose to use modality-independent receptive fields for different modalities. The independent receptive fields for different modalities, learnable embedding, text, and visual, denoted as $K^{(E)}$, $K^{(T)}$, and $K^{(V)}$ are selected based on validation set performance. (2) **Sampling-based Global Transformer (SGT)** is designed to exploit global information. SGT performs self-attention using a few vertices uniformly sampled from the entire graph to enrich vertex representations. Unlike typical Transformers, which compute attention scores for every pair of vertices—resulting in significant time and space complexity—SGT only needs to compute the attention scores between each vertex and the few sampled vertices, thereby enhancing efficiency. We also propose a Transformer Unsmooth Regularization (TUR) for optimization.

Modality-Independent Receptive Fields

We apply separate GNNs to features from different modalities. Each GNN follows an encode-then-propagate framework (Klicpera, Bojchevski, and Günnemann 2019), using an MLP to encode vertex features into d-dimensional vectors for message propagation on the graph. ($d \ll |N|$ is the dimensionality of the final vertex representations.) For message propagation, we utilize MGDN, which propagates vertex features without feature transformations.

Specifically, we use $X^{(M)} \in \mathbb{R}^{|N| \times d^{(M)}}$ to denote the raw vertex feature matrix of modality M, where $|N|$ is the number of vertices, and $d^{(M)}$ is the feature dimension specific to modality M. $X^{(E)}$, $X^{(T)}$, and $X^{(V)}$ represent the learnable embedding, text, and visual modalities, respectively. The encoded feature vectors by the MLP are denoted as:

$$\tilde{X}^{(M)} = \text{MLP}(X^{(M)}) \in \mathbb{R}^{|N| \times d} \tag{2}$$

One exception is the learnable embedding modality, where $X^{(E)} \in \mathbb{R}^{|N| \times d}$ can be directly optimized. Therefore, it does not require an extra MLP, and $\tilde{X}^{(E)} = X^{(E)}$. Note that treating each modality independently may cause certain vertices to have missing features. For example, when dealing with either text or visual modalities, the user vertices may lack features under certain modalities. In such cases, we simply assign zero vectors as the encoded feature vectors for these featureless vertices (in $\tilde{X}^{(M)}$), with their dimensionality matching that of other vertices' features.

For message propagation, we use MGDN separately for each modality with different receptive fields $K^{(M)}$ for each. We first compute a normalized adjacency matrix for MGDN, shared across all modalities, $\hat{A} = \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}}$, where $\tilde{A} = A + I$ and $\tilde{D}_{ii} = \sum_{j=0}^{N} \tilde{A}_{ij}$ is the degree matrix of $\tilde{A}$. With it, MGDN learns vertex representations by incorporating neighbor information within $K^{(M)}$ hops as follows:

$$Z^{(M)} = f_{MGDN}(\tilde{X}^{(M)}, A)$$
$$= (\beta^K \hat{A}^{K^{(M)}} + \sum_{k=0}^{K^{(M)}-1} \alpha \beta^k \hat{A}^k) \tilde{X}^{(M)} / \Gamma \tag{3}$$
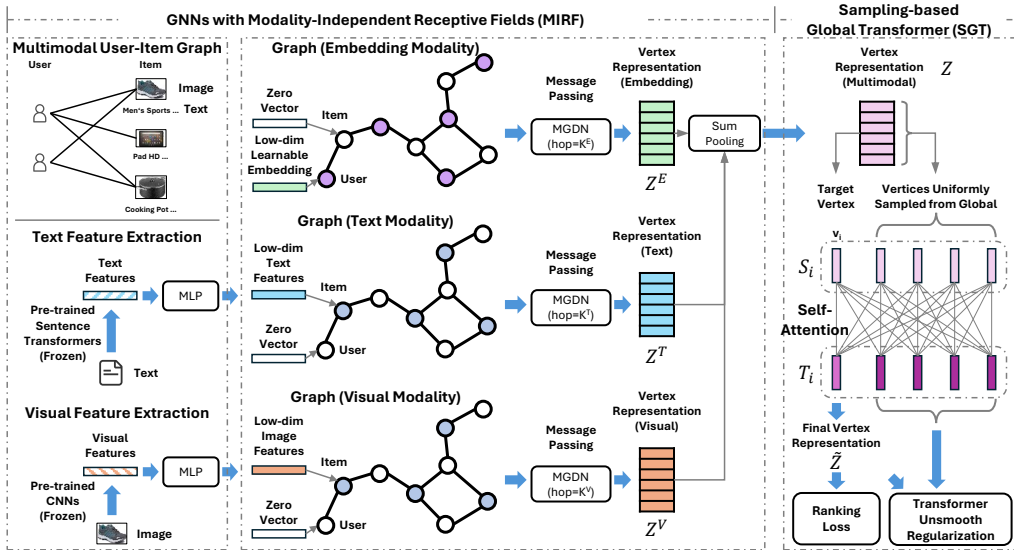
Figure 3: Overall Framework of Modality-Independent Graph Neural Networks with Global Transformers (MIG-GT).

