



Title	Research on Improving GNN-based Recommender System Involving Multi-modal Information
Author(s)	Jiao, Shengzhe
Citation	大阪大学, 2026, 博士論文
Version Type	VoR
URL	https://doi.org/10.18910/105505
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

<https://ir.library.osaka-u.ac.jp/>

The University of Osaka

Research on Improving GNN-based
Recommender System Involving Multi-modal
Information

Submitted to
Graduate School of Information Science and Technology
The University of Osaka

January 2026

Shengzhe JIAO

List of Publications

1. Journal Paper

1. Shengzhe Jiao, Yihong Zhang, and Takahiro Hara. “Multi-task Multi-modal Graph Neural Network for Recommender System.” *Knowledge and Information Systems* (2025): Volume 67, pages 8059–8084.

2. International Conference Paper

1. Shengzhe Jiao, Yihong Zhang, and Takahiro Hara. “Multi-modal Relational Side Information Graph Attention Networks for Recommender System.” 2023 International Conference on Machine Learning and Applications (ICMLA). IEEE, 2023.
2. Shengzhe Jiao, Yihong Zhang, Yuanyuan Wang, Shingo Mabu, Haoyang Xia, and Takahiro Hara. “Multi-Modal Contrastive Learning for Medical Image Classification with Limited Training Data.” 2024 International Conference on Machine Learning and Applications (ICMLA). IEEE, 2024.
3. Shengzhe Jiao, Yihong Zhang, Aoran Chen, and Takahiro Hara. “LLM Augmentation Enhanced Counterfactual Attention for Graph-based Recommender System.” 2025 International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT). IEEE/WIC, 2025.

Abstract

With the rapid growth of online platforms and the large number of products and services they provide, users are increasingly overwhelmed by information and face difficulties in identifying items that match their preferences. Recommender systems (RSs) have been developed to address this problem by modeling user-item interactions, such as views, clicks, and purchases, and generating personalized recommendations. By predicting user preferences from historical behaviors, RSs improve user experience and provide substantial commercial value to service providers. As a result, RSs have become a fundamental component of modern online ecosystems, including e-commerce platforms, streaming services, and social networks.

Recent advances in deep learning have significantly enhanced the ability of RSs to capture complex and nonlinear relationships between users and items. Deep models can automatically learn high-dimensional representations that reflect both explicit and implicit factors of user preferences, leading to improved recommendation accuracy. However, their effectiveness is often limited by the sparsity of interaction data: in real-world systems, users typically interact with only a small subset of items, and many items receive limited feedback. This sparsity poses a major challenge for learning robust representations and reduces the practical performance of deep learning-based RSs.

To alleviate data sparsity, multi-modal recommender systems have been proposed to incorporate additional side information, such as images, textual descriptions, and categorical attributes. These heterogeneous modalities provide complementary perspectives on user preferences and item characteristics, enabling richer representations. Despite their potential, existing multi-modal approaches still face several fundamental challenges that limit their effectiveness, robustness, and explainability.

First, many methods integrate different modalities in a shallow manner, often processing each modality independently and combining them through simple fusion strategies. Such designs fail to capture the underlying relationships between modalities and overlook modality-specific relations, such as visual similarity or textual correlation among items, leading to incomplete item representations. Second, there is often a misalignment between user behaviors and side information: interaction-based embeddings and content-based features are learned separately, which may result in representations

that are semantically inconsistent with actual user preferences. Third, existing RSs generally lack explainability. While achieving high accuracy, they operate as black boxes and provide limited insight into why particular items are recommended, which reduces user trust and restricts their applicability in sensitive domains.

In this thesis, we investigate multi-modal graph-based learning approaches to address these challenges and improve both the accuracy and explainability of recommender systems. Specifically, we develop three models that target different aspects of the recommendation process.

In Chapter 2, we propose a framework that deeply integrates multi-modal information by constructing modality-specific relational graphs that jointly model user-item interactions and item-item relations derived from side information. By enabling information propagation across modalities, the model captures richer dependencies between users and items and improves robustness under sparse interaction settings.

In Chapter 3, we address the misalignment between behavioral and semantic representations by introducing a joint learning strategy that explicitly aligns interaction-based embeddings with similarity structures derived from side information. Through multi-task optimization, the model learns representations that are consistent with both user behaviors and item semantics, leading to improved generalization when interaction data are limited.

In Chapter 4, we focus on enhancing explainability by enriching item semantics with large language model (LLM)-generated entities and introducing a personalized reasoning mechanism. By augmenting side information and estimating the causal influence of semantic entities on user preferences, the proposed approach improves both recommendation accuracy and transparency.

Experimental results on multiple real-world datasets demonstrate that the proposed methods consistently outperform state-of-the-art baselines, confirming their effectiveness in modeling sparse interactions, heterogeneous modalities, and explainable reasoning.

Finally, Chapter 5 summarizes the overall contributions of this thesis, discusses the limitations of the three proposed models, and outlines several promising future research directions.

Contents

1	Introduction	1
1.1	Background	1
1.2	Motivation and Scope	3
1.3	Fundamental Challenges in RSs	5
1.3.1	Limited Use of Complementary Information Across Modalities	7
1.3.2	Inconsistent Representations Across Different Modalities	8
1.3.3	Lack of Rich Side Information and Inaccurate Attention Weight- ing	9
1.4	Overview of Research Contents	10
1.5	Organization	11
2	Utilizing Multi-modal Side Information for Better Recommendation	15
2.1	Introduction	15
2.1.1	Challenge	16
2.1.2	Contribution	17
2.2	Related Work	18
2.2.1	Multi-modal Recommendation	18
2.2.2	GNN-based Recommendation	19
2.3	Preliminaries	20
2.4	Proposed Framework	21
2.4.1	Overview	21
2.4.2	Graph Structures	21
2.4.3	Aggregation Layers	24
2.4.4	Combination Layers	26

2.4.5	Prediction Layer	27
2.4.6	Optimization	27
2.5	Experiments	28
2.5.1	Datasets	28
2.5.2	Baselines	29
2.5.3	Evaluation Metrics	30
2.5.4	Parameter Settings	30
2.5.5	Performances Comparison (RQ1)	30
2.5.6	Ablation Study (RQ2)	32
2.5.7	Further Model Analysis (RQ3)	33
2.6	Conclusion and Future Work	35
3	Multi-task Learning for Multi-modal Representation Enhancement	37
3.1	Introduction	37
3.1.1	Challenge	38
3.1.2	Contribution	39
3.2	Related Work	40
3.2.1	Multi-modal Recommendation	40
3.2.2	Multi-task Recommendation	42
3.2.3	GNN-based Recommendation	43
3.3	Preliminaries	44
3.4	Proposed Method	45
3.4.1	Overview	45
3.4.2	Representations from GNNs	46
3.4.3	Incorporating Side Information	51
3.4.4	Combined Model Training and Inference	54
3.5	Experiments	57
3.5.1	Datasets	57
3.5.2	Baselines	58
3.5.3	Evaluation Metrics	59
3.5.4	Parameter Settings	59
3.5.5	Performances Comparison (RQ1)	60

3.5.6	Ablation Study (RQ2-4)	60
3.5.7	Further Model Analysis	65
3.6	Conclusion	66
4	LLM-augmented Counterfactual Attention for Reliable Recommendations	69
4.1	Introduction	69
4.1.1	Challenge	70
4.1.2	Contribution	71
4.2	Related Work	72
4.2.1	Multi-modal Recommendation	72
4.2.2	GNN-based Recommendation	73
4.3	Preliminaries	74
4.4	Proposed Method	75
4.4.1	Overview	75
4.4.2	Survey-Guided LLM Data Augmentation	76
4.4.3	Item-entity KG Structure and Training	77
4.4.4	User-Item Interaction Graph Structure	78
4.4.5	Counterfactual Attention Mechanism	79
4.4.6	BPR Loss	81
4.4.7	Combined Model Training	82
4.5	Experiment	82
4.5.1	Datasets	82
4.5.2	Baselines	83
4.5.3	Evaluation Metrics	84
4.5.4	Parameter Settings	84
4.5.5	Performances Comparison (RQ1)	84
4.5.6	Ablation Study (RQ2)	86
4.5.7	Effect of Different LLM models (RQ3)	87
4.6	Conclusion	87
5	Summary	89
5.1	Summary of Contributions	89
5.2	Limitations	91

5.3	Future Work	92
5.3.1	Enhancing the Robustness of LLM-augmented Knowledge . . .	93
5.3.2	Scalable and Efficient Graph Learning	93
5.3.3	Extending to Cross-domain Recommendations	93
5.3.4	Integrated Framework Combining All Proposed Techniques . .	93
Acknowledgment		95

Chapter 1

Introduction

1.1 Background

With the explosive growth of online information and services, efficiently exploring useful content from the overwhelming amount of available options has become a tough and important task in our daily lives [1, 2]. Personalized recommendations provided by online platforms can handle the overload of information and effectively reduce the time costs of users' exploration process by automatically identifying items that fit their preferences and needs.

Consequently, recommender systems (RSs) have become a central part of user experience in a wide range of online services. Many modern platforms, including e-commerce sites (e.g., Amazon, Rakuten), streaming platforms (e.g., Netflix, YouTube), social networking services (e.g., X, Instagram), and app stores (e.g., Google Play, Apple App Store), rely heavily on RSs to provide tailored experiences. These systems not only improve user satisfaction by helping users efficiently locate relevant items but also contribute significantly to business success by increasing engagement and conversion rates. Because of their crucial role in both user experience and business revenue, RSs have attracted continuous research interest in both academia and industry over the past two decades [3, 4, 5].

Figure 1.1 illustrates a general paradigm of a RS. In practice, service providers collect the historical interactions between users and items within their platforms, such as

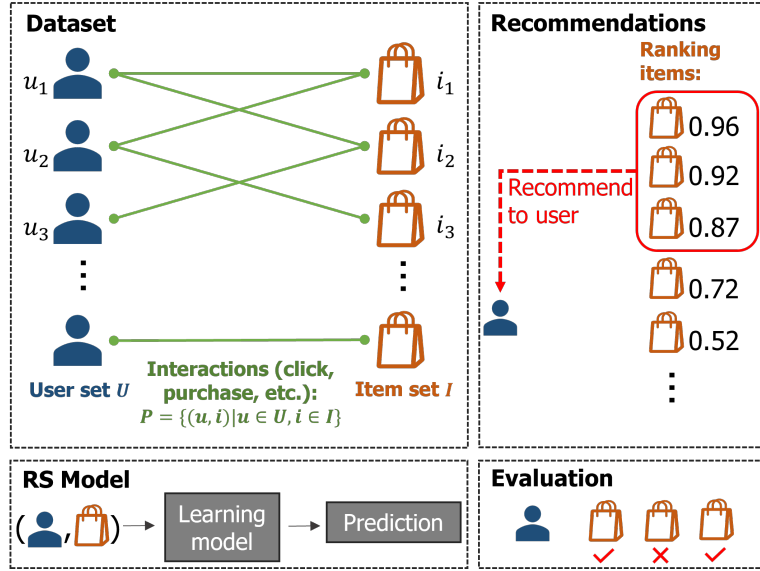


Figure 1.1: A general paradigm of a RS. Service providers collect historical user-item interactions and train a learning model to predict users’ preferences toward items. The trained model outputs a ranked list of items for each user based on predicted preference scores.

the records of items that users have clicked, purchased, viewed, or rated. These interaction records are then used to train a learning model that predicts users’ preferences toward items that they have not yet interacted with. The model outputs a numerical preference score for each user-item pair, based on which all items are ranked, and the top-ranked ones are presented as recommendations to the user. When users’ preferences are accurately captured, the recommendations are expected to satisfy user interests, thus improving the user experience [6, 7].

In industrial applications, the performance of RSs is usually assessed through online A/B tests, which compare key metrics such as click-through rate (CTR), conversion rate, or sales revenue under different recommendation strategies. Empirical results consistently confirm that well-designed RSs can have a major impact on business outcomes. For example, [8] reported that replacing a non-personalized recommendation module with a purchase-based recommendation system increased sales by 35 percentage for an online DVD retailer. Similarly, studies conducted at YouTube showed that the dif-

ference in CTR between two RS variants could reach as high as 200 percentage [9], while experiments at Google News demonstrated a 38 percentage improvement in CTR when replacing popularity-based recommendations with personalized ones [10]. Such examples highlight the significant commercial potential of accurate RSs. Accordingly, service providers consistently demand high-performance RSs to further enhance both user satisfaction and commercial profit [11, 12, 13, 14].

In contrast, academic research on RSs typically evaluates models through offline experiments on benchmark datasets, where the goal is to measure how accurately the model can predict or rank items that users actually interacted with in held-out test data. The underlying assumption is that when a RS ranks the items a user interacted with higher, it better captures the user’s true interests. Although offline evaluation cannot perfectly replicate real-world user behavior, existing studies have shown that offline ranking metrics such as Recall@K and NDCG@K are strongly correlated with online indicators such as CTR [1]. Thus, improving offline recommendation accuracy is widely recognized as an important step toward improving the real-world performance of RSs.

Both in industry and academia, improving the performance of RSs serves two purposes: (i) it enhances user experience by reducing the effort and time required to discover relevant items, and (ii) it increases business profitability by driving engagement and purchases. As such, continual improvement of RSs has become one of the most active and enduring research topics in the field of information retrieval and machine learning [15, 16].

1.2 Motivation and Scope

In recent years, with the success of deep learning, RSs have undergone a rapid transformation. Deep neural architectures, such as graph neural networks (GNNs), have demonstrated remarkable ability in modeling complex and nonlinear relationships between users and items [17, 18], which users and items can be represented as nodes and interactions can be represented as edges. Figure 1.2 shows an example of graph based RS.

Since these approaches are known to be data hungry, they require a large amount of interaction data to learn millions of parameters effectively [19, 20, 21]. In practice,

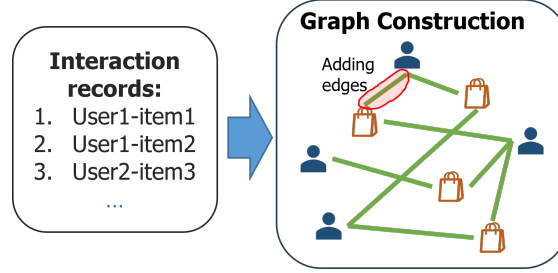


Figure 1.2: An example of graph structure based on the user-item interactions, where users and items can be represented as nodes and interactions can be represented as edges.

most real-world online services face data sparsity, where only a small subset of users actively interact with a limited number of items. This issue is existed not only in newly launched services or small-scale companies but also in big and well-known platforms with massive user bases [22, 23]. Under such sparse scenarios, the performance of deep learning-based RSs declines sharply, making it difficult to achieve accurate and reliable recommendations from interaction data alone.

To handle the limitations caused by sparse user-item interactions, researchers have increasingly turned to the use of side information, such as product images, textual descriptions, categories, and other attributes, which provide valuable complementary knowledge about items and users [24, 25, 26, 27]. This trend has led to the emergence of MMRSs (MMRSs), which integrate multiple types of information or modalities to gather richer and more robust representations of users and items [28, 29, 30, 31, 32]. Different modalities can capture different aspects of user preference: for example, visual data (e.g., product or movie posters) capture appearance and style, textual data (e.g., titles, descriptions, or reviews) capture semantics and opinions, and categorical metadata (e.g., genre, brand, or price range) capture contextual structure. By jointly leveraging these modalities, RSs can not only enhance accuracy but also improve robustness against missing or sparse interaction data [33, 34].

The incorporation of multi-modal information has achieved promising results in a variety of domains [35, 36, 37, 38]. In e-commerce, visual features extracted from product images help model user preferences for color, shape, or fashion style [39, 40, 41]. In movie and music recommendations, textual descriptions and genre metadata provide

important semantic characters about thematic content. Similarly, in social media and news recommendations, combining text, image, and temporal data enables the system to better understand user interests and adapt to fast-changing contexts [42, 43]. These applications demonstrate the effectiveness of multi-modal information in improving personalization quality and recommendation reliability.

However, despite these advancements, existing MMRSs still face several critical challenges that limit their effectiveness in real-world applications [44, 45, 46]. First, many approaches [47, 48, 49, 50] integrate different modalities only at the feature level, simply concatenating or averaging modality-specific embeddings. Such shallow fusion overlooks the rich relational structure among modalities and cannot fully capture how different types of information interact with each other. Second, the representations learned from user interactions and those learned from side information are often not well aligned, leading to inconsistencies between behavioral and semantic spaces. This misalignment can result in poor performance, especially when side information contains noise or incomplete data. Third, while deep learning-based MMRSs can achieve high accuracy, they often operate as black boxes, offering little explainability about why certain items are recommended. The lack of transparency decrease user trust and limits the practical deployment of RSs in domains where explanations are essential. Also, the types of side information contained in the existing dataset are still limited, which can not fully represent the users' preferences.

In summary, while deep learning and multi-modal fusion have greatly advanced RSs, current approaches still struggle to integrate modalities deeply, align user behavior with item semantics, and provide explainable reasoning behind their recommendations. Addressing these issues is critical for building the next generation of RSs that are accurate, robust, and explainable—which is the main motivation of this thesis.

1.3 Fundamental Challenges in RSs

Although MMRSs have achieved remarkable progress in recent years, several fundamental challenges remain unsolved when deploying them in real-world scenarios. The goal of this thesis is to develop RSs that can effectively utilize multi-modal side information, align heterogeneous data with user behavior, enrich the item side information,

and provide explainable reasoning. Considering the practical demand for accurate and explainable recommendations, this thesis investigates the essential challenges that limit current systems and aims to propose feasible solutions to address them.

In real-world recommendation environments, users interact with only a small fraction of items, and side information often comes from multiple modalities, such as text, images, and structured attributes. These modalities differ in feature space, noise level, and semantic meaning, which makes their integration challenging. In addition, user behaviors (clicks, views, purchases) and item content (textual or visual information) are often modeled independently, resulting in inconsistent representations that fail to accurately reflect user interests. Moreover, as models become more complex, explainability becomes a crucial and underexplored issue: deep learning models can achieve high accuracy but offer little insight into why a particular recommendation is made. Also, the lack of side information will affect the recommendation accuracy and explainability due to poor understanding of users' preferences. To build systems that are both reliable and trustworthy, it is essential to address these limitations comprehensively.

From a broader perspective, making RSs practical in real-world applications involves multiple fundamental dimensions, including data sparsity, representation learning, system scalability, robustness, and user trust. Among these, representation-related challenges play a central role in modern MMRs, as they directly determine how heterogeneous information is perceived, integrated, and utilized by the model.

In particular, recent advances in MMRs have shown that simply incorporating additional modalities is insufficient. The effectiveness of such systems critically depends on how multi-modal information is structured, how semantic content is aligned with user behavior, and how the learned representations can be interpreted and trusted. Therefore, this thesis focuses on representation-centric challenges that consist of multi-modal learning, graph modeling, and explainable recommendation.

Under this representation-centric perspective, this thesis focuses on three key challenges that are frequently encountered in the design of MMRs:

- The shallow integration of multi-modal information that leads to incomplete item understanding.
- The misalignment between user behaviors and side information that reduces the

consistency of learned representations.

- The lack of information richness and explainability that affect user trust and model transparency.

These challenges are interconnected—improving one often affects the others—and collectively determine how well a RS can perform in sparse and complex environments.

We briefly describe each challenge below and explain why it remains an open problem in current research. The first challenge concerns the limited ability of existing systems to deeply integrate heterogeneous information sources in a unified framework. The second challenge arises from the difficulty of aligning behavioral and semantic information when modalities differ in scale, representation, and context. The third challenge involves the lack of data richness and explainability, and the need for explanation mechanisms that can provide human-understandable reasoning for each recommendation. Addressing these issues is the main motivation of this thesis.

In the following subsections, we elaborate on these three challenges individually and discuss their corresponding motivations in detail.

1.3.1 Limited Use of Complementary Information Across Modalities

In modern online services, each item is often associated with diverse side information such as product images, textual descriptions, and categorical attributes. These different types of data, or modalities, contain complementary signals that can provide a more comprehensive understanding of items and user preferences. For example, an image can represent the visual appearance and style of a product, while the related text may describe its functions or specifications. By jointly using these heterogeneous sources, RSs are expected to generate more accurate and reliable recommendations, especially when direct user-item interactions are limited.

However, effectively combining multiple modalities into a single framework remains a fundamental challenge. Existing studies often process each modality independently and simply merge their representations at a later stage. Such shallow fusion overlooks the deeper relationships between modalities. For example, how textual mean-

ings align with visual features, this issue prevents the model from fully exploiting the complementary nature of multi-modal data. As a result, the learned representations of users and items are often incomplete and may not accurately reflect users' true interests. The challenge becomes even greater in real-world environments, where side information can be noisy, missing, or unbalanced across modalities. Some items may have rich textual descriptions but poor image quality, while others may contain detailed visuals but lack metadata or attributes. When the system cannot properly adapt to these differences, its performance becomes unstable and biased toward dominant modalities. Essential difficulty lies not only in fusing information from different modalities but also in learning how to integrate them adaptively according to their reliability and relevance in each context.

Therefore, the first challenge of this thesis is to design a unified framework that can deeply and flexibly integrate multi-modal information, allowing meaningful interactions between modalities and enabling the system to make accurate recommendations even under sparse and noisy data conditions.

1.3.2 Inconsistent Representations Across Different Modalities

In RSs, user behaviors—such as clicks, views, and purchases—reflect what users are interested in, while side information such as images, textual descriptions, and attributes describes what the items actually are. In an ideal scenario, these two sources of information should be closely aligned: the semantic content of an item should explain why a user interacts with it, and the behavioral patterns of users should reinforce the semantic meaning of the items they prefer. Achieving this alignment allows the system to understand why certain interactions occur and to make more meaningful predictions for future recommendations.

However, most existing RSs learn user-item interactions and side information separately. Behavioral features are often captured through collaborative filtering or graph-based models, while side information is processed independently using visual or textual encoders. Without an explicit connection between these two representations, the learned embedding spaces may not correspond to each other. This misalignment causes the system to rely on features that are semantically irrelevant to user preferences or to

overlook the true factors driving interactions. For example, two products might share similar textual keywords but appeal to entirely different user groups, or visually similar items may have distinct functional purposes that are not reflected in the interaction data. Furthermore, side information in multi-modal settings is heterogeneous and unbalanced: different modalities may contribute unequally to user decisions depending on the context. When behavioral and semantic spaces are not aligned, the system may overfit to abundant modalities or noise while ignoring subtle behavioral factors. This weak correlation between user actions and item semantics often leads to unstable performance, especially when interaction data are sparse or when items have incomplete side information.

Therefore, the second challenge of this thesis is to bridge the gap between user behaviors and item semantics by establishing a consistent relationship between what users do and what items represent. Aligning these two spaces allows the RS to learn user preferences that are both behaviorally guided and semantically meaningful, improving generalization and reliability across different domains and data conditions.

1.3.3 Lack of Rich Side Information and Inaccurate Attention Weighting

As RSs become increasingly complex with the adoption of deep neural networks and multi-modal learning frameworks, understanding how these systems generate their recommendations has become a critical challenge. Modern models can predict users' preferences with high accuracy, but they often operate as black boxes, providing little insight into the reasoning behind their outputs. In practical applications, users, developers, and service providers all require explainable recommendations—users need to know why certain items are suggested to them, while developers need transparency to ensure fairness, detect bias, and maintain trust in the system.

In most existing approaches, the internal representations learned by deep models are abstract and difficult to interpret. Although they successfully encode user-item relationships, these embeddings do not directly correspond to human-understandable concepts such as product features, styles, or attributes. As a result, when a system recommends a particular item, it is often impossible to determine which specific factors—visual ap-

pearance, textual description, or brand—contributed most to the decision. This lack of explainability not only reduces user trust but also makes it challenging to analyze or improve system behavior in industrial deployments. The problem is further complicated by the fact that side information in real-world services is usually incomplete or inconsistent. When data about items are limited, traditional attention or visualization methods fail to provide clear and reliable explanations. Moreover, even when side information exists, most systems lack mechanisms to link their learned representations to explicit, human-readable entities. Without this connection, explanations remain abstract, limiting their usefulness in real decision-making scenarios.

Therefore, the third challenge of this thesis is to enhance the explainability and transparency of RSs by enriching side information with meaningful semantic knowledge and enabling the model to identify which factors drive individual recommendations. Addressing this challenge is essential for building trustworthy, user-centered RSs that not only provide accurate predictions but also deliver clear and understandable reasoning behind their recommendations.

1.4 Overview of Research Contents

We address each of the challenges and develop new RSs to overcome them. Our proposed methods utilize multi-modal and graph-based learning frameworks to solve the research issues discussed in Section 1.3. The outlines of the research conducted in this thesis are summarized below.

- Utilizing Multi-modal Side Information for Better Recommendation.

In Chapter 2, we address the challenge of effectively integrating information from multiple modalities in a unified framework. To overcome the limitations of shallow fusion, we construct a relational structure that connects user-item interactions with modality-specific relations such as visual similarity, textual correlation, and attribute similarity. This structure allows information to flow between different modalities, enabling the system to utilize complementary signals and better understand the relationships among users and items. By learning the importance of

each modality, the system achieves more accurate and stable recommendations, particularly under sparse interaction conditions.

- **Multi-task Learning for Multi-modal Representation Enhancement.**

In Chapter 3, we address the challenge of aligning user behaviors with item semantics to ensure that learned representations are both meaningful and consistent. We introduce a joint learning approach that simultaneously captures interaction patterns from behavioral data and semantic relations from side information. Through this process, the system minimizes discrepancies between the behavioral and semantic spaces, ensuring that the features learned from content truly reflect the factors influencing user preferences. This alignment not only improves generalization and stability but also enhances the explainability of the learned representations.

- **LLM-augmented Counterfactual Attention for Reliable Recommendations.**

In Chapter 4, we address the challenge of improving explainability and transparency in RSs. We develop a large-language-model (LLM) based data augmentation pipeline that automatically enriches items with fine-grained, human-readable semantic entities derived from textual descriptions or metadata [51]. These entities provide the foundation for an explainable reasoning mechanism that estimates the influence of each factor on the user’s predicted preference. By simulating counterfactual conditions—such as removing or modifying specific entities—the system can identify which features most strongly drive recommendations. This approach enhances both the accuracy and explainability of the recommendations, contributing to more trustworthy and user-centered systems.

1.5 Organization

This thesis consists of five chapters. The remainder of the thesis is organized as follows.

In Chapter 2, we study the challenge of limited use of complementary information across modalities. In Section 2.1, we introduce motivation of leveraging relational structures across modalities and discuss why existing MMRSs fail to fully utilize com-

plementary side information. Section 2.2 reviews prior works related to the topic of this chapter. Section 2.3 introduces some preliminaries of this chapter. Our proposed framework is presented in Section 2.4. We first give an overview of the framework, and then elaborate on our proposed graph attention-based model that captures both user-item interactions and relationships among item side information. Then, we elaborate on our model that predicts user preferences and our novel framework. The experiment setup and the results, including the comparison with state-of-the-art methods and the ablation study, are described in Section 2.5. Finally, a conclusion of this chapter is given in Section 2.6. The study in this chapter is based on our works published in [52].

In Chapter 3, we introduce and discuss the problem of inconsistent representations across different modalities. In Section 3.1, we describe the motivation of aligning behavioral and semantic spaces in multi-modal learning and explain why existing fusion methods fail to ensure consistency between user-item interactions and side information. Next, we review some related works on this topic in Section 3.2. In Section 3.3, we give some preliminaries of this chapter. Our proposed framework in this chapter is presented in Section 3.4. We first exhibit an overview of the framework and then describe each component of the proposed multi-task learning architecture in detail, including the modality-specific graph construction and the similarity-based semantic alignment task. We compare our proposed method with existing methods and conduct an ablation study to evaluate the effectiveness of each component in our proposal in Section 3.5. Finally, this chapter is summarized in Section 3.6. The study in this chapter is based on our work published in [53].

In Chapter 4, we focus on lack of rich side information and inaccurate attention weighting. We describe the issue of prior works in Section 4.1, highlighting the restricted semantic coverage of traditional knowledge graphs and the need for personalized reasoning. In Section 4.2, we review some related works on this topic. In Section 4.3, we give some preliminaries of this chapter. Section 4.4 presents our proposed framework. We first give an overview of our LLM-augmented counterfactual attention model and then elaborate on its two main components: entity enrichment through large language models and personalized causal reasoning via counterfactual attention. We conduct experiments to examine the performance of our proposed method, and the setting and results of experiments are described in Section 4.5. Finally, this chapter is

concluded in Section 4.6. The study in this chapter is based on our work published in WI-IAT 2025.

Finally, in Chapter 5, we summarize the overall contributions of this thesis in Section 5.1 and provide a unified discussion of the remaining limitations identified across the three proposed models in Section 5.2. We further highlight several open challenges that arise from our findings such as the need for richer semantics, more scalable graph learning, and deeper integration across modalities, and outline multiple promising research directions for advancing multi-modal, graph-based, and LLM-augmented RSs in Section 5.3.

Chapter 2

Utilizing Multi-modal Side Information for Better Recommendation

2.1 Introduction

As discussed in Section 1.3.1, integrating multi-modal side information is one of the fundamental challenges in developing RSs. Items in modern online services are often accompanied by various types of side information, such as textual descriptions, images, and categorical attributes. These heterogeneous data sources contain complementary signals that can enhance item understanding and alleviate the sparsity of user-item interactions. As a result, multi-modal information has become an essential component of many practical RSs, especially in scenarios where interaction data alone is insufficient to represent item semantics or user preferences.

To leverage such information, multi-modal RSs (MMRSs) employ deep learning models to extract features from different modalities and generate user and item representations [26, 54]. For example, convolutional neural networks (CNNs) are commonly used for visual signals such as product images, while recurrent or transformer-based models are applied to textual data including titles, descriptions, and reviews. Existing MMRSs [29, 55] typically follow a two-stage pipeline: each modality is processed independently to obtain modality-specific embeddings, and the outputs are subsequently fused through concatenation, weighted averaging, or shallow attention mechanisms.

While these methods enrich the representational space, they often treat modalities as independent sources of information and overlook the relational structure that exists among users, items, and side information.

Recent advances in GNNs have shown strong potential to capture complex relationships in RSs[48, 56]. User–item interactions can be described as bipartite graphs, where graph convolutions propagate collaborative signals across multi-hop neighbors. This graph formulation enables the model to better exploit the underlying structure of user behaviors and item co-occurrences. Motivated by this progress, some studies have attempted to incorporate multi-modal information into graph-based frameworks. A common approach is to build a separate user–item bipartite graph for each modality and aggregate the representations across graphs.

This chapter focuses on addressing this limitation by developing a relational framework that models the interactions among users, items, and their multi-modal side information. By analyzing item–item similarity patterns within each modality and explicitly incorporating these relationships into graph structures, our approach aims to move beyond shallow feature fusion toward a more expressive and explainable multi-modal recommendation model.

2.1.1 Challenge

Despite the increasing interest in multi-modal recommendation, existing MMRSs still face several key limitations. Most of them process each modality separately and fuse the representations only at a late stage through simple concatenation or averaging. Such shallow fusion fails to capture the complex semantic relationships between modalities, limiting the model’s ability to represent users’ diverse and fine-grained preferences. Moreover, most graph-based RSs only focus on user-item interactions and overlook implicit relationships between items derived from side information, such as textual or visual similarity. In particular, side information is not only auxiliary content but often encodes meaningful relational patterns. For example, items with similar visual appearance may appeal to similar user groups, and items sharing textual attributes may appear together in user behavior sequences. Existing MMRSs rarely incorporate such relational dependencies. They overlook the fact that side information and user–item interactions

are related: similarities among items can shape user preferences, and user interactions implicitly reflect how side information influences decision making. As a result, information from side modalities cannot be fully utilized, especially when user-item interactions are sparse.