where Γ is used for normalization to ensure the sum of coefficients of $\hat{A}^k \tilde{X}^{(M)}$ in Equation 3 is 1.0:

$$\Gamma = \beta^{K^{(M)}} + \sum_{k=0}^{K^{(M)}-1} \alpha\beta^k \tag{4}$$

For efficiency, $Z^{(M)}$ is computed in a step-wise manner:

$$H^{(M,0)} = \tilde{X}^{(M)} \tag{5}$$

$$H^{(M,k)} = \beta\hat{A}H^{(M,k-1)} + \alpha H^{(M,0)} \tag{6}$$

$$Z^{(M)} = H^{(M,K^{(M)})}/\Gamma \tag{7}$$

In each iteration, the updated vertex representations are formed by incorporating both the propagation result, $\hat{A}H^{(M,k-1)}$, and the input embeddings, $H^{(M,0)} = \tilde{X}^{(M)}$. The relative importance of these components is modulated by the hyperparameters β and α.

After obtaining modality-independent vertex representations $Z^{(E)}$, $Z^{(T)}$, and $Z^{(V)}$, we perform sum-pooling to obtain the multimodal vertex representations $Z \in \mathbb{R}^{|N|\times d}$:

$$Z = Z^{(E)} + Z^{(T)} + Z^{(V)} \tag{8}$$

We define z_i to denote the i-th row in Z, representing the multimodal representation of the i-th vertex.

For selecting modality-independent receptive fields $K^{(M)}$ ($K^{(E)}$, $K^{(T)}$, and $K^{(V)}$), we show in the experiments section that grid search on validation datasets is feasible, with the impact of these hyperparameters on validation datasets being almost consistent with that on test datasets.

Sampling-Based Global Transformer

Typical Transformers require computing the attention scores between every pair of samples, thus making it impractical to apply them directly to graphs. To alleviate this, for each vertex z_i, we uniformly sample C vertex representations from Z, and apply Transformers only on the current vertex and the sampled vertices. Note that, for each vertex, C vertices are sampled independently at each training step. Therefore, throughout the training, each vertex is likely to be sampled

along with many vertices uniformly sampled from the global set, capturing more global information. This global information will then be learned by the MLP encoder of the GNNs and the learnable embeddings.

Formally, given z_i, we construct a matrix $S_i \in \mathbb{R}^{(C+1)\times d}$, where the first row $s_{i1} = z_i$ represents the representation of the i-th vertex. Each subsequent row s_{ij} (for $1 < j \leq C+1$) is assigned a vertex representation z_k from Z denoted as $s_{ij} = z_k$, and $k \sim \mathrm{Uniform}(1, |N|)$.

We apply a simplified Transformer on S_i to perform self-attention to enrich the semantics of the vertex representations in S_i, obtaining $T_i \in \mathbb{R}^{(C+1)\times d}$:

$$T_i = (1-\gamma)\mathrm{softmax}\left(\frac{\mathcal{Q}\mathcal{K}^{\top}}{\sqrt{d}}\right)\mathcal{V} + \gamma S_i \tag{9}$$

where $\mathrm{softmax}$ denotes the row-wise softmax normalization, and $0 \leq \gamma \leq 1$ is a hyperparameter that modifies Transformer's residual connection to offer more flexibility, allowing for adjustable integration of the Transformer's output with the original input. $\mathcal{Q} = S_iW^{(\mathcal{Q})}$, $\mathcal{K} = S_iW^{(\mathcal{K})}$, and $\mathcal{V} = S_i$, where $W^{(\mathcal{Q})} \in \mathbb{R}^{d\times d^{(\mathrm{att})}}$ and $W^{(\mathcal{K})} \in \mathbb{R}^{d\times d^{(\mathrm{att})}}$ are learnable parameters, and $d^{(\mathrm{att})}$ is a hyperparameter.

After self-attention, we extract the first row of T_i as the final representation of the i-th vertex. We denote the final vertex representation matrix as $\tilde{Z} \in \mathbb{R}^{|N|\times d}$, where $\tilde{z}_i$ represents the i-th vertex, thus $\tilde{z}_i = T_{i1}$. Note that the remaining rows $\{T_{ij}|j > 1\}$ are not taken as the final vertex representations but are used for regularization, introduced later. Besides, we use $\tilde{u}_i = \tilde{z}_i$ and $\tilde{v}_j = \tilde{z}_{j+|U|}$ to denote the final vertex representation of the i-th user and j-th item.

Transformer Unsmooth Regularization Our method samples vertices globally and uses the Transformer to enrich them by extracting complementary information from each other. This process may cause an issue known as smoothing, making it difficult to distinguish the representations of certain vertices. Given the i-th vertex and the corresponding output of the Transformer, T_i, it contains the enriched representation of not only the i-th vertex but also C uniformly sampled vertices. Since they complement each other

via self-attention, their representations are likely to become smooth and indistinguishable from one another. Our un-smooth regularization aims to ensure our model can still distinguish between them. This regularization relies on edges in graphs. Sampling one of the i-th vertex's neighbors, denoted as the k-th vertex (that is, $A_{ik} = 1$), we take its final representation $\tilde{z}_k$ and apply the regularization as follows:

$$\mathcal{L}_{TUR} = -\sum_{A_{ik}=1} \log\left(\frac{\exp(\tilde{z}'_k T_{i1})}{\sum_{j=1}^{C+1} \exp(\tilde{z}'_k T_{ij})}\right) \quad (10)$$

Our experiments show that with a small number of vertices sampled from the global graph ($C \leq 20$), our model can already achieve significant performance improvements. Note that, unlike a typical Transformer, which requires computing attention scores between each vertex and all vertices in the graph—resulting in significant time and space complexity—our method computes attention scores only between each vertex and the C sampled vertices, where C is very small, thereby enhancing efficiency.

Model Optimization

We optimize a combined loss with Adam optimizer (Kingma and Ba 2015), where the combined loss is as follows:

$$\mathcal{L}_{rec} = \mathcal{L}_{rank}(\tilde{Z}) + \mathcal{L}_{TUR}(\tilde{Z}, T) + \Psi_{L2}\mathcal{L}_{L2}(\tilde{Z}) \quad (11)$$

where $\mathcal{L}_{rank}(\tilde{Z})$ is the ranking loss, and $\mathcal{L}_{L2}(\tilde{Z})$ is the L2 regularization with coefficient Ψ_{L2}. For the ranking loss, we adopt the popular BPR loss (Rendle et al. 2009):

$$\mathcal{L}_{BPR} = -\sum_{B_{ij}=1} \mathbb{E}_{v_k \sim p(v)} \log \sigma(\tilde{u}'_i \tilde{v}_j - \tilde{u}'_i \tilde{v}_k) \quad (12)$$

where σ denotes the sigmoid activation function, and $v_k \sim p(v)$ represents a vertex randomly sampled from the graph.

Experiments

We conduct experiments on three public datasets using a Linux system with two Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz processors, 128GB of RAM, and a GeForce GTX 1080 Ti GPU (11GB). The model is implemented via PyTorch [1] and DGL (Wang et al. 2020), with our code included in the supplementary materials.

Datasets

Following prior work (Zhang et al. 2021; Zhou and Shen 2023), we conduct our studies on three public datasets from the Amazon review datasets (He and McAuley 2016a), specifically abbreviated as Baby, Sports, and Clothing, respectively. These datasets provide multimodal data (textual and visual) for the items and vary in item count per category. We utilize the preprocessed data from previous studies, where the raw data for each category was filtered using a 5-core threshold for both products and users. Regarding multimodal features, we adopt the text and visual embeddings extracted and published by prior work (Zhou and Shen 2023), with visual features as 4,096-dimensional embeddings obtained from pre-trained Convolutional Neural Networks, and

---

[1] https://pytorch.org/

Dataset	Users	Items	Interactions	Sparsity
Baby	19,445	7,050	160,792	99.88%
Sports	35,598	18,357	296,337	99.95%
Clothing	39,387	23,033	278,677	99.97%

Table 1: Statistics of the three datasets.

text features as 384-dimensional embeddings from sentence-transformers, derived from item titles, descriptions, categories, and brands. Dataset statistics are shown in Table 1.

Baselines

All baselines utilize BPR as the ranking loss. The first set of baselines comprises models that rely solely on user-item interactions and do not incorporate multimodal data: **MF** (Koren, Bell, and Volinsky 2009), **LightGCN** (He et al. 2020), **ApeGNN** (Zhang et al. 2023), and **MGDN** (Hu et al. 2024). The second set includes multimodal recommendation models that leverage both user-item interactions and multimodal data: **VBPR** (He and McAuley 2016b), **MMGCN** (Wei et al. 2019), **GRCN** (Wei et al. 2020), **DualGNN** (Wang et al. 2023), **SLMRec** (Tao et al. 2023), **LATTICE** (Zhang et al. 2021), and **FREEDOM** (Zhou and Shen 2023).

Model Evaluation and Parameter Settings

To ensure a fair comparison, we adopt the evaluation settings from previous studies (Tao et al. 2023; Wang et al. 2023; Zhang et al. 2021; Zhou and Shen 2023). Our evaluation criteria include two widely-used metrics, abbreviated as R for recall and N for Normalized Discounted Cumulative Gain (NDCG). We report these metrics for top 10 and top 20 recommendations, denoted as R@10, R@20, N@10, and N@20. Regarding data split, we allocated 80% of known user interactions for training, 10% for validation, and the remaining 10% for testing. The reported performance is the mean results obtained using five different random seeds.