In addition, side information across modalities is often noisy or incomplete: images may lack meaningful visual information, while textual descriptions may be short or inconsistent. Using the same structure and equal weights for all modalities makes the models less robust and easy to bias from unreliable information. Therefore, this chapter focuses on the following challenges:

- Designing a unified relational framework that can jointly model user-item interactions and item-item relationships derived from multiple modalities.
- Developing an adaptive mechanism that can assign appropriate importance to different modalities according to their informativeness and reliability.

2.1.2 Contribution

To address the above challenges, we develop a Multi-modal Relational Side Information Graph Attention Network (MRSI-GAT). The proposed framework constructs a unified graph where users and items are represented as nodes, and multiple types of edges are defined to capture both user-item interactions and modality-specific item-item relations. An attention mechanism is applied to dynamically assign weights to different edges, allowing the model to focus on the most informative relations in each modality. Through message propagation across modalities, the model can learn richer and more coherent user and item representations that better reflect users' true preferences.

The main contributions of this chapter are summarized as follows:

- We propose a unified multi-modal relational framework that jointly models user-item interactions and item-item connections derived from textual, visual, and categorical information.
- We introduce an attention-based message passing mechanism to adaptively integrate information from different modalities, improving the robustness of recommendations under sparse and noisy conditions.

- Experimental results on two public datasets (Amazon Musical Instruments and MovieLens) demonstrate that our method significantly outperforms several state-of-the-art baselines in Recall@K and NDCG@K.

The organization of this chapter is as follows. We introduce related works in Section 2.2. Next, we describe some preliminaries of this chapter in Section 2.3. Then, we present the detail of our model framework in Section 2.4. The experimental setup and results are presented in Section 2.5. Finally, a conclusion of this chapter is given in Section 2.6.

2.2 Related Work

In this section, we introduce some related works, including multi-modal recommendations and GNN-based recommendations.

2.2.1 Multi-modal Recommendation

Methods of the multi-modal RS are mainly based on the idea of the combination of multi-media models and CF. Matrix factorization (MF) is an effective CF-based method, which makes predictions by the representations of users and items. Due to the success of deep learning models for multi-media data, analyzing the side information can enrich the representations and improve performance. In most e-commerce platforms, side information generally includes visual, textual, and auxiliary data (e.g., brand, category, and manufacturer). For each type of side information, there are specific models to analyze the features of data, the methods mainly include convolutional neural network (CNN), recurrent neural network (RNN), and auto-encoder (AE). Zhang et al. [57] constructed a joint representation learning framework. In this method, each type of side information was adopted to learn the corresponding user and item representations and then integrated various information together to make predictions. Guan et al. [58] constructed specific models for different types of side information. In this method, visual features were processed by a stacked convolutional auto-encoder, and textual features were processed by a stacked auto-encoder. After generating the representations for various side information, they designed an attention layer called the individualized cogni-

tion layer to assign different weights of various side information for each user and then integrated the latent representations from multiple models. This method enriched the representations of users and items and made predictions. Although this method used multi-models for each type of side information, the stack auto-encoder just extracted the features from original side information vectors, so these features only contained the features of side information and trained separately with the prediction model, and not analyze the relationships between side information features and interactions. Hong et al. [59] constructed multi-models for user reviews and item’s various side information. In this method, visual features were processed by CNN and textual features were processed by RNN. Instead of randomly initializing the user’s representations, this method also analyzed the user’s review records to enrich the representations, which contained more information to improve the recommendation.

Although these methods analyzed multiple types of side information based on multi-models, they failed to explore the deeper relationships between side information and interactions because they did not analyze the high-order connectivity.

2.2.2 GNN-based Recommendation

Recently, the GNN techniques have been widely used in RSs because GNN techniques have been explored to be powerful in representation learning for graph data [60]. In most of the cases of recommendation, data can be represented as graph structure [61]. The common graph structures in RSs are user-item bipartite, sequence, social relationship, and knowledge graphs. Based on the different types of data, it is a challenge to design suitable models for RSs. For our specific task of recommendation, most of the existing methods directly applied GNNs on original user-item bipartite graph [62, 63]. Wang et al. [64] constructed a graph structure to integrate the user-item interactions into the embedding process. This method exploited the user-item bipartite graph by propagating the embeddings from higher-order connectivity to encode the collaborative filtering signal into the representation. Inspired by the ideas of multi-modal, Wei et al. [48] constructed a multi-modal graph convolution network framework to learn the specific representations of users and items in various modalities. This method analyzed different types of side information by multiple graphs, and the graph structures of each

side information were the same user-item bipartite graph. For each specific modality, user and item nodes can aggregate specific features to enrich the representations.

Although these methods took advantage of the GNN techniques and ideas of multi-modal, they just generated user and item representations by the user-item bipartite graph. Existing methods did not focus on the relationships between interactions and various side information. They just aggregated the embeddings by the edges of interactions such as purchase and click. Graph structures ignoring edges of relationships between different types of item’s side information are relatively shallow and can not contain deeper connectivity between users and items. Especially in the case of sparse data, users may interact with a small number of items. Even though existing methods tried to enrich the representations by analyzing side information in multiple models, they are still analyzing the shallow relationships between users and items.

2.3 Preliminaries

We consider a standard recommendation scenario consisting of a set of users $\mathcal{U}$ and items $\mathcal{I}$. The observed user-item interactions (e.g., clicks or purchases) are represented as binary pairs $(\mathbf{u}, \mathbf{i})$, which indicate user preferences. For each user $\mathbf{u}$ and item $\mathbf{i}$, let $\mathcal{P} = \{p_{ui} \mid \mathbf{u} \in \mathcal{U}, \mathbf{i} \in \mathcal{I}\}$ be the observed interactions, where $p_{ui} = 1$ denotes the user $\mathbf{u}$ interacted with item $\mathbf{i}$. The recommendation task aims to predict missing interactions by learning representations that capture user interests and item characteristics.

In addition to interaction data, each item may contain several types of side information that describe its properties from different perspectives. In this chapter, we focus on three representative modalities: visual features (e.g., product images), textual descriptions (e.g., titles or reviews), and categorical metadata (e.g., brand or category). These heterogeneous signals complement user-item interactions and help alleviate data sparsity by providing additional semantic context. We use $\mathcal{M} = \{\mathbf{v}, \mathbf{t}, \mathbf{a}\}$ to denote the modalities where $\mathbf{v}$, $\mathbf{t}$ and $\mathbf{a}$ refer to visual, textual and auxiliary information.

Graph-based recommendation further enables the model to capture high-order relationships between users and items through neighborhood propagation. When side information is introduced, items can be connected not only via user interactions but also through semantic similarities derived from different modalities. This motivates the con-

struction of multiple modality-specific graphs that jointly encode user-item interactions and item-item relationships.

2.4 Proposed Framework

2.4.1 Overview

In this section, we elaborate on our framework. Our framework consists of four components: graph structures, aggregation layers, combination layers, and prediction layers. For each type of side information, we construct specific graphs and stack multiple aggregation layers to gather the representations from various modalities. Lastly, we combine the representations from each modality and make the predictions of interaction. The overall framework is illustrated in Figure 2.1.

2.4.2 Graph Structures

Since we have multiple types of side information, we design a specific graph consisting of user-item interactions and relationships between item’s side information. Beyond the interactions, items also contain multiple types of side information. The representations of the input user $\mathbf{u}$ in modality $\mathbf{m}$ are random initialized as $\mathbf{u}_m^0 \in \mathbb{R}^d$ where d is the unified embedding dimension used for users and items, and the representations of the original item $\mathbf{i}$ in modality $\mathbf{m}$ are extracted by the pre-trained models as $\mathbf{i}_m^p \in \mathbb{R}^{d_m}$ and d_m is the original feature dimension of modality m .

Besides the edges of interactions between users and items, we also add edges between items based on their side information in each modality. First, for visual information, we use the pre-trained ResNet50 model [65] to extract the features, and for textual information, we use the pre-trained GloVe model [66] to extract the features. To analyze the relationships between the item’s visual and textual side information, we use a metric to calculate the similarities between the items in each specific modality $\mathbf{m}$. In our framework, we use the cosine similarity [67] metric to measure the similarity between two items. Mathematically, cosine similarity measures the similarity between two vectors of an inner product space. It is measured by the cosine of the angle between vectors

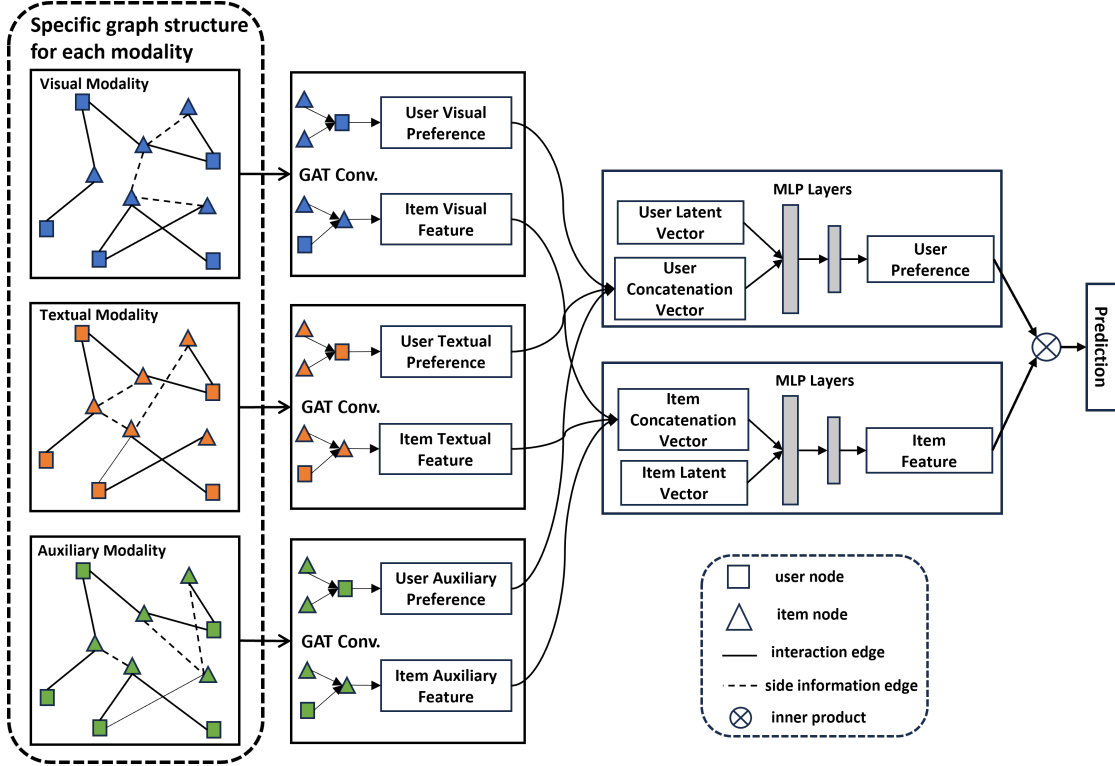


Figure 2.1: The overall framework of our proposed model MRSI-GAT. It consists the specific graphs based on the similarities between items of each modality, GAT aggregation layers, and linear combination layers to generate the user and item representations and make the predictions.

and determines the directions of the vectors. Cosine similarity commonly belongs to the interval $[-1, 1]$, where a higher value means that two vectors are more similar to each other. Given two vectors $\mathbf{a}$ and $\mathbf{b}$, cosine similarity can be formulated as:

$$\text{Similarity}(\mathbf{a}, \mathbf{b}) = \cos(\theta) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|}. \quad (2.1)$$

In the modalities of visual and textual, we calculate the cosine similarities between items, and we set the thresholds s^m for modality $\mathbf{m}$. If the similarity value between two items is higher than the threshold s^m , we add an edge between them which indicates they are similar items in that specific modality. In the modalities of auxiliary information such as brand, category, and genres, these types of side information can be directly

compared and add edges between items. Given a set of items $\mathcal{I}$, for each pair of item $\mathbf{i}$ and $\mathbf{i}'$, let $\mathcal{R}^m = \{r_{ii'}^m \mid \mathbf{i} \in \mathcal{I}, \mathbf{i}' \in \mathcal{I}\}$ be the relationships between side information, where $r_{ii'}^m = 1$ denotes the item $\mathbf{i}$ is similar with item $\mathbf{i}'$ in modality $\mathbf{m}$. For each modality $\mathbf{m}$, we have a specific graph $\mathcal{G}^m$ consists of user-item interaction edges $\mathcal{P}$ and item-item relationship edges $\mathcal{R}^m$, which is $\mathcal{G}^m = \{\mathcal{V}, \mathcal{E}^m\}$. Where $\mathcal{V} = \mathcal{U} \cup \mathcal{I}$ denotes the nodes of users and items, and $\mathcal{E}^m = \mathcal{P} \cup \mathcal{R}^m$ denotes two types of edges. In each modality, we have the same edges of interactions and different edges of relationships between items to construct specific graph structures for different types of side information. Figure 2.2 shows an example of the graph structure of adding edges between items.

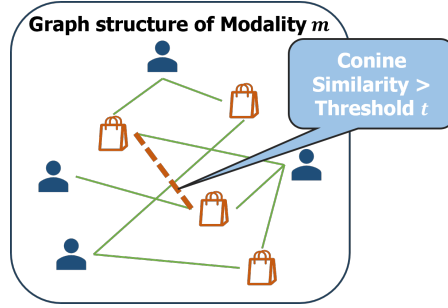


Figure 2.2: The sample graph structure of containing the edges of interactions between users and items and edges between items based on the similarities.

As mentioned before, we extract the features of visual and textual information from the pre-trained models. Based on the different architectures of the pre-trained model, the dimensions of the outputs are different from the user’s dimension $\mathbb{R}^d$. To unify the dimensions of users and items, we apply the multi-layer perception (MLP) to map the item’s input dimensions $\mathbb{R}^{d_m}$ in modalities of visual and textual to the same dimensions as user’s input $\mathbb{R}^d$ as:

$$\mathbf{i}_m^0 = \mathbf{W}^2(\sigma(\mathbf{W}^1 \mathbf{i}_m^p + \mathbf{b})), \quad (2.2)$$

where $\mathbf{W}^1$ and $\mathbf{W}^2$ are the trainable transformation matrix and σ is the activation function. Based on the different dimensions of input vectors, the MLP layer will output the vectors of the same dimension as the user’s input vector $\mathbb{R}^d$, and the outputs of MLP will be the input vectors $\mathbf{i}_m^0 \in \mathbb{R}^d$ of the GNN in modality $\mathbf{m}$. For the auxiliary information that does not process by the pre-trained models, we random initialized the input

representations as $\mathbf{i}_m^0 \in \mathbb{R}^d$ in that modality.

2.4.3 Aggregation Layers

Based on the graph structure, we can utilize the various edges to enrich the representations of users and items. Since we both have edges of interactions and relationships between side information, each edge may aggregate different information into nodes. To be more specific, historical interactions of a user can capture the interest, and edges between items can capture information from deep relationships of other items. These two types of edges can explore deeper connections between side information and interactions.

Basically, other frameworks use the message-passing mechanism of GCN to propagate the information within the bipartite graph. Since bipartite graphs only contain the edges of interactions between users and items, it is reasonable to use GCN in which the nodes are influenced equally by their neighbors. In our framework, we have two types of edges, and they contain different information between nodes, so the message-passing mechanisms of existing GNN efforts (e.g., GCN [68], GraphSAGE [69], LightGCN [70]) are relatively shallow and can not pass the deeper information from the edges. Inspired by the message-passing with attention mechanism of GAT [71], we can calculate different attention values for the neighbors of each node. Given a node of user/item $\mathbf{p}$ and its neighbors $\mathbf{q} \in \mathcal{N}_p$, we can calculate the importance of node $\mathbf{q}$'s features to node $\mathbf{p}$:

$$\mathbf{e}_{pq} = \text{att}(\mathbf{W}\mathbf{h}_p, \mathbf{W}\mathbf{h}_q), \quad (2.3)$$

where $\mathbf{W}$ is a shared linear transformation, parametrized by a weight matrix, and att is a shared attentional mechanism to compute attention coefficients. Based on the graph structure, we can calculate the attention weight of node $\mathbf{p}$'s neighbors $\mathcal{N}_p$:

$$\alpha_{pq} = \text{Softmax}(\mathbf{e}_{pq}) = \frac{\exp(\mathbf{e}_{pq})}{\sum_{k \in \mathcal{N}_p} \exp(\mathbf{e}_{pk})}. \quad (2.4)$$

And then, we can calculate the updated representation of node $\mathbf{p}$ by aggregating from its neighbors $\mathcal{N}_p$ with attention:

$$\mathbf{h}'_p = \sigma\left(\sum_{q \in \mathcal{N}_p} \alpha_{pq} \mathbf{W}\mathbf{h}_q\right). \quad (2.5)$$

Based on our unique graph structure, we have two types of edges, so different types of edges may influence the representations in different aspects. Specifically, for user nodes, they can aggregate neighbor item nodes that they interacted with before with attention, and for item nodes, they can aggregate information from neighbor users and items. Figure 2.3 shows an example of the effect of adding edges between items when aggregating neighbors from 2 hop neighbors.

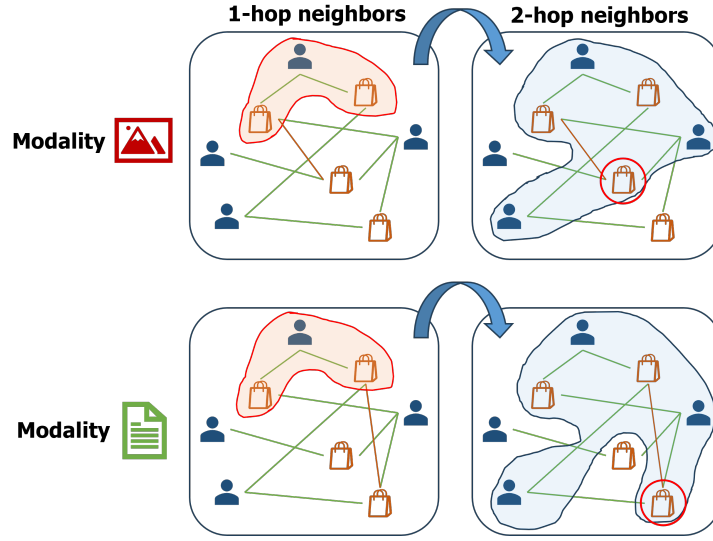


Figure 2.3: An example of effect of adding edges between items. For different modality, the neighbors aggregated to nodes will also be different.

Concretely, item representations will be influenced not only by users but also by similar items in each specific modality. By stacking GAT layers, item nodes will aggregate information from other items and analyze the influence by the attention mechanism, user nodes will aggregate information not only from their directly interacted nodes but also from other related item nodes, which can enrich the user’s preferences and item’s features. Due to the complex relationships between interactions and side information, one single aggregation with attention is relatively shallow. For example, we should not assume that one particular feature will remain the same importance in all cases of message passing. Even though two users purchase the same items, they may not be attracted by the same features, so the features passing to each user should be different. Also, even

though two items have similar features, the importance of features may be different in different cases. Concretely, we need to aggregate the representations with multiple attention mechanisms for various cases. Compared with other GNNs, GAT employed a multi-head attention structure which is suitable for our framework because we need to analyze the representations in multiple aspects. Specifically, we have K attention heads, the output of node $\mathbf{p}$ will be:

$$\mathbf{h}'_p = \parallel_{k=1}^K \sigma \left(\sum_{q \in \mathcal{N}_p} \alpha_{pq}^k \mathbf{W}^k \mathbf{h}_q \right), \quad (2.6)$$

where $\parallel$ represents concatenation, α_{pq}^k is the attention of k -th attention head, and $\mathbf{W}^k$ is the linear transformation. Finally, we can gather the representations of users and items from GAT layers. Since we have different graph structures for various modalities, the output of node $\mathbf{p}$ in modality $\mathbf{m}$ will be:

$$\mathbf{h}_p^m = \parallel_{k=1}^K \sigma \left(\sum_{q \in \mathcal{N}_p^m} \alpha_{pq}^k \mathbf{W}_m^k \mathbf{h}_q^m \right). \quad (2.7)$$

2.4.4 Combination Layers

Assume we have l GAT layers. After generating the representations from multiple GAT layers, we have output vectors of users and items $\mathbf{u}_m^l$ and $\mathbf{i}_m^l$ for each modality $\mathbf{m}$. And then, we combine these vectors to get the representations. First, we combine the output vectors for users and items from each modality $\mathbf{m}$:

$$\mathbf{u}^l = \parallel_m^M (\mathbf{u}_m^l), \quad (2.8)$$

$$\mathbf{i}^l = \parallel_m^M (\mathbf{i}_m^l). \quad (2.9)$$

As we mentioned before, representations for each modality were generated separately from their specific GNNs. So the vectors $\mathbf{u}^l$ and $\mathbf{i}^l$ mainly contain the representations based on the side information. To enrich the representations, we also combine the latent factors of users and items as follows:

$$\mathbf{u}^c = \mathbf{u}^l \parallel \mathbf{u}^e, \quad (2.10)$$

$$\mathbf{i}^c = \mathbf{i}^l \parallel \mathbf{i}^e, \quad (2.11)$$

where $\parallel$ represents concatenation and $\mathbf{u}^c$ and $\mathbf{i}^c$ are the representations.

To find the relationships between modalities and latent factors, we apply the MLP layers to $\mathbf{u}^c$ and $\mathbf{i}^c$ to get the final representations $\mathbf{u}^*$ and $\mathbf{i}^*$:

$$\mathbf{u}^* = \mathbf{W}_u^2(\sigma(\mathbf{W}_u^1 \mathbf{u}^c + \mathbf{b}^u)), \quad (2.12)$$

$$\mathbf{i}^* = \mathbf{W}_i^2(\sigma(\mathbf{W}_i^1 \mathbf{i}^c + \mathbf{b}^i)). \quad (2.13)$$

2.4.5 Prediction Layer

After gathering the final representations of users and items, we can make the final predictions of interactions by multiplying the user and item's vectors:

$$\mathbf{y}' = \sigma(\mathbf{u}^* \cdot \mathbf{i}^{*\top}), \quad (2.14)$$

where σ is a sigmoid activation function to map the predictions to range $[0, 1]$.

2.4.6 Optimization

We apply Bayesian Personalized Ranking (BPR) [72], which is a commonly used pairwise ranking optimization framework, as our learning method. In particular, we model a triplet of one user and two items, one of the items is a positive item which the user interacted, and the other one is a negative item which the user did not interact before:

$$\mathcal{S} = \{(\mathbf{u}, \mathbf{i}, \mathbf{i}') | (\mathbf{u}, \mathbf{i}) \in \mathcal{P}, (\mathbf{u}, \mathbf{i}') \notin \mathcal{P}\}. \quad (2.15)$$

We assume that the user's preference for the positive item is higher than the negative item. Hence, the BPR Loss can be formulated as:

$$\mathcal{L} = - \sum_{(u, i, i') \in \mathcal{S}} \ln \sigma(\mathbf{u}^* \mathbf{i}^{*\top} - \mathbf{u}^* \mathbf{i}'^{*\top}), \quad (2.16)$$

where σ is the sigmoid activation function. We can then train our model based on this loss function.

2.5 Experiments

In this section, we conduct experiments on our model and state-of-the-art methods. We also conduct studies on our proposed model and answer the following questions:

- **RQ1:** How does our model perform compared with the state-of-the-art methods on the real-world datasets?
- **RQ2:** How do different components (e.g., edges between items, latent factors) of our model affect the performance?
- **RQ3:** How do different designs (e.g., thresholds, number of GAT layers) of our model affect the performance?

2.5.1 Datasets

To evaluate our model, we experimented with two public datasets: Amazon Musical Instruments and MovieLens. The characteristics of these datasets are summarized in Table 2.1.

Amazon Musical Instruments. This dataset is publicly available and commonly used in RSs. This dataset contains reviews of the product that is sold on the Amazon Musical Instruments e-commerce platform. Besides interaction records between users and items, this dataset also contains image URLs, descriptions, brand information, and categories as the side information of items. We use ResNet50 model [65] and GloVe model [66] to extract the visual and textual features from images (extracted from image URLs) and descriptions.

MovieLens. This public dataset is also commonly used in RSs. This dataset contains ratings of the movies from the MovieLens website. Besides interaction records between users and movies, this dataset also contains the titles and genres of movies. We use GloVe model [66] to extract the textual features from titles.

For each dataset, we randomly split the training set, validation set, and testing set with the ratio 8:1:1. For each user in the training set, we randomly sampled an item that the user has not interacted with as the negative item. The validation set is used to tune the hyper-parameters, and the testing set is used to evaluate the performance.

Table 2.1: Characteristics of datasets.

Datasets	Interactions	Users	Items	V	T	A
Amazon Musical Instruments	15673	3520	4729	2048	100	32
MovieLens	11269	3006	5017	-	100	32

2.5.2 Baselines

To demonstrate the performances, we compared our proposed method with the following method:

VBPR [73]. This is a benchmark framework in the multi-media recommendation. It combines the side information into the collaborative filtering framework. For a fair comparison, we concentrate the multi-modal features into the item’s representations to make the predictions.

Deep-MINE [58]. This is a multi-modal framework for the multi-media recommendation. It uses Auto Encoder to generate the item’s representations for each modality. It also applies a cognition layer for each user to represent the attention for various side information.

LightGCN [70]. This is an effective model for recommendation by including the most essential ingredients of GCN, which is neighborhood aggregation for collaborative filtering. It learns user and item embeddings by linearly propagating them on the user-item bipartite graph and uses the weighted sum of the embeddings generated from all layers as the final representations. For a fair comparison, we construct a multi-modal Light-GCN, which applies Light-GCN for each specific side information to generate the representations.

GAT [71]. This is a novel neural network architecture that operates on graph-structured data with an attention mechanism. It assigns different attention weights to neighbors when aggregating, which can calculate the importance of each neighbor. It also can be extended to the multi-head attention to gathering deeper relationships between the node and its neighbors. For a fair comparison, we construct a multi-modal GAT, which applies GAT for each specific side information to generate the representations.

MMGCN [48]. This is a GCN-based multi-modal framework. It refines user and item representations by user-item bipartite graph for each type of side information and specific combination layers. The graph structure for each modality remains the same to learn the representations by performing the GCN propagation.

2.5.3 Evaluation Metrics

We use the commonly used evaluation methods in recommendation works: Recall@K and Normalized Discounted Cumulative Gain (NDCG@K) to evaluate the performance of our proposed model, where @K denotes the top-K candidate items of each user. Specially, we set $K = 5, 10$ and calculated the average scores of the above metrics for all users during the testing phase.

2.5.4 Parameter Settings

To train our proposed model, we randomly initialize the parameters and optimize the model with Adam [74], based on the mini-batch size of 64. We searched the learning rate in $\{0.01, 0.001, 0.0001\}$ and the regularization weight in $\{0.001, 0.0001, 0.00001\}$. We searched the number of attention heads in $\{3, 4, 5\}$, note that the number of attention heads is fixed to 1 in the last GAT layer. We show the results of the learning rate 0.001, regularization weight 0.0001, 4 attention heads, and we unified the dimension to 32 as the input and latent vectors of all models.

2.5.5 Performances Comparison (RQ1)

The comparative results are summarized in Table 2.2. Note that “Improv.” denotes the improvement of MRSI-GAT’s result compared with the best results of the baseline models. From the table, we have the following observations:

All GNN-based models outperformed the CF-based models: VBPR and DEEP-MINE in almost all metrics. It demonstrates that graph convolutional networks can better learn the representations from the interactions because most of the data in RSs can be structured into graphs. In our datasets, the interactions can be structured as user-item bipartite graphs.

Table 2.2: Performance comparison between our model and baselines on the Amazon Musical Instruments and MovieLens datasets.

Amazon Musical Instruments				
Model	Recall@5	Recall@10	NDCG@5	NDCG@10
VBPR	0.093	0.147	0.065	0.085
DEEP-MINE	0.094	0.151	0.068	0.094
Light-GCN	0.101	0.201	0.071	0.104
GAT	0.109	0.163	0.079	0.097
MMGCN	0.092	0.244	0.077	0.131
MRSI-GAT	0.134	0.268	0.107	0.165
Improv.	22.93%	9.84%	35.44%	25.95%
MovieLens				
Model	Recall@5	Recall@10	NDCG@5	NDCG@10
VBPR	0.056	0.088	0.037	0.044
DEEP-MINE	0.050	0.091	0.038	0.051
Light-GCN	0.066	0.156	0.059	0.086
GAT	0.074	0.138	0.055	0.059
MMGCN	0.060	0.135	0.049	0.066
MRSI-GAT	0.097	0.189	0.074	0.099
Improv.	31.08%	21.15%	25.42%	15.12%

MRSI-GAT obviously outperforms the other GNN-based baseline models in all metrics, which can verify the effectiveness of our model. The reasons are: first, our model contains the relationships between items in each modality. Compared with baseline models, our model design specific graph for different modalities instead of using the same graph structure, which can analyze the relationships between interactions and side information. Second, our model contains unique combination layers for the user and item concatenation vectors. Instead of simply concatenating the outputs of each modality, our model applies the MLP layers to analyze the relationships between modalities

and the latent factors. Thus the MRSI-GAT framework can improve the performance of the recommendations.

2.5.6 Ablation Study (RQ2)

Effects of Edges Between Items

To demonstrate the effectiveness of the edges between items in each modality, we compared the results of the model without side information edges with our model. The results are summarized in Table 2.3. Note we name the model without side information edge “MMGAT.” From the table, we can find that the side information edges can improve the performance in both datasets.

Effects of Latent Factors

To demonstrate the effectiveness of the latent factors in the combination layer, we compared the results of the model without combining the latent factors with our model. The results are summarized in Table 2.4. Note we name the model without latent factors “GNN vectors only.” From the table, we can find that combining the latent factors can improve the performance in both datasets.

Table 2.3: Performance comparison between our model and the model without side information edges on the Amazon Musical Instruments and MovieLens datasets.

Amazon Musical Instruments				
Model	Recall@5	Recall@10	NDCG@5	NDCG@10
MMGAT	0.123	0.257	0.094	0.144
MRSI-GAT	0.134	0.268	0.107	0.165
MovieLens				
Model	Recall@5	Recall@10	NDCG@5	NDCG@10
MMGAT	0.086	0.170	0.061	0.090
MRSI-GAT	0.097	0.189	0.074	0.099

Table 2.4: Performance comparison between our model and the model without latent factors on the Amazon Musical Instruments and MovieLens datasets.

Amazon Musical Instruments				
Model	Recall@5	Recall@10	NDCG@5	NDCG@10
GNN vectors only	0.126	0.259	0.095	0.146
MRSI-GAT	0.134	0.268	0.107	0.165
MovieLens				
Model	Recall@5	Recall@10	NDCG@5	NDCG@10
GNN vectors only	0.088	0.172	0.063	0.091
MRSI-GAT	0.097	0.189	0.074	0.099

2.5.7 Further Model Analysis (RQ3)

Effects of Thresholds

The thresholds for different modalities will affect the number of edges between items and lead to different performances of the framework. It is important to set the appropriate relationships between items because too many edges may cause information redundancy, and too few edges may cause information scarcity. For different datasets, we need to set different thresholds for various modalities. We searched the threshold of visual modality in $\{0.75, 0.80, 0.85\}$, and textual modality in $\{0.85, 0.90, 0.95\}$ for Amazon Musical Instruments Dataset, the results are summarized in Figure 2.4. From the figure, we can find that the threshold of visual modality in 0.80, and textual modality in 0.95 can have the best performance. We searched the threshold of textual modality in $\{0.75, 0.80, 0.85\}$ for MovieLens Dataset, the results are summarized in Table 2.5. From the table, we can find that the threshold of textual modality in 0.75 can have the best performance. The best thresholds for the two datasets are different because of the specific characters of each dataset.

Effects of Number of GAT Layers

The number of GAT layers will affect the information aggregate to the nodes. It is important to set the appropriate number of GAT layers because too many layers may aggregate redundant nodes, and too few layers may not aggregate enough nodes from the graph. We searched the number of GAT layers in $\{2, 3, 4\}$, the results are summarized in Table 2.6. From the table, we can find that 3 GAT layers can have the best performance.