In terms of hyperparameters, we tune each hyperparameter within a range, and take the combination achieving best performance in the validation set, reporting the corresponding test performance. For the independent receptive fields, we search number $K^{(M)} \leq 4$. For the γ in the transformer, we search within 0.8 and 0.9. The learning rate and L2 regularization coefficient are searched from $\{1 \times 10^{-2}, 1 \times 10^{-3}\}$ and $\{1 \times 10^{-4}, 1 \times 10^{-5}\}$, respectively.

Performance Analysis

We report the performance of baseline methods and our method in Table 2. Some baseline results are directly cited from the original literature, while others, which may slightly differ, are derived from running the provided source code without modifications. The column 'Multimodal' indicates whether the method utilizes multimodal data, and the column 'GNN' denotes whether it is GNN-based. From the reported performance, we make the following observations:

- The methods exploiting multimodal data generally demonstrate an advantage over those that do not utilize multimodal data. Notably, VBPR, which extends MF via multimodal features, shows a significant improvement in performance, underscoring the importance of leveraging multimodal data in recommendation systems.

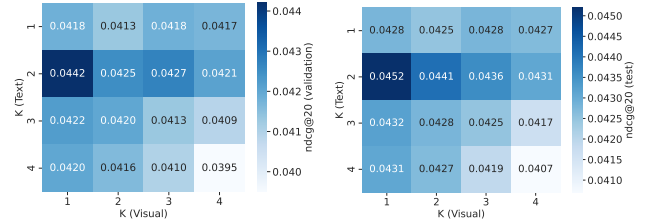|			Baby				Sports				Clothing			
Method	Multimodal	GNN	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20
MF	✗	✗	0.0357	0.0575	0.0192	0.0249	0.0432	0.0653	0.0241	0.0298	0.0206	0.0303	0.0114	0.0138
LightGCN	✗	✓	0.0479	0.0754	0.0257	0.0328	0.0569	0.0864	0.0311	0.0387	0.0361	0.0544	0.0197	0.0243
ApeGNN	✗	✓	0.0501	0.0775	0.0267	0.0338	0.0608	0.0892	0.0333	0.0407	0.0378	0.0538	0.0204	0.0244
MGDN	✗	✓	0.0495	0.0783	0.0272	0.0346	0.0614	0.0932	0.0340	0.0422	0.0362	0.0551	0.0199	0.0247
VBPR	✓	✗	0.0423	0.0663	0.0223	0.0284	0.0558	0.0856	0.0307	0.0384	0.0281	0.0415	0.0158	0.0192
MMGCN	✓	✓	0.0421	0.0660	0.0220	0.0282	0.0401	0.0636	0.0209	0.0270	0.0227	0.0361	0.0154	0.0154
GRCN	✓	✓	0.0532	0.0824	0.0282	0.0358	0.0599	0.0919	0.0330	0.0413	0.0421	0.0657	0.0224	0.0284
DualGNN	✓	✓	0.0513	0.0803	0.0278	0.0352	0.0588	0.0899	0.0324	0.0404	0.0452	0.0675	0.0242	0.0298
SLMRec	✓	✓	0.0521	0.0772	0.0289	0.0354	0.0663	0.0990	0.0365	0.0450	0.0442	0.0659	0.0241	0.0296
LATTICE	✓	✓	0.0547	0.0850	0.0292	0.0370	0.0620	0.0953	0.0335	0.0421	0.0492	0.0733	0.0268	0.0330
FREEDOM	✓	✓	0.0627	0.0992	0.0330	0.0424	0.0717	0.1089	0.0385	0.0481	0.0626	0.0932	0.0338	0.0416
MIG-GT	✓	✓	**0.0665**	**0.1021**	**0.0361**	**0.0452**	**0.0753**	**0.1130**	**0.0414**	**0.0511**	**0.0636**	**0.0934**	**0.0347**	**0.0422**
Improv.			6.06%	2.92%	9.39%	6.6%	5.02%	3.76%	7.53%	6.24%	1.6%	0.21%	2.66%	1.44%

Table 2: Performance comparison of different recommendation models.

- GNN-based methods consistently outperform non-GNN methods (MF and VBPR), with or without multimodal data, highlighting the efficacy of GNNs in this domain.

- MMGCN is an early and typical study on applying GNNs for multimodal recommendation, proposing to apply different GNNs to different modalities separately. Compared to MMGCN, GRCN, DualGNN, and SLMRec consider more nuanced properties, like noisy user-item interactions, resulting in better performance. This shows that there is room to improve GNNs based on the specific properties in multimodal recommendation systems.

- Besides explicit user-item interactions, LATTICE and FREEDOM learn implicit item-item relationships and build graphs to explicitly utilize them. FREEDOM also introduces a denoising mechanism for the learned item-item relations. Both models outperform most other baselines, demonstrating that explicitly modeling item-item relationships is an effective alternative approach.