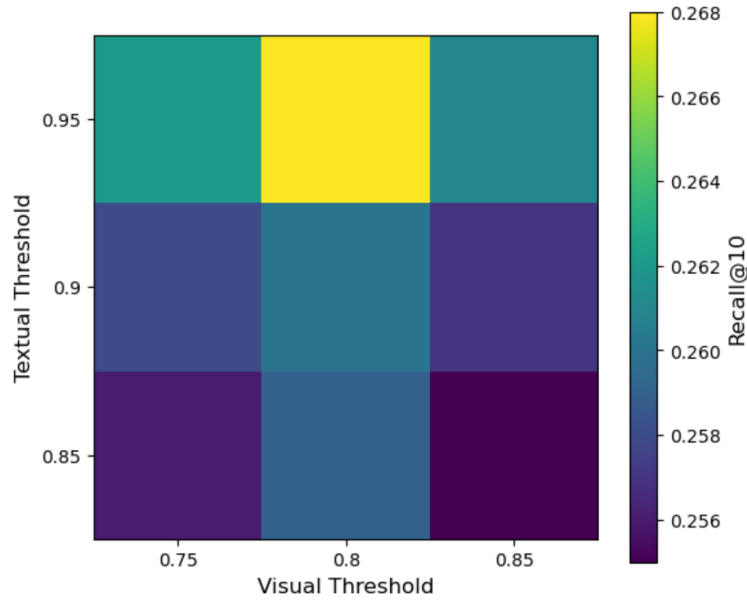


Figure 2.4: Performances comparison on Recall@10 between different thresholds of visual and textual modality for Amazon Musical Instruments dataset.

Table 2.5: Performances comparison on Recall@10 between different thresholds of textual modality for MovieLens dataset.

Threshold	0.75	0.80	0.85
Recall@10	0.189	0.178	0.181

Table 2.6: Performances comparison on Recall@10 between different numbers of GAT layers over two datasets.

Number of Layers	2	3	4
Amazon Musical Instruments	0.259	0.268	0.255
MovieLens	0.177	0.189	0.176

2.6 Conclusion and Future Work

In this chapter, we developed a multi-modal relational side information graph attention network (MRSI-GAT) to improve recommendation performance by modeling the deeper relationships between user-item interactions and multi-modal side information. Different from the multi-modal RSs that process each modality independently and apply shallow feature fusion, MRSI-GAT explicitly constructs modality-specific graph structures based on user-item interactions and item-item relationships derived from side information. By leveraging an attention-based message passing mechanism, the proposed framework adaptively assigns different importance to various relations and modalities, thereby capturing users’ fine-grained preferences and latent correlations among items in a unified framework. This relational design enables the system to utilize complementary semantic characters from textual, visual, and auxiliary modalities, enhancing both the accuracy and explainability of recommendations.

To verify the effectiveness of the proposed model, we conducted experiments on Amazon Musical Instruments and MovieLens. The experimental results show that MRSI-GAT consistently outperforms several state-of-the-art methods. The results confirm that incorporating modality-specific relational structures can effectively enrich user and item representations. Furthermore, experiment results show that the attention mechanism improves explainability by analyzing which modalities and neighboring items contribute most to each recommendation. These findings highlight the importance of integrating relational structures among side information into multi-modal RSs.

Although MRSI-GAT effectively enhances recommendation performance by modeling intra-modality relationships, it still faces several limitations. First, in the current framework, item-item relationships are derived using fixed similarity metrics with

manually defined thresholds, which may not fully capture the relationships between items. Second, each modality is modeled independently, and cross-modal interactions are not explicitly learned, which may limit the system’s ability to represent complementary information between modalities. In the future, we plan to explore adaptive graph construction strategies that can automatically infer item relationships through learnable similarity metrics and dynamic thresholding. We will also investigate cross-modal fusion mechanisms to jointly model dependencies between modalities, thus enhancing representation consistency and robustness.

Chapter 3

Multi-task Learning for Multi-modal Representation Enhancement

3.1 Introduction

RSs have become important tools in modern e-commerce and digital content platforms, helping users efficiently find items that match their interests among an large number of available options. Traditional RSs, such as collaborative filtering (CF) and hybrid models, learn from historical user-item interactions to predict future preferences [75, 76]. However, in practical applications, interaction data are often extremely sparse, as most users interact with only a limited number of items. This sparsity issue limits the model’s ability to learn meaningful user and item representations. To overcome this limitation, recent studies have explored the integration of various types of side information, such as item images, textual descriptions, brands, and categories, to enrich the feature space and improve recommendation accuracy [17, 77].

The emergence of multi-modal RSs (MMRSs) has provided an effective way to leverage such heterogeneous data. These systems extract modality-specific features using deep learning models. Such as convolutional neural networks (CNNs) for visual data and recurrent or transformer-based networks for textual data, and then combine them through fusion layers to form unified user and item representations [55]. Early MMRSs primarily employed late fusion strategies, in which features from each modal-

ity are processed independently and aggregated by concatenation or weighted averaging at the final stage. Although this approach enhances representation richness, it fails to capture the deeper relationships between different modalities. To address this limitation, some research has introduced co-attention and cross-modal mechanisms that allow different modalities to influence each other during the feature learning process [78, 79, 80]. While these methods partially improve feature alignment, they still struggle to model the relationship between multi-modal side information and user behaviors. In other words, they learn how to fuse modalities, but not how user preferences are shaped by each modality.

Meanwhile, GNNs have shown remarkable performance in recommendation tasks due to their ability to capture higher-order connectivity between users and items [49]. By representing users and items as nodes and their interactions as edges, GNN-based RSs can effectively propagate information over multi-hop neighbors, improving collaborative signal utilization. Some studies have combined GNNs with multi-modal learning by constructing multiple user-item bipartite graphs one per modality to capture modality-specific interaction patterns [48, 81]. For example, in a visual modality graph, user-item edges represent visual preferences (e.g., color or style), while in a textual modality graph, they represent content preferences (e.g., descriptions or reviews). After aggregating these graphs, the system fuses modality-specific embeddings to make the final recommendation. However, these methods still rely on shallow fusion and often ignore the implicit correlations between side information and interactions, leading to insufficient alignment between users' behavioral patterns and the semantic characteristics of items.

3.1.1 Challenge

Despite the progress achieved by MMRSs and GNN-based RSs, two critical challenges remain unresolved in effectively integrating side information with user-item interactions. First, most existing methods depend on predefined or manually tuned similarity thresholds to determine whether two items are related based on the side information. Such manually defined metrics can be unstable and fail to generalize across domains or modalities, leading to inefficient and inconsistent graph structures. Second, while

GNN-based MMRs are effective in propagating user-item interaction signals, they rarely ensure that the learned representations are semantically aligned with the original side information. This misalignment between the behavior space (interactions) and the semantic space (features) prevents the model from accurately understanding how user decisions are influenced by different types of side information.

The challenges addressed in this chapter are summarized as follows:

- The first challenge is to automatically learn meaningful relationships between item side information and interactions without relying on manually defined similarity thresholds.
- The second challenge is to align the behavioral representations learned from user-item interactions with the semantic relationships derived from multi-modal side information, enabling the model to better capture modality-specific user preferences.

3.1.2 Contribution

To overcome these challenges, we propose a novel model named Multi-task Multi-modal GNN (MTMM-GNN). MTMM-GNN integrates multi-modal representation learning, GNNs, and multi-task optimization to explore the deeper relationships between user-item interactions and side information. The framework consists of two main tasks: the first task constructs a modality-specific graph for each type of side information, using the edges of user-item interactions to generate modality-aware representations. The second task compares the similarity distributions between item features derived from side information and the corresponding embeddings learned from GNNs, thereby measuring and aligning the consistency between the semantic and behavioral spaces. Through joint optimization, MTMM-GNN can automatically identify which types of side information are most influential for each user’s decision-making process, achieving more accurate and explainable recommendations. From a graph learning perspective, MTMM-GNN is designed to jointly model interaction-driven representations and modality-specific semantic representations through multi-task optimization. Unlike traditional GNN-based RSs that focus solely on propagating collaborative signals, MTMM-

GNN introduces an auxiliary graph-based alignment task to explicitly regularize the consistency between behavioral embeddings and semantic features.

We evaluate MTMM-GNN on three publicly available datasets-Amazon Clothing, ModCloth, and MovieLens. The results show that our model significantly outperforms several state-of-the-art baselines in terms of Recall@K, NDCG@K, and Precision@K. These results confirm the effectiveness of our approach in modeling the complex relationships between user behaviors and multi-modal side information.

To summarize, our contributions are as follows:

- We study the problem of representation inconsistency between interaction-driven embeddings and semantic features from multiple modalities, and formulate it as a multi-task graph learning problem.
- We propose MTMM-GNN, a novel multi-task GNN framework that jointly optimizes interaction-based representation learning and semantic alignment on modality-specific graphs. The auxiliary alignment task acts as a regularization mechanism to guide GNN representation learning toward semantic consistency.
- We conduct experiments on three public datasets, and the results of our proposed model outperform several state-of-the-art models in Recall@K, NDCG@K and Precision@K.

In Section 3.2, we review related works. Then, we introduce some preliminaries in Section 3.3. Section 3.4 elaborates on the details of our proposed method. The experimental results are summarized in Section 3.5. Section 3.6 presents our conclusion.

3.2 Related Work

In this section, we introduce some related works, including multi-modal recommendations, multi-task recommendations, and GNN-based recommendations.

3.2.1 Multi-modal Recommendation

Methods of the multi-modal RS are mainly based on the idea of the combination of multi-media models and CF [55]. Matrix factorization (MF) is an effective CF-based

method, that makes predictions by the representations of users and items. Also, MF can be replaced by the inner product with a neural architecture that can learn the arbitrary function from interactions [82]. Kumar et al. [83] analyzed the recent trends in RSs and the characteristics of various deep learning-based recommendation models. Due to the success of deep learning models for multi-media data, analyzing the side information can enrich the representations and improve performance. In most e-commerce platforms, side information generally includes visual, textual, and auxiliary data (e.g., brand, category, and manufacturer). For each type of side information, there are specific models to analyze the features of data, the methods mainly include convolutional neural network (CNN) [84], recurrent neural network (RNN) [85], and auto-encoder (AE) [86]. Zhang et al. [57] constructed a joint representation learning framework. In this method, each type of side information was adopted to learn the corresponding user and item representations and then integrated various information to make predictions. Truong et al. [87] constructed the framework of combining the visual information and textual information to make the predictions. This method extracted the features to generate the reviews of the user to item based on the user's representations and image features. Then combined the modalities of visual and textual to make predictions. Conceicao et al. [88] constructed a multi-modal framework based on deep learning, pre-trained models, and AE. This method extracted textual and visual features from pre-trained models and used AE to generate low-dimensional sparse representations. Then it produced a ranking of the top-N most relevant items to the user. Guan et al. [58] constructed specific models for different types of side information. In this method, visual features and textual features were processed by stacked auto-encoder. After generating the representations for various side information, they designed an attention layer to assign different weights of various side information for each user and then integrated the latent representations from multiple models. This method used multi-models and stack auto-encoders to extract the features from original side information vectors and trained separately with the prediction model.

Although these methods analyzed multiple types of side information based on multi-models, they failed to explore the deeper relationships between side information and interactions because they did not analyze the high-order connectivity. Concretely, they just extracted the features from various modalities and analyzed separately with deep

learning models to generate the representations, and did not analyze the connections of interactions and the effect of side information.

3.2.2 Multi-task Recommendation

Methods of the multi-task RS are mainly based on the user's behaviors. Also, various types of side information can enrich the representations and improve the performance. Hong et al. [59] constructed multi-models for user reviews and item's various side information. In this method, visual features were processed by CNN and textual features were processed by RNN. Instead of randomly initializing the user's representations, this method also analyzed the user's review records to enrich the representations. Long et al. [89] constructed a multi-task model to analyze the user's historical behavior data. This method divided the behavior data into explicit feedback and implicit feedback, instead of separately analyzing these two types of data, it combined these data as the multi-task model to make to final prediction. This framework had three tasks: two auxiliary tasks, and one main task. Each auxiliary task contained the collaborative filtering framework with neural networks to generate the representations. Lu et al. [90] constructed a multi-task framework that combines probabilistic matrix factorization and adversarial Seq2Seq model. It used the matrix factorization to generate the rate predictions and used the Seq2Seq model to generate user comments on items based on the user's reviews and item comments, which is the explainable recommendation. Gao et al. [91] constructed a multi-task framework based on various types of user behaviors such as views and clicks. This method accounted for the cascading relationship among different types of behaviors such as users needing to click and then make the purchase, and each type of behavior can be treated as a task. Hadash et al. [92] constructed a multi-task framework for ranking and rating. Instead of learning user and item representations to optimize both tasks simultaneously in a multi-task framework, this method generated the predictions based on the facts of the decision process, which first predicts the ranking score and then predicts the rating score.

Although these methods analyzed the user's behaviors and preferences, the side information of items was just aggregated to enrich the item's representations and ignored the relationships between various types of side information and interactions. More con-

cretely, they only analyzed the relationships between various types of interactions between users and items and ignored the relationships between items.

3.2.3 GNN-based Recommendation

Recently, GNN techniques have been widely studied in representation learning for graph-structured data in many computer science tasks beyond RSs, such as node classification, link prediction, and graph-level prediction [61]. In particular, there is an increasing trend of extending GNNs to handle multi-task learning and multi-modal inputs, where different tasks or heterogeneous signals are jointly modeled to improve generalization and robustness. These studies suggest that designing task objectives and representations that properly connect heterogeneous signals is critical for achieving effective graph learning in realistic settings.

Motivated by the strong representation capability of GNNs, GNN-based models have also become popular in RSs, since user behavior data can naturally be represented as graphs [60]. Kumar et al. [93] analyzed the efficacy of graph representation learning by GCNs compared with traditional rule-based methods. In recommendation scenarios, common graph structures include user-item bipartite graphs, sequential graphs, social graphs, and knowledge graphs. For the user-item interaction setting, many existing methods directly apply GNNs on the original user-item bipartite graph [94]. Wang et al. [64] integrated user-item interactions into the embedding learning process by propagating embeddings through higher-order connectivity, thereby encoding collaborative filtering signals into representations.

With the increasing availability of multi-modal side information, recent studies further extend GNN-based recommenders to incorporate heterogeneous item content. Wei et al. [48] constructed a multi-modal GCN framework to learn modality-specific representations, where each modality shares the same user-item bipartite structure while using modality features to enrich representations. Wang et al. [81] additionally introduced a user-user graph to connect users with similar behaviors and strengthen message passing. Jiao et al. [52] constructed modality-specific graph structures by combining user-item interaction edges with item-item relational edges derived from side information similarities using manually set thresholds.

Although these methods demonstrate the effectiveness of combining GNNs with multi-modal information, most of them still focus on propagating signals mainly through interaction edges, and the connection between behavioral representations learned from interactions and the semantic structures implied by side information is not explicitly modeled. Moreover, item-item relations built from side information are often constructed by simple rules (e.g., thresholding similarities), which may not generalize well across datasets and may not provide a principled way to ensure that the learned embeddings remain consistent with the semantic space. These limitations motivate the need for a framework that not only performs modality-aware graph learning, but also explicitly aligns behavioral and semantic spaces under a unified objective, which is the focus of this chapter.

3.3 Preliminaries

In contrast to Chapter 2, which focuses on modeling relational structures across multiple modalities, this chapter emphasizes learning consistent representations between user behaviors and semantic item features. We first introduce the basic recommendation setting and then describe how multi-modal information is incorporated from the perspective of representation alignment.

Let $\mathcal{U}$ and $\mathcal{I}$ denote the sets of users and items, respectively. The historical interactions between users and items are represented by $\mathcal{P} = \{p_{ui} \mid \mathbf{u} \in \mathcal{U}, \mathbf{i} \in \mathcal{I}\}$, where $p_{ui} = 1$ indicates that user $\mathbf{u}$ has interacted with item $\mathbf{i}$. The goal of an RS is to learn a prediction function $f(\mathbf{u}, \mathbf{i})$ that estimates the likelihood of future interactions based on these observed behaviors.

Each item $\mathbf{i} \in \mathcal{I}$ is associated with multiple types of side information derived from different modalities, such as visual, textual, and auxiliary attributes (e.g., brand or category). We denote the modality set as $\mathcal{M} = \{\mathbf{v}, \mathbf{t}, \mathbf{a}\}$. These modalities provide complementary semantic descriptions of items and reflect different aspects that may influence user preferences. In this chapter, multi-modal learning aims not only to incorporate heterogeneous information but also to ensure that the learned representations remain semantically consistent with the original side features.

To model interactions between users, items, and multi-modal information, we adopt

a graph-based formulation where users and items are represented as nodes and interactions form edges. Unlike Chapter 2, where graphs are primarily used to capture relational structures within each modality, the graph formulation here serves as a foundation for learning modality-specific interaction patterns whose representations can be explicitly aligned with semantic similarities.

Finally, the objective of this chapter is to learn low-dimensional embeddings for users and items that jointly encode collaborative signals from interaction data and semantic information from multiple modalities. By enforcing consistency between behavioral embeddings and modality-derived semantic similarities, the learned representations can better reflect user preferences and improve recommendation accuracy.

3.4 Proposed Method

3.4.1 Overview

In this section, we elaborate on our framework. The overall framework is illustrated in Figure 3.1. Unlike existing multi-modal GNN-based RSs that primarily focus on representation aggregation, our framework is designed to explicitly address the inconsistency between behavioral signals derived from user-item interactions and semantic information extracted from item side information.

The key motivation is: although multi-modal features and interaction-based embeddings are often learned jointly, they are typically optimized under different objectives and may encode incompatible similarity structures. As a result, representations learned from graph propagation may not faithfully reflect the semantic relationships implied by side information, even when both are derived from the same items. Our framework not only generates modality-aware representations but also explicitly regularizes their alignment across behavioral and semantic spaces.

To this end, MTMM-GNN adopts a multi-task architecture with two complementary objectives. The first task focuses on learning modality-specific user and item representations through GNNs constructed from user-item interactions, capturing collaborative signals under each modality. The second task evaluates the consistency between these interaction-derived representations and the original modality features by com-

paring their similarity distributions, thereby providing a direct supervisory signal for semantic alignment.

By jointly optimizing these two tasks, the proposed framework enables interaction-driven representation learning while constraining the learned embeddings to preserve modality-specific semantic structures. This design distinguishes MTMM-GNN from prior multi-modal GNN approaches, which typically rely solely on graph propagation or feature fusion without explicitly modeling the consistency between behavioral and semantic representations.

3.4.2 Representations from GNNs

In this module, we introduce the GNN structure, including graph structures based on interactions and aggregation layers; combination layers of representations from GNNs and latent factors, prediction layers, and BPR loss.

Graph Structures

Since we have multiple types of side information, we design a graph consisting of user-item interactions for each modality. For each modality m , we have a graph $\mathcal{G}^m$ consists of user-item interaction edges $\mathcal{P}$, which is $\mathcal{G}^m = \{\mathcal{V}, \mathcal{E}\}$. Where $\mathcal{V} = \mathcal{U} \cup \mathcal{I}$ denotes the nodes of users and items, and $\mathcal{E} = \mathcal{P}^+$ denotes the interactions. The representations of the input user u and item i in modality m are random initialized on $\mathbb{R}^d$ as $u_m^0 \in \mathbb{R}^d$ and $i_m^0 \in \mathbb{R}^d$. Figure 3.2 shows an example of the graph structure based on the interaction records between users and items.

Aggregation Layers

Based on the graph structure of each modality $\mathcal{G}^m$, we can utilize the interaction edges to enrich the representations of users and items. To be more specific, historical interactions can capture the interest of users, and also generate the representations of items of each modality.

Basically, other frameworks use the message-passing mechanism of GCN to propagate the information within the bipartite graph. Since bipartite graphs contain the edges of interactions between users and items, it is reasonable to use GCN in which the nodes

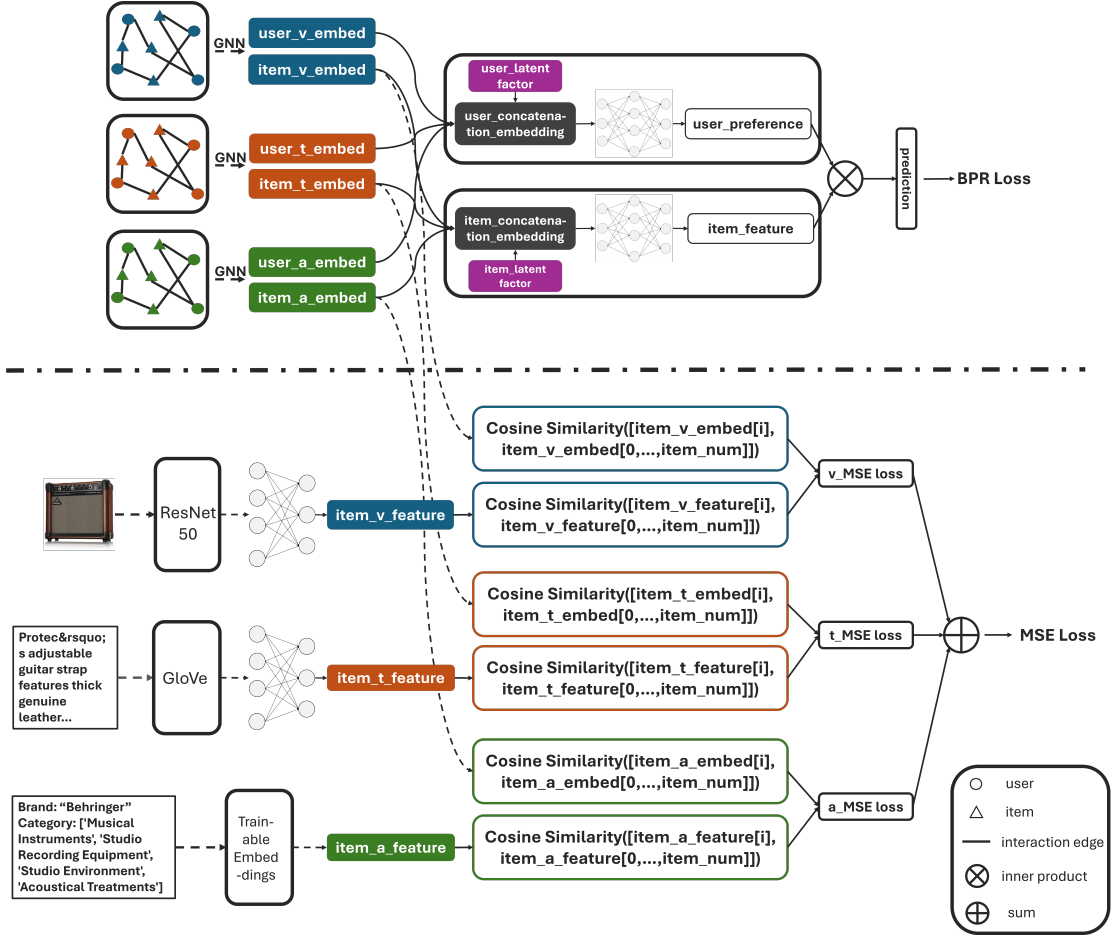


Figure 3.1: The overall framework of our proposed model MTMM-GNN. It consists of 2 tasks: we first generate the representations of each modality from interaction records by GNNs, then we compare the cosine similarities between the representations and the features of the side information and calculate the loss of 2 vectors of similarities. We train our model based on the BPR Loss and MSE loss, which can both analyze the interactions and the effects of different types of side information.

are influenced by their neighbors. In our framework, we have multiple modalities, and they may contain different information between nodes, so the message-passing mechanisms of existing GNN efforts (e.g., GCN [68], GraphSAGE [69], LightGCN [70]) are relatively shallow and cannot pass the deeper information from the edges, especially for different modalities. Hence, we proposed a novel GNN architecture, which integrates

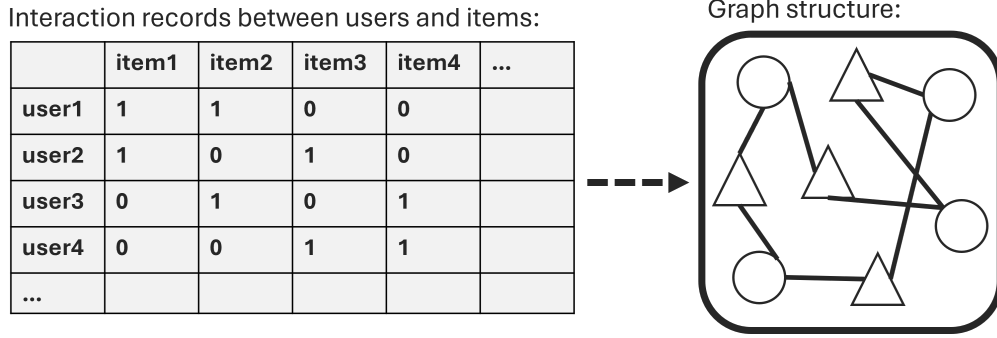


Figure 3.2: The sample graph structure based on the interaction records between users and items.

attention mechanisms into the traditional GCN to enhance the model's ability to capture and focus on important node features during message passing. Given a node of user/item $\mathbf{p}$ and its neighbors $\mathbf{q} \in \mathcal{N}_p$, we first apply a linear transformation to the input node features $\mathbf{h}_p, \mathbf{h}_q$ using the weight matrix and then performs message passing:

$$\mathbf{x}_{pq} = \mathbf{W}^g \cdot \mathbf{h}_q, \quad (3.1)$$

where $\mathbf{W}^g$ is a shared linear transformation, parameterized by a weight matrix, and $\mathbf{x}_{pq}$ is the updated features of node $\mathbf{p}$ and its neighbor $\mathbf{q}$. After calculating the message $\mathbf{x}_{pq}$, we introduce an attention mechanism to compute the attention weights. We apply an attention weight to $\mathbf{x}_p$ and its neighbors $\mathbf{x}_q$:

$$\mathbf{att}_{pq} = \mathbf{W}^a(\mathbf{x}_p, \mathbf{x}_q), \quad (3.2)$$

where $\mathbf{W}^a$ is the learnable attention weights, and then we calculate the attention score α_{pq} of $\mathbf{x}_p$ and its neighbors $\mathbf{x}_q$:

$$\alpha_{pq} = \text{Softmax}(\text{LeakyReLU}(\mathbf{att}_{pq})). \quad (3.3)$$

The attention-weighted message $\mathbf{x}'_{pq}$ is then computed as:

$$\mathbf{x}'_{pq} = \mathbf{x}_{pq} \cdot \alpha_{pq}. \quad (3.4)$$

Finally, the aggregated message $\mathbf{h}'_p$ for node $\mathbf{p}$ is obtained through the following aggregation methods:

1. **Sum Aggregation:**

$$\mathbf{h}'_p = \sum_{q \in \mathcal{N}_p} \mathbf{x}'_{pq}. \quad (3.5)$$

The sum aggregation computes the aggregated message by summing the attention-weighted messages from all neighboring nodes. This is a straightforward aggregation method that preserves the total information from the neighbors.

2. **Mean Aggregation:**

$$\mathbf{h}'_p = \frac{1}{|\mathcal{N}_p|} \sum_{q \in \mathcal{N}_p} \mathbf{x}'_{pq}, \quad (3.6)$$

where $|\mathcal{N}_p|$ is the degree of node $\mathbf{p}$, representing the number of its neighbors. The mean aggregation computes the aggregated message by taking the average of the attention-weighted messages from all neighboring nodes. This method provides a normalized representation of the neighbors and can be beneficial when the number of neighbors varies across nodes.

This proposed message-passing method integrates attention mechanisms into the GCN, allowing the model to focus on relevant node features during message passing. The attention mechanism enhances the model's ability to capture complex relationships and dependencies between nodes in the graph, which can enrich node representations and enhance performance in the specific modality. Figure 3.3 shows the visualization of the attention mechanism of the aggregation layers.

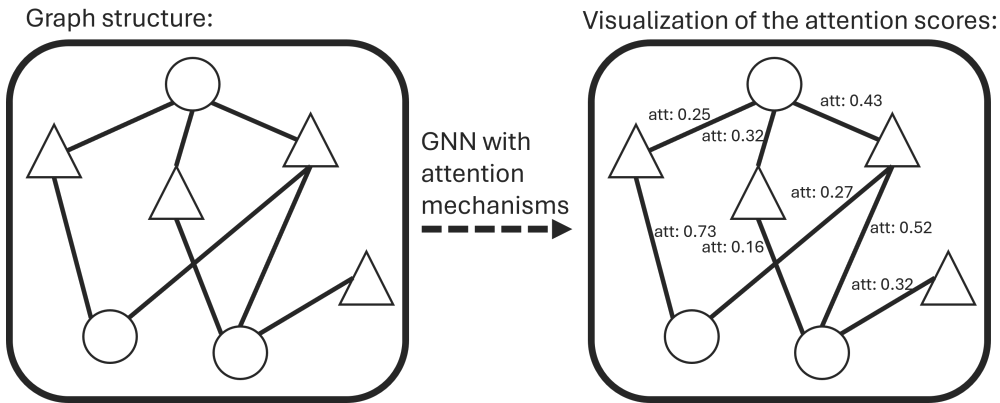


Figure 3.3: The visualization of the attention scores.

Combination Layers

Assume we have l GNN layers. After generating the representations from multiple GNN layers, we have output vectors of users and items $\mathbf{u}_m^l$ and $\mathbf{i}_m^l$ for each modality $\mathbf{m}$. And then, we combine these vectors to get the representations. First, we combine the output vectors for users and items from each modality $\mathbf{m}$:

$$\mathbf{u}^l = \parallel_m^M(\mathbf{u}_m^l), \quad (3.7)$$

$$\mathbf{i}^l = \parallel_m^M(\mathbf{i}_m^l). \quad (3.8)$$

As we mentioned before, representations for each modality were generated separately from their specific GNNs. So the vectors $\mathbf{u}^l$ and $\mathbf{i}^l$ mainly contain the representations based on the side information. To enrich the representations, we also combine the latent factors of users and items $\mathbf{u}^e$ and $\mathbf{i}^e$ as follows:

$$\mathbf{u}^c = \mathbf{u}^l \parallel \mathbf{u}^e, \quad (3.9)$$

$$\mathbf{i}^c = \mathbf{i}^l \parallel \mathbf{i}^e, \quad (3.10)$$

where $\parallel$ represents concatenation and $\mathbf{u}^c$ and $\mathbf{i}^c$ are the representations of all modalities and latent factors.

To find the relationships between modalities and latent factors, we apply the multi-layer perception (MLP) to $\mathbf{u}^c$ and $\mathbf{i}^c$ to get the final representations $\mathbf{u}^*$ and $\mathbf{i}^*$:

$$\mathbf{u}^* = \mathbf{W}_u^2(\text{ReLU}(\mathbf{W}_u^1 \mathbf{u}^c + \mathbf{b}^u)), \quad (3.11)$$

$$\mathbf{i}^* = \mathbf{W}_i^2(\text{ReLU}(\mathbf{W}_i^1 \mathbf{i}^c + \mathbf{b}^i)), \quad (3.12)$$

where $\mathbf{W}_u$ and $\mathbf{W}_i$ are the learnable weight matrix for the fusion MLP layers of users and items.

Prediction Layers

After gathering the final representations of users and items, we can make the final predictions of interactions by multiplying the user and item's vectors with the trainable bias:

$$\mathbf{y}' = \sigma(\mathbf{u}^* \cdot \mathbf{i}^{*\top}) + \mathbf{b}^{u*} + \mathbf{b}^{i*}, \quad (3.13)$$

where σ is a sigmoid activation function to map the predictions to range $[0, 1]$, $\mathbf{b}^{u*}$ and $\mathbf{b}^{i*}$ are trainable bias weights for users and items.