- MIG-GT outperforms the SOTA baseline (FREEDOM) with an improvement of around 5% on two datasets (Baby and Sports) and slightly outperforms FREEDOM on Clothing. Note that, unlike other GNN baselines, MIG-GT does not rely on commonly used components like denoising and explicit modeling of item-item relations. It relies solely on our modality-independent receptive fields and a sampling-based Global Transformer, yet it still outperforms or matches the performance of baselines, showing the effectiveness of our method.
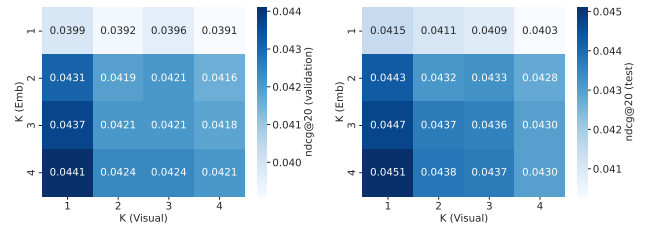
Detailed Analysis

Impact and Selection of Modality-Independent Receptive Fields Firstly, we examine the impact of modality-independent receptive fields, denoted as $K^{(M)}$, which include $K^{(E)}$, $K^{(T)}$, and $K^{(V)}$. Given the extensive number of possible combinations, we present a subset of these on the Amazon Baby dataset for brevity. In each figure, we fix $K^{(M)}$ of one modality to the optimal value and vary the others from 1 to 4, resulting in 16 combinations. We report both validation and test performance for each combination.

To clearly visualize the impact of MIRFs, we use several heatmaps in Figure 4 and Figure 5, with the x-axis and y-axis representing the numbers of two modality receptive fields, respectively. Each cell in the heatmap displays the performance (ndcg@20), highlighting performance variations across different combinations. In Figure 4, we fix $K^{(E)} = 4$

(a) Performance on Valid Set. (b) Performance on Test Set.

Figure 4: Heatmaps showing the NDCG@20 scores for different combinations of $K^{(T)}$ and $K^{(V)}$.

(a) Performance on Valid Set. (b) Performance on Test Set.

Figure 5: Heatmaps showing the NDCG@20 scores for different combinations of $K^{(E)}$ and $K^{(V)}$.

and vary $K^{(T)}$ and $K^{(V)}$. Another set of heatmaps in Figure 5 maintains $K^{(T)} = 2$ while varying $K^{(E)}$ and $K^{(V)}$. The results indicate that different combinations yield different performances, with the optimal MIRF configuration (which diverges from the identity receptive field setting across all modalities) achieving the best performance.

Selection of MIRFs. Furthermore, we demonstrate the feasibility of using grid search on the validation dataset for selecting MIRFs. Figure 4a and Figure 5a depict the heatmaps generated from the validation dataset, while Figure 4b and Figure 5b correspond to the test dataset. Comparing the two sets of heatmaps allows us to assess the consistency of the hyperparameters' impact between the validation and test datasets. Our findings confirm that the patterns observed during validation are generally representative of the test phase, validating grid search as a viable method for selecting independent receptive fields for different modalities.

Impact of Sampling-Based Global Transformers We perform ablation tests with a variant of our model, MIG, which removes SGT and retains only the MIRF components. We compare our model against the variant in Figure 6, in-
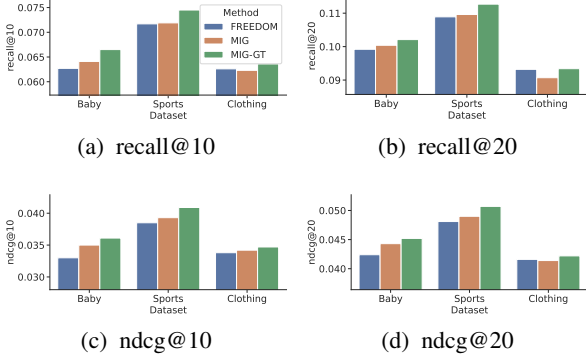
(a) recall@10   (b) recall@20

(c) ndcg@10   (d) ndcg@20

Figure 6: Impact of Sampling-based Global Transformers.

	Baby		Sports		Clothing	
Method	R@20	N@20	R@20	N@20	R@20	N@20
MIG-SGFormer	0.0863	0.0376	0.0887	0.0392	0.0827	0.0363
MIG-Polynormer	0.0997	0.0436	0.1048	0.0461	0.0864	0.0386
MIG-GT	**0.1021**	**0.0452**	**0.1130**	**0.0511**	**0.0934**	**0.0422**

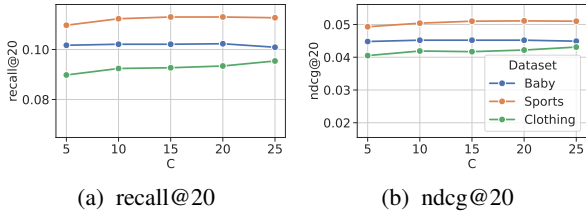Table 3: Impact of Different Global Transformers.