BPR Loss

We apply Bayesian Personalized Ranking (BPR) [72], which is a commonly used pairwise ranking optimization framework, as one of our learning methods. In particular, we model a triplet of one user and two items, one of the items is a positive item with which the user interacted, and the other one is a negative item with which the user did not interact before:

$$\mathcal{S} = \{(\mathbf{u}, \mathbf{i}, \mathbf{i}') | (\mathbf{u}, \mathbf{i}) \in \mathcal{P}, (\mathbf{u}, \mathbf{i}') \notin \mathcal{P}\}. \quad (3.14)$$

We assume that the user's preference for the positive item is higher than the negative item. Hence, the BPRLoss can be formulated as:

$$\mathcal{L}_{BPR} = - \sum_{(\mathbf{u}, \mathbf{i}, \mathbf{i}') \in \mathcal{S}} \ln \sigma((\mathbf{u}^* \mathbf{i}^{*\top} + \mathbf{b}^{u^*} + \mathbf{b}^{i^*}) - (\mathbf{u}^* \mathbf{i}'^{*\top} + \mathbf{b}^{u^*} + \mathbf{b}^{i'^*})), \quad (3.15)$$

where σ is the sigmoid activation function. This $\mathcal{L}_{BPR}$ function contains the characters from the interaction records and it is one of the loss functions of our framework.

3.4.3 Incorporating Side Information

In this module, we introduce the relationships between interactions and various types of side information. We first process the features of side information and then compare the cosine similarities of representations generated from GNNs based on interactions and processed features of side information. Lastly, we calculate the MSE loss of these similarities.

Side Information Pre-processing

Since we have various types of side information, we use the pre-trained models to extract the features from visual and textual information. For visual information, we use the pre-trained ResNet50 model [65] to extract the features, and for textual information, we use the pre-trained GloVe model [66] to extract the features. For the auxiliary information (e.g., brand and category) that are not extracted from pre-trained models, we use a trainable embedding vector $\mathbf{e}^a$ to represent the features.

As mentioned before, we extract the features of visual and textual information from the pre-trained models. Because of the different architectures of the pre-trained model, the dimensions of the outputs are different. Also, the features directly extracted from pre-trained models may contain raw and abstract features that are not suitable for our specific task. To address this issue, we propose to further process the raw features generated from pre-trained models using the MLP layer. For modality $\mathbf{m}$, if we have features generated from pre-trained models $\mathbf{i}_m^p$, we can process the features as:

$$\mathbf{i}_m^{p'} = \mathbf{W}^m \mathbf{i}_m^p + \mathbf{b}_m^p, \quad (3.16)$$

where $\mathbf{W}^m$ is the trainable transformation matrix for modality $\mathbf{m}$, and $\mathbf{i}_m^{p'}$ represents the updated features; otherwise, $\mathbf{i}_m^{p'} = \mathbf{e}^a$. The MLP layer is designed to transform the high-dimensional raw features into a more compact and informative representation that is more suitable for our specific task.

Loss between Similarities

In reality, user's purchase activities are influenced by the similarities of different types of side information between items. For example, some users may be interested in some specific categories and not interested in visual features, and some users may be interested in some specific visual features and not interested in categories. For example, when purchasing clothes, some users are more interested in some specific visual information such as color, and less interested in other side information such as descriptions or categories. Also, some users are more interested in some specific categories such as suits and uniforms, and less interested in colors. For different items, the effects of various types of side information are different, as the example shown in figure 3.4, given 3 items, item1, and item2 have a similar category, and item1 and item3 have similar visual information. Based on the interaction records, item1 and item2 are always been purchased together because of the similar category, but item1 and item3 are not likely to be purchased together even they have similar visual information. So in different modalities, the effects of similarities can also be different depending on various items.

Hence, we first analyze the similarities between the representations generated from GNNs and the side information of each modality. Concretely, we use the cosine similarity [67] metric to measure the similarity between two items. Mathematically, cosine

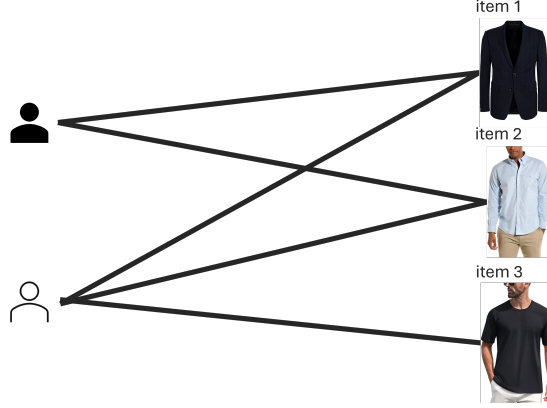


Figure 3.4: Relationships between interaction and different types of side information.

similarity measures the similarity between two vectors of an inner product space. It is measured by the cosine of the angle between vectors and determines the directions of the vectors. Cosine similarity commonly belongs to the interval $[-1, 1]$, where a higher value means that two vectors are more similar to each other. Given two vectors $\mathbf{a}$ and $\mathbf{b}$, cosine similarity can be formulated as:

$$\text{Similarity}(\mathbf{a}, \mathbf{b}) = \cos(\theta) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|}. \quad (3.17)$$

In our framework, we analyze the relationships between the representations and side information features based on the loss between cosine similarities. Concretely, given an item $\mathbf{i}$, we have the representation $\mathbf{i}_m^l$ in modality $\mathbf{m}$, we calculate the cosine similarities with other items in $\mathcal{I}$:

$$\mathbf{S}_r^m = [\cos(\mathbf{i}_m^l[\mathbf{i}], \mathbf{i}_m^l[\mathbf{0}]), \cos(\mathbf{i}_m^l[\mathbf{i}], \mathbf{i}_m^l[\mathbf{1}]), \dots, \cos(\mathbf{i}_m^l[\mathbf{i}], \mathbf{i}_m^l[\mathbf{I}])]. \quad (3.18)$$

Also, given an item $\mathbf{i}$, we have the side information features $\mathbf{i}_m^{p'}$ in modality $\mathbf{m}$, we can calculate the cosine similarities with other items in $\mathcal{I}$:

$$\mathbf{S}_f^m = [\cos(\mathbf{i}_m^{p'}[\mathbf{i}], \mathbf{i}_m^{p'}[\mathbf{0}]), \cos(\mathbf{i}_m^{p'}[\mathbf{i}], \mathbf{i}_m^{p'}[\mathbf{1}]), \dots, \cos(\mathbf{i}_m^{p'}[\mathbf{i}], \mathbf{i}_m^{p'}[\mathbf{I}])], \quad (3.19)$$

where $\mathbf{I}$ is the total number of items in $\mathcal{I}$. Then we can analyze the relationships between these 2 vectors of similarities by the mean squared error (MSE) [95], which is defined as the average of the squared differences between the predicted and actual values. As we

mentioned before, even though items with similar side information are more likely to be purchased, the effect of multiple types of side information for various items can also be different. For example, for items such as books or movies, the textual information (e.g., descriptions) may have a larger effect than the auxiliary information (e.g., brands). Hence, for each item, we apply a trainable bias weight to the MSE loss of similarities between representations and side information features in each modality:

$$\mathcal{L}_{MSE}^m = \text{MSE}(\mathbf{S}_r^m, \mathbf{S}_f^m) \cdot \mathbf{b}^m, \quad (3.20)$$

where $\mathbf{b}^m$ is a trainable bias weight for item i in modality m . We then sum the $\mathcal{L}_{MSE}^m$ of all modalities:

$$\mathcal{L}_{MSE} = \sum_{m \in M} \mathcal{L}_{MSE}^m. \quad (3.21)$$

This $\mathcal{L}_{MSE}$ contains the relationships between item representations generated from interactions and features of side information, and it is another loss function of our framework.

3.4.4 Combined Model Training and Inference

Since we have the loss from interaction records and relationships between side information, we can sum the $\mathcal{L}_{BPR}$ and $\mathcal{L}_{MSE}$ as:

$$\mathcal{L} = \mathcal{L}_{BPR} + \mathcal{L}_{MSE}. \quad (3.22)$$

Then we can train our model with $\mathcal{L}$ as Algorithm 1. Concretely, we first construct the graph for each type of side information based on the interaction records and initialize the inputs of user and item nodes (line 1-4). We then extract the features from the pre-train models, if no pre-train features exist, we randomly initialize the embeddings (line 5-6). For the training loop, we first generate the representations from each modality by GNNs (line 10-11), and then we combine the representations with the latent factors to generate the final representations by MLP layers (line 12-14). We then make the predictions and calculate the $\mathcal{L}_{BPR}$ based on the final representations (line 15-16). Then we process the item's pre-trained features by MLP (line 17-18) and compare the cosine similarities between items on representations generated from GNNs and processed side information

features (line 19-21). Then we calculate and sum the $\mathcal{L}_{MSE}$ of the similarity vectors of each modality with bias (line 22-24). Lastly, we sum the $\mathcal{L}_{BPR}$ and $\mathcal{L}_{MSE}$ to $\mathcal{L}$ to train the model (line 25-27).

Algorithm 1 MTMM-GNN parameter learning**Input:** Set of users and items $\mathcal{U}$ and $\mathcal{I}$, modalities $\mathcal{M}$ and Interaction matrix $\mathcal{P}$ **Output:** Final user representations $\mathbf{u}^*$ and item representations $\mathbf{i}^*$ **Graph Construction:**Construct $\mathcal{G}^m = \{\mathcal{V}, \mathcal{E}\}$ for each modality $\mathbf{m}$, where $\mathcal{V} = \mathcal{U} \cup \mathcal{I}$, $\mathcal{E} = \mathcal{P}^+$ **Initialization:**Initialize the representations $\mathbf{u}_m^0$ and $\mathbf{i}_m^0$ for each modality $\mathbf{m}$ **Feature Extraction:**Extract features using pre-trained models to get $\mathbf{i}_m^p$, or $\mathbf{i}_m^p = \mathbf{e}^a$ if no pre-trained features exist**Training Loop:**epoch = 0 **while** $epoch < train_epochs$ **do****GNN Execution:**Generate representations $\mathbf{u}_m^l$ and $\mathbf{i}_m^l$ using GNN for each modality $\mathbf{m}$ and combine them to generate $\mathbf{u}^l$ and $\mathbf{i}^l$ **Combination and MLP:**Combine latent factors $\mathbf{u}^e$ and $\mathbf{i}^e$ to get $\mathbf{u}^c$ and $\mathbf{i}^c$ Apply MLP to $\mathbf{u}^c$ and $\mathbf{i}^c$ to get final representations $\mathbf{u}^*$ and $\mathbf{i}^*$ **Prediction and BPRLoss:**Compute prediction and $\mathcal{L}_{BPR}$ **Feature Processing:**For modality $\mathbf{m}$, process the features as with MLP layers to get $\mathbf{i}_m^{p'}$ **Similarity Calculation:**Calculate cosine similarities of representations generated from GNNs with other items in $\mathcal{I}$ to get $\mathbf{S}_r^m$ Calculate cosine similarities of side information features with other items in $\mathcal{I}$ to get $\mathbf{S}_f^m$ **MSELoss Calculation:**Calculate the $\mathcal{L}_{MSE}$ of $\mathbf{S}_r^m$ and $\mathbf{S}_f^m$ with trainable bias weight $\mathbf{b}^m$ to get the $\mathcal{L}_{MSE}^m$ Calculate the $\mathcal{L}_{MSE}$ by summing all of $\mathcal{L}_{MSE}^m$ **Combined Model Training and Inference:**Sum $\mathcal{L}_{BPR}$ and $\mathcal{L}_{MSE}$ to get $\mathcal{L}$ Train the model with $\mathcal{L}$

epoch = epoch + 1

3.5 Experiments

In this section, we conduct experiments on our model and state-of-the-art methods. We also conduct studies on our proposed model and answer the following questions:

- **RQ1:** How does our model perform compared with the state-of-the-art methods on the real-world datasets?
- **RQ2:** How do multiple modalities (v, t, a) of our model effect the interactions?
- **RQ3:** How do relationships between interactions and side information of our model affect the performance?
- **RQ4:** How do the item bias weights of each modality affect the task of analyzing the relationships between interactions and side information?

3.5.1 Datasets

To evaluate our model, we experimented with three public datasets: Amazon Clothing, ModCloth and MovieLens. The characteristics of these datasets are summarized in Table 3.1.

Amazon Clothing. This dataset is publicly available and commonly used in RSs. This dataset contains reviews of the products that were sold on the Amazon Clothing e-commerce platform. Besides interaction records between users and items, this dataset also contains image URLs, descriptions, brands, and categories as the side information of items. We use the ResNet50 model [65] and the GloVe model [66] to extract the visual and textual features from images (extracted from image URLs) and descriptions.

Modcloth. This public dataset is commonly used in RSs. This dataset contains reviews of the products that were sold on the Modcloth website. Besides interaction records between users and movies, this dataset contains various types of information on items and review text of users. We use the GloVe model [66] to extract the textual features from the basic information and reviews of the items.

MovieLens. This public dataset is also commonly used in RSs. This dataset contains ratings of the movies from the MovieLens website. Besides interaction records

between users and movies, this dataset also contains the titles and genres of movies. We use the GloVe model [66] to extract the textual features from the titles of the movies.

For each dataset, we randomly split the training set, validation set, and testing set with the ratio 8:1:1. For each user in the training set, we randomly sampled an item that the user has not interacted with as the negative item. The validation set is used to tune the hyper-parameters, and the testing set is used to evaluate the performance.

Table 3.1: Characteristics of datasets.

Datasets	Interactions	Users	Items	V	T	A
Amazon Clothing	16839	3950	5059	2048	100	32
Modcloth	12570	2720	965	-	100	32
MovieLens	13191	3367	5496	-	100	32

3.5.2 Baselines

To demonstrate the performances, we compared our proposed method with the following method:

VBPR [73]. This is a classical baseline for incorporating side information into collaborative filtering. It extends matrix factorization by integrating item-level content features to enhance user and item representations. In our experiments, multi-modal features are aggregated into unified item representations to ensure a fair comparison with MTMM-GNN.

Deep-MINE [58]. This is a deep multi-modal recommendation framework that learns modality-specific item representations using auto-encoders. It further introduces a cognition layer to model user-specific attention over different modalities. This model serves as a representative approach that focuses on feature-level fusion and user-aware modality weighting without explicit graph modeling.

LightGCN [70]. This is an efficient graph-based collaborative filtering model that captures high-order connectivity by linear neighborhood aggregation on the user-item interaction graph. To adapt it to the multi-modal setting, we construct a multi-modal

variant by applying LightGCN separately to interaction graphs associated with different types of side information, and then combining the resulting embeddings.

GAT [71]. This introduces an attention mechanism into GNNs, enabling nodes to assign different importance weights to their neighbors during aggregation. This allows the model to capture heterogeneous influences in graph-structured data. For comparison, we extend GAT to a multi-modal setting by learning modality-specific representations through independent GAT models for each type of side information.

MMGCN [48]. This is a graph-based multi-modal RS that refines user and item representations by performing GCN propagation over user-item bipartite graphs for each modality. Item node features are initialized using modality-specific side information, and the learned representations are fused through dedicated combination layers. MMGCN represents a typical approach that jointly exploits interaction structures and raw modality features.

MRSI-GAT [52]. This is a GAT-based multi-modal framework. It constructs different graph structures based on the characters of various modalities to enrich the representations. It also analyzes the relationships between interactions and side information by adding edges based on the thresholds of the similarities between the item’s side information. Each modality contains specific graph structures that combine the edges of interactions and similarities between items.

3.5.3 Evaluation Metrics

We use the commonly used evaluation methods in recommendation works: Recall@K, Normalized Discounted Cumulative Gain (NDCG@K) and Precision@K to evaluate the performance of our proposed model, where @K denotes the top-K candidate items of each user. Specifically, we set $K = 5, 10$ and calculated the average scores of the above metrics for all users during the testing phase.

3.5.4 Parameter Settings

To train our proposed model, we randomly initialize the parameters and optimize the model with Adam [74], based on the mini-batch size of 64. We searched the learning rate in $\{0.01, 0.001, 0.0001\}$ and the regularization weight in $\{0.001, 0.0001, 0.00001\}$.

We show the results of the learning rate 0.001, and regularization weight 0.0001, and we unified the dimension to 32 as the input and latent vectors of all models.