(a) recall@20   (b) ndcg@20

Figure 7: Impact of Number of Global Samples (C) for SGT.

cluding the performance of the SOTA method FREEDOM for better comparison, to show the effectiveness of SGT. The results show that MIG, even without SGT, already outperforms FREEDOM on Baby and Sports. With SGT, MIG-GT further enhances MIG's performance. On Clothing, although MIG is slightly outperformed by FREEDOM, with SGT, MIG-GT enhances it to achieve better performance than FREEDOM, showing the effectiveness of SGT.

Additionally, we replace SGT with existing Graph Transformer methods, SGFormer and Polynormer, to construct variants MIG-SGFormer and MIG-Polynormer, and compare them with MIG-GT in Table 3. The results show that MIG-GT outperforms these variants, demonstrating the effectiveness of our sampling-based approach for developing Global Transformers in recommendation contexts.

Impact of Number of Global Samples for SGT To investigate the impact of the number of global samples C for SGT, we vary it from 5 to 25 and report the corresponding performance. Results in Figure 7 show that when increasing from 5 to 10, we observe a performance improvement across datasets. Further increasing it beyond 10, performance increments can be observed on certain datasets. Overall, this demonstrates that with only 10 or 20 global samples, our SGT can significantly improve performance.

Training Efficiency of MIG-GT To demonstrate the training efficiency of our method, we visualize test perfor-
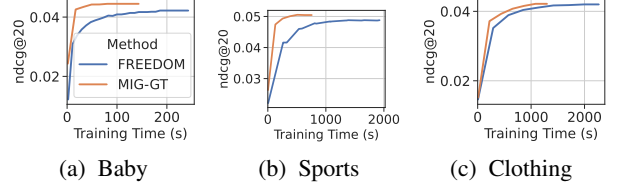

(a) Baby   (b) Sports   (c) Clothing

Figure 8: Test performance (ndcg@20) during training.

	Baby		Sports		Clothing	
Method	R@20	N@20	R@20	N@20	R@20	N@20
MMSSL	0.0971	0.0420	0.1013	0.0474	0.0797	0.0359
MGCN	0.0964	0.0427	0.1106	0.0496	0.0945	**0.0428**
LGMRec	0.1002	0.0440	0.1068	0.0480	0.0828	0.0371
MIG-GT	0.1021	**0.0452**	**0.1130**	**0.0511**	0.0934	0.0422
MIG-GT-CL	**0.1022**	0.0451	0.1120	0.0505	**0.0946**	**0.0428**

Table 4: Comparison with CL-Based Methods.

mance (ndcg@20) against training time in seconds during training in Figure 8. The choice to report training time instead of epochs is deliberate. In recommendation tasks, the definition of an 'epoch' can vary, being defined as a full iteration over either users (vertices) or user-item interactions (edges). Thus, training time serves as a more consistent and comparable measure of efficiency across different methods.

When compared with FREEDOM – known for its efficiency – our method consistently achieves higher performance more rapidly across all tested graphs by avoiding a complex denoising mechanism over item-item relations. On Baby and Sports, our method surpasses FREEDOM's final performance at early stages. On Clothing, while our final performance matches FREEDOM's, our method converges faster and reaches the optimal results earlier.

Comparison with Contrastive Learning (CL)-Based Methods Another research direction focuses on improving GNNs via CL. We build MIG-GT-CL, integrating our model with the typical CL loss, InfoNCE (van den Oord, Li, and Vinyals 2019), and compare it against CL-based methods, MMSSL (Wei et al. 2023), MGCN (Yu et al. 2023), and LGMRec (Guo et al. 2024). Results in Table 4 show that MIG-GT already surpasses most baselines and outperforms all with the simple addition of the CL loss (MIG-GT-CL).

Conclusions

In this study, we explored GNNs for multimodal recommendation systems. We observe that optimal receptive fields for GNNs can vary across different modalities. To capitalize on this, we introduced GNNs with Modality-Independent Receptive Fields, employing separate GNNs for each modality with independent receptive fields to enhance performance. To address the challenge where the optimal receptive field size, which can be quite low, restricts GNNs' ability to capture global information, we proposed a Sampling-based Global Transformer (SGT). It utilizes uniform global sampling to more efficiently integrate global information within GNN frameworks. Experiments show that the SGT improves performance even with a small number of sampled vertices, confirming sampling as an effective method for applying Global Transformers in multimodal recommendations.

Acknowledgements

References

Cai, D.; Qian, S.; Fang, Q.; Hu, J.; and Xu, C. 2022. Adaptive Anti-Bottleneck Multi-Modal Graph Learning Network for Personalized Micro-video Recommendation. In *MM '22, Lisboa, Portugal, October 10 - 14, 2022*, 581–590. ACM.

Chen, X.; Chen, H.; Xu, H.; Zhang, Y.; Cao, Y.; Qin, Z.; and Zha, H. 2019. Personalized Fashion Recommendation with Visual Explanations based on Multimodal Attention Network: Towards Visually Explainable Recommendation. In *ACM SIGIR 2019, Paris, France, July 21-25, 2019*, 765–774. ACM.

Deng, C.; Yue, Z.; and Zhang, Z. 2024. Polynormer: Polynomial-Expressive Graph Transformer in Linear Time. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Fang, Q.; Zhang, X.; Hu, J.; Wu, X.; and Xu, C. 2023. Contrastive Multi-Modal Knowledge Graph Representation Learning. *IEEE Trans. Knowl. Data Eng.*, 35(9): 8983–8996.

Gao, C.; Zheng, Y.; Li, N.; Li, Y.; Qin, Y.; Piao, J.; Quan, Y.; Chang, J.; Jin, D.; He, X.; and Li, Y. 2023. A Survey of Graph Neural Networks for Recommender Systems: Challenges, Methods, and Directions. *Trans. Recomm. Syst.*, 1(1): 1–51.

Gao, J.; Zhang, T.; and Xu, C. 2021. Learning to Model Relationships for Zero-Shot Video Classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(10): 3476–3491.

Guo, Z.; Li, J.; Li, G.; Wang, C.; Shi, S.; and Ruan, B. 2024. LGMRec: Local and Global Graph Learning for Multimodal Recommendation. In *AAAI 2024, February 20-27, 2024, Vancouver, Canada*, 8454–8462. AAAI Press.

Han, N.; Chen, J.; Xiao, G.; Zhang, H.; Zeng, Y.; and Chen, H. 2021. Fine-grained Cross-modal Alignment Network for Text-Video Retrieval. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, 3826–3834. ACM.

He, R.; and McAuley, J. J. 2016a. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. In *ACM WWW 2016, Montreal, Canada, April 11 - 15, 2016*, 507–517. ACM.

He, R.; and McAuley, J. J. 2016b. VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. In *AAAI 2016, Phoenix, Arizona, USA*, 144–150. AAAI Press.

He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; and Wang, M. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *ACM SIGIR 2020, Virtual Event, China, July 25-30, 2020*, 639–648. ACM.

Hu, J.; Hooi, B.; Qian, S.; Fang, Q.; and Xu, C. 2024. MGDCF: Distance Learning via Markov Graph Diffusion for Neural Collaborative Filtering. *IEEE Transactions on Knowledge and Data Engineering*, 1–16.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Klicpera, J.; Bojchevski, A.; and Günnemann, S. 2019. Predict then Propagate: Graph Neural Networks meet Personalized PageRank. In *ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Koren, Y.; Bell, R. M.; and Volinsky, C. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer*, 42(8): 30–37.

Liu, H.; Wei, Y.; Liu, F.; Wang, W.; Nie, L.; and Chua, T. 2024. Dynamic Multimodal Fusion via Meta-Learning Towards Micro-Video Recommendation. *ACM Trans. Inf. Syst.*, 42(2): 47:1–47:26.

Liu, K.; Xue, F.; Guo, D.; Wu, L.; Li, S.; and Hong, R. 2023a. MEGCF: Multimodal Entity Graph Collaborative Filtering for Personalized Recommendation. *ACM Trans. Inf. Syst.*, 41(2): 30:1–30:27.

Liu, Q.; Wu, S.; and Wang, L. 2017. DeepStyle: Learning User Preferences for Visual Recommendation. In *ACM SIGIR 2017, Shinjuku, Tokyo, Japan, August 7-11, 2017*, 841–844. ACM.

Liu, Z.; Yu, X.; Fang, Y.; and Zhang, X. 2023b. GraphPrompt: Unifying Pre-Training and Downstream Tasks for Graph Neural Networks. In *ACM WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, 417–428. ACM.

Mao, K.; Zhu, J.; Xiao, X.; Lu, B.; Wang, Z.; and He, X. 2021. UltraGCN: Ultra Simplification of Graph Convolutional Networks for Recommendation. In *ACM CIKM*, 1253–1262. ACM.

Qian, S.; Xue, D.; Fang, Q.; and Xu, C. 2023. Integrating Multi-Label Contrastive Learning With Dual Adversarial Graph Neural Networks for Cross-Modal Retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(4): 4794–4811.

Rendle, S.; Freudenthaler, C.; Gantner, Z.; and Schmidt-Thieme, L. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *UAI 2009, Montreal, QC, Canada, June 18-21, 2009*, 452–461. AUAI Press.

Shang, Y.; Gao, C.; Chen, J.; Jin, D.; Wang, M.; and Li, Y. 2023. Learning Fine-grained User Interests for Micro-video Recommendation. In *ACM SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, 433–442. ACM.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Sun, P.; Wu, L.; Zhang, K.; Chen, X.; and Wang, M. 2024. Neighborhood-Enhanced Supervised Contrastive Learning for Collaborative Filtering. *IEEE Trans. Knowl. Data Eng.*, 36(5): 2069–2081.

Tao, Z.; Liu, X.; Xia, Y.; Wang, X.; Yang, L.; Huang, X.; and Chua, T. 2023. Self-Supervised Learning for Multimedia Recommendation. *IEEE Trans. Multim.*, 25: 5107–5116.

van den Berg, R.; Kipf, T. N.; and Welling, M. 2017. Graph Convolutional Matrix Completion. *arXiv preprint arXiv:1706.02263*.

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

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *NeurIPS 2017, December 4-9, 2017, Long Beach, CA, USA*, 5998–6008.

Wang, M.; Zheng, D.; Ye, Z.; Gan, Q.; Li, M.; Song, X.; Zhou, J.; Ma, C.; Yu, L.; Gai, Y.; Xiao, T.; He, T.; Karypis, G.; Li, J.; and Zhang, Z. 2020. Deep Graph Library: A Graph-Centric, Highly-Performant Package for Graph Neural Networks. arXiv:1909.01315.

Wang, Q.; Wei, Y.; Yin, J.; Wu, J.; Song, X.; and Nie, L. 2023. DualGNN: Dual Graph Neural Network for Multimedia Recommendation. *IEEE Trans. Multim.*, 25: 1074–1084.

Wang, X.; He, X.; Wang, M.; Feng, F.; and Chua, T. 2019. Neural Graph Collaborative Filtering. In *ACM SIGIR*, 165–174. ACM.

Wei, W.; Huang, C.; Xia, L.; and Zhang, C. 2023. Multi-Modal Self-Supervised Learning for Recommendation. In *ACM WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, 790–800. ACM.

Wei, Y.; Wang, X.; Nie, L.; He, X.; and Chua, T. 2020. Graph-Refined Convolutional Network for Multimedia Recommendation with Implicit Feedback. In *ACM MM '20, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, 3541–3549. ACM.

Wei, Y.; Wang, X.; Nie, L.; He, X.; Hong, R.; and Chua, T. 2019. MMGCN: Multi-modal Graph Convolution Network for Personalized Recommendation of Micro-video. In *ACM MM 2019, Nice, France, October 21-25, 2019*, 1437–1445. ACM.

Wu, Q.; Zhao, W.; Yang, C.; Zhang, H.; Nie, F.; Jiang, H.; Bian, Y.; and Yan, J. 2023a. Simplifying and Empowering Transformers for Large-Graph Representations. In *NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Wu, S.; Sun, F.; Zhang, W.; Xie, X.; and Cui, B. 2023b. Graph Neural Networks in Recommender Systems: A Survey. *ACM Comput. Surv.*, 55(5): 97:1–97:37.

Ying, R.; He, R.; Chen, K.; Eksombatchai, P.; Hamilton, W. L.; and Leskovec, J. 2018. Graph Convolutional Neural Networks for Web-Scale Recommender Systems. In *ACM SIGKDD*, 974–983. ACM.

Yu, P.; Tan, Z.; Lu, G.; and Bao, B. 2023. Multi-View Graph Convolutional Network for Multimedia Recommendation. In *ACM MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, 6576–6585. ACM.

Zhang, D.; Zhu, Y.; Dong, Y.; Wang, Y.; Feng, W.; Kharlamov, E.; and Tang, J. 2023. ApeGNN: Node-Wise Adaptive Aggregation in GNNs for Recommendation. In *ACM WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, 759–769. ACM.

Zhang, J.; Zhu, Y.; Liu, Q.; Wu, S.; Wang, S.; and Wang, L. 2021. Mining Latent Structures for Multimedia Recommendation. In *ACM MM '21, Virtual Event, China, October 20 - 24, 2021*, 3872–3880. ACM.

Zhou, H.; Zhou, X.; Zeng, Z.; Zhang, L.; and Shen, Z. 2023a. A Comprehensive Survey on Multimodal Recommender Systems: Taxonomy, Evaluation, and Future Directions. arXiv:2302.04473.

Zhou, X.; and Miao, C. 2024. Disentangled Graph Variational Auto-Encoder for Multimodal Recommendation With Interpretability. *IEEE Trans. Multim.*, 26: 7543–7554.

Zhou, X.; and Shen, Z. 2023. A Tale of Two Graphs: Freezing and Denoising Graph Structures for Multimodal Recommendation. In *ACM MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, 935–943. ACM.

Zhou, X.; Zhou, H.; Liu, Y.; Zeng, Z.; Miao, C.; Wang, P.; You, Y.; and Jiang, F. 2023b. Bootstrap Latent Representations for Multi-modal Recommendation. In *ACM WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, 845–854. ACM.

Zhu, Y.; Cong, F.; Zhang, D.; Gong, W.; Lin, Q.; Feng, W.; Dong, Y.; and Tang, J. 2023. WinGNN: Dynamic Graph Neural Networks with Random Gradient Aggregation Window. In *ACM KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, 3650–3662. ACM.