3.5.5 Performances Comparison (RQ1)

The comparative results are summarized in Table 3.2. Note that “Improv.” denotes the improvement of MTMM-GNN’s result compared with the best results of the baseline models. From the table, we have the following observations:

All GNN-based models outperformed the CF-based models: VBPR and DEEP-MINE in almost all metrics. It demonstrates that graph convolutional networks can better learn the representations from the interactions because most of the data in RSs can be structured into graphs. In our datasets, the interactions can be structured as user-item bipartite graphs for each modality. The graph structures contain the nodes of users and items and edges of interaction records (users purchase items).

MTMM-GNN obviously outperforms the other GNN-based baseline models in all metrics, which can verify the effectiveness of our model. The reasons are: first, our model analyzes not only the interaction records but also the relationships between interactions and various types of side information. We use the similarities between items to analyze the relationships between interactions and side information, which can also avoid manually setting up the thresholds. Also, our model contains the latent factors of users and items and combined with the vectors generated from GNNs, which leverages the strengths of both latent factor models and GNNs, enabling the model to better generalize from observed data to make accurate predictions.

3.5.6 Ablation Study (RQ2-4)

We analyze the effect of multi-modalities, multi-task learning, and bias weight on the performance of our model.

Effect of Multi-modalities (RQ2)

To demonstrate the effectiveness of multiple modalities, we compared the results of the model with the single modality model. In the Amazon Clothing dataset, the side infor-

Table 3.2: Performances comparison between our model and baseline models over three datasets.

Model	Amazon Clothing					
	Recall@5	Recall@10	NDCG@5	NDCG@10	Precision@5	Precision@10
VBPR	0.1020	0.1997	0.0932	0.1012	0.0610	0.0592
DEEP-MINE	0.1052	0.2035	0.0989	0.1116	0.0617	0.0601
LightGCN	0.2002	0.3045	0.1524	0.1928	0.1203	0.1133
GAT	0.1978	0.2994	0.1491	0.1849	0.1118	0.1062
MMGCN	0.2104	0.3221	0.1823	0.2003	0.1245	0.1195
MRSI-GAT	0.2206	0.3551	0.2001	0.2108	0.1374	0.1328
MTMM-GNN	0.2505	0.4082	0.2303	0.2429	0.1503	0.1459
Improv.	13.55%	14.95%	15.09%	15.23%	9.39%	9.86%
Model	Modcloth					
	Recall@5	Recall@10	NDCG@5	NDCG@10	Precision@5	Precision@10
VBPR	0.0912	0.1771	0.0782	0.0823	0.0550	0.0530
DEEP-MINE	0.0984	0.1820	0.0852	0.0912	0.0561	0.0539
LightGCN	0.1292	0.2021	0.1232	0.1461	0.0799	0.0606
GAT	0.1303	0.2059	0.1249	0.1490	0.0811	0.0627
MMGCN	0.1554	0.2290	0.1402	0.1648	0.0944	0.0722
MRSI-GAT	0.1701	0.2507	0.1607	0.1809	0.1022	0.0861
MTMM-GNN	0.1944	0.2859	0.1828	0.2070	0.1150	0.0950
Improv.	14.29%	14.04%	13.75%	14.42%	12.52%	10.34%
Model	MovieLens					
	Recall@5	Recall@10	NDCG@5	NDCG@10	Precision@5	Precision@10
VBPR	0.0732	0.1261	0.0591	0.0702	0.0426	0.0377
DEEP-MINE	0.0801	0.1340	0.0629	0.0829	0.0433	0.0386
LightGCN	0.1273	0.2249	0.1121	0.1309	0.0751	0.0675
GAT	0.1280	0.2250	0.1197	0.1401	0.0755	0.0681
MMGCN	0.1402	0.2403	0.1305	0.1523	0.0837	0.0702
MRSI-GAT	0.1567	0.2521	0.1498	0.1689	0.0939	0.0797
MTMM-GNN	0.1753	0.2828	0.1675	0.1877	0.1049	0.0879
Improv.	11.87%	12.18%	11.82%	11.13%	11.71%	10.29%

mation includes image, description, brand, and category. In the Modcloth dataset, the side information includes size, category and review. In the MovieLens dataset, the side

information includes title and genre. The results of models only contain one modality and the multi-modalities are shown in Figure 3.5. Note that we named the single modality of image, description, brand, and category of Amazon Clothing as “model-i, model-d, model-b, model-c.” We named the single modality of size, category and review of Modcloth as “model-s, model-c, model-r.” We named the single modality of title and genre of MovieLens as “model-t, model-g.”

From the results in Figure 3.5, we can find that the model with multiple modalities outperforms the model with a single modality, which means that the purchase activities are based on multiple aspects of items. Also, models with different types of side information have different performances, which shows the importance of side information in different datasets. Specifically, in the Amazon Clothing dataset, the image modality got the best performance and the category modality got the worst performance; in the Modcloth dataset, the review modality got the best performance and the size modality got the worst performance; in the MovieLens dataset, the title modality got the best performance and the genre modality got the worst performance. Even though different modalities have various distributions, fusing representations from different modalities can improve performance.

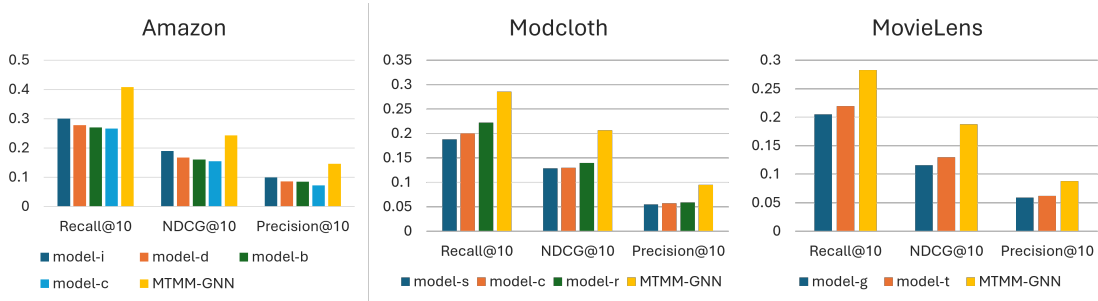


Figure 3.5: Effect of multi-modalities.

Effect of Multi-task Learning (RQ3)

As we mentioned before, user’s purchase activities are influenced by the similarities of different types of side information between items. Concretely, in some specific modality, items with similar side information may affect the interaction which means they are more likely to be purchased together. Also, in some modalities, similarities between

side information may not affect the interaction. To demonstrate the effectiveness of similarities of various types of side information, we compared the results of our model with the single task model. The results of models only containing task1 and multi-task model are shown in Figure 3.6. Note that we named the single task model of task1 as “model-task1.”

From the results in Figure 3.6, we can find that the model with multi-task learning outperforms the model with single-task learning, which can verify the effectiveness of the similarities between side information. Concretely, in the first task of our model, we generate the representations only from interactions, which means that the inputs of the GNNs of each modality are randomly initialized embeddings and are updated based on the interaction records. Hence, the representations of task1 only contain the characters of interactions of each modality. In the second task, we compare the similarities between the representations from task1 and the features of side information to analyze the relationships between interactions and side information, which can improve the performance.

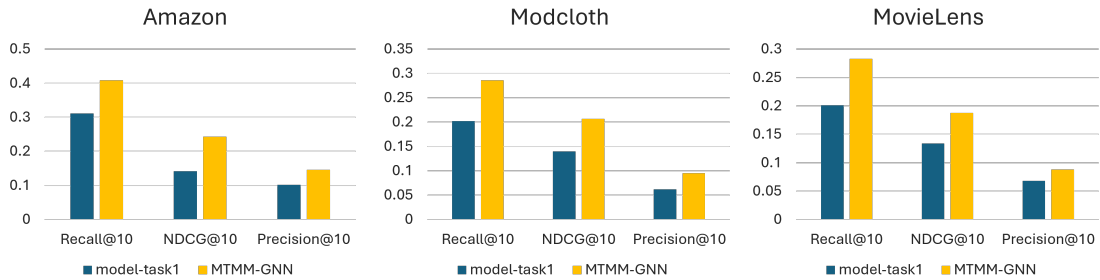


Figure 3.6: Effect of multi-task learning.

Effect of Bias Weight of Each Modality (RQ4)

Even though in some modalities, similar items are more likely to be purchased together, the effects of similarities between features of side information on interactions for different items are also different. To demonstrate the effectiveness of bias weight of different items in task2, we compared the results of our model with the model without item bias. The results of models without item bias and models with item bias are shown in Figure 3.7. Note that we named the model without item bias as “model-no-bias.”

From the results in Figure 3.7, we can find that the model with item bias outperforms the model without item bias, which can verify the effectiveness of the item bias. Concretely, the effects of different types of side information for different items on interactions are not the same. For example, even though two items may have similar features in the specific modality, they are not likely to be purchased together. Figure 3.8 shows an example of the effect of bias. We visualize the embeddings of 3 items in modalities of brand and category (Modality-b and Modality-c), and these items contain similar side information. We can find that in the modality of brand, item1 and item2 have similar representations, and in the modality of category, item1 and item3 have similar representations. Even though these items have similar side information, their representations are not similar because the bias weights of each item in different modalities are different.

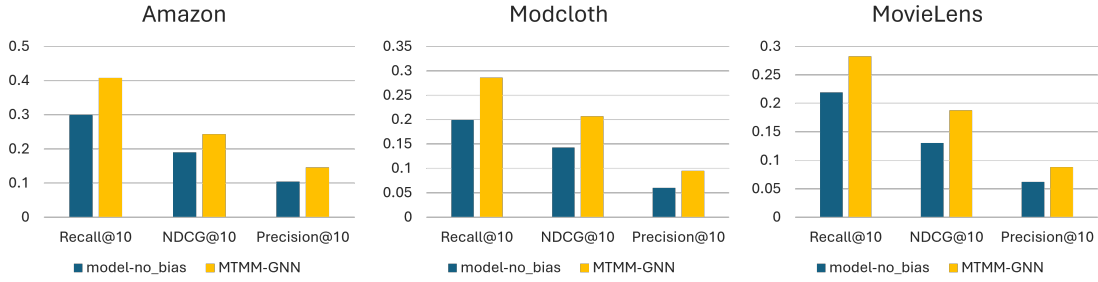


Figure 3.7: Effect of item bias weight.

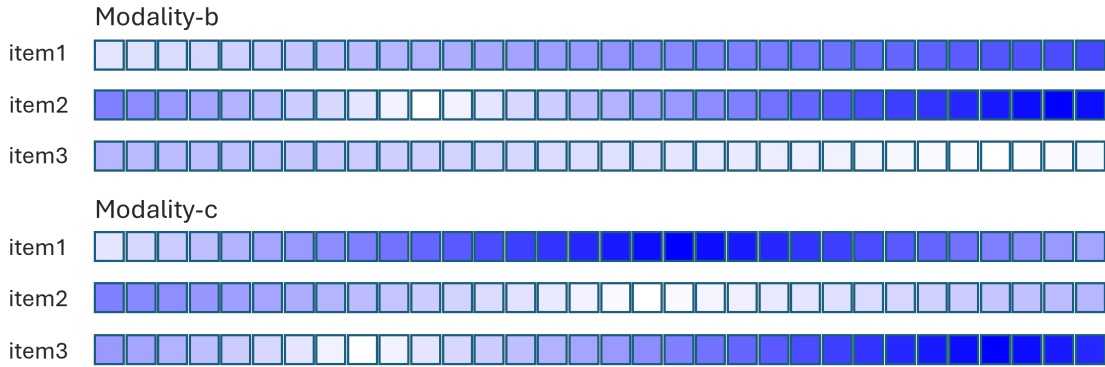


Figure 3.8: Examples of effect of bias.

3.5.7 Further Model Analysis

Here we further analyze the effect of the number of GNN layers and the aggregation methods.

Effect of GNN Layers

The number of GNN layers will affect the information aggregated to the nodes. It is important to set the appropriate number of GNN layers because too many layers may aggregate redundant nodes, and too few layers may not aggregate enough nodes from the graph. We searched the number of GNN layers in $\{2, 3, 4\}$, and the results are summarized in Figure 3.9. From the table, we can find that 3 GNN layers can have the best performance.

Theoretically, shallow GNNs struggle to capture high-order collaborative signals, while deeper architectures suffer from over-smoothing, where node embeddings become indistinguishable due to excessive neighborhood aggregation. Several research[49] also demonstrates that 3-layer GNNs achieve optimal performance because they can balance the trade-off between capturing multi-hop user-item interactions and avoiding performance degradation from over-smoothing.

Comparison of Aggregation Methods

To further investigate the impact of aggregation methods in GNN layers, we compare the performance using both sum and mean aggregation methods. The results are summarized in Table 3.3.

Table 3.3: Performances comparison on Recall@10 using different aggregation methods over three datasets.

Aggregation Method	Amazon Clothing	Modcloth	MovieLens
Sum	0.3926	0.2785	0.2746
Mean	0.4082	0.2859	0.2828

From the table, we can find that the mean aggregation method outperforms the sum aggregation method on all datasets. The mean aggregation method calculates the av-

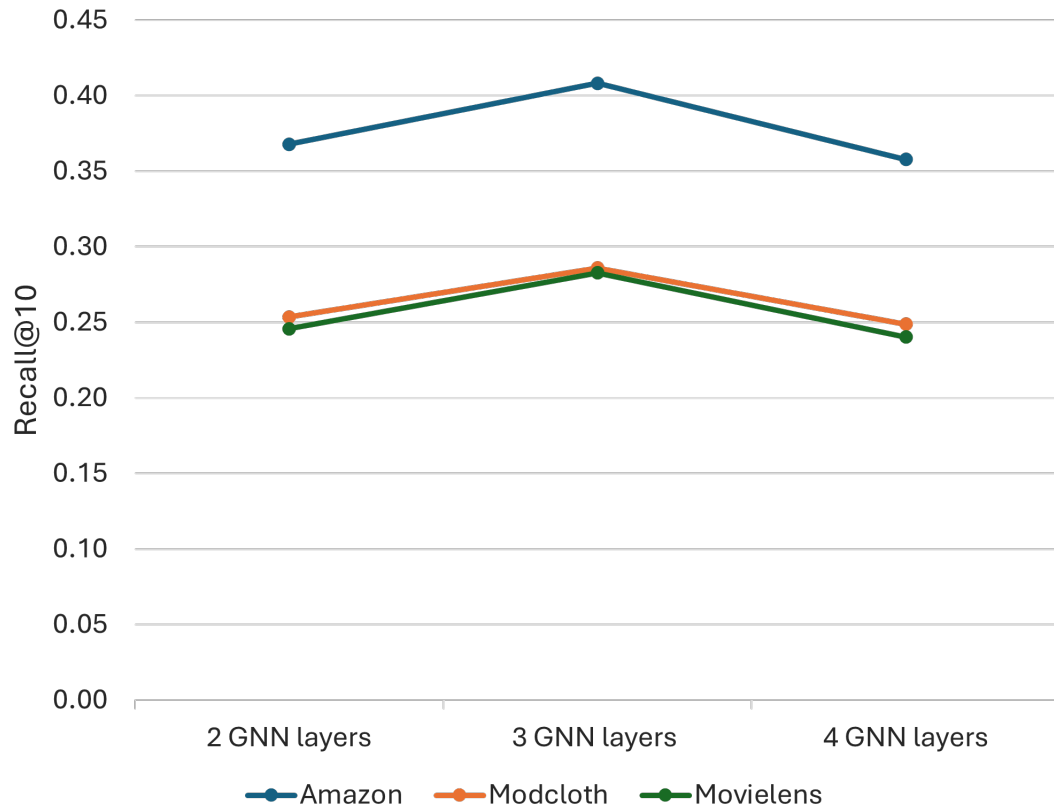


Figure 3.9: Performances comparison on Recall@10 between different numbers of GNN layers over three datasets.

erage of the neighbor node features, which can preserve more information and is less likely to be dominated by nodes with extremely high or low values. Theoretically, mean aggregation can better normalize node degrees and reduce bias toward high-degree entities, which is critical for representation learning in graphs. This leads to a more balanced and effective aggregation of node

3.6 Conclusion

In this chapter, we explored the deeper relationships between user-item interactions and side information in multi-modal RSs. To achieve this, we proposed a Multi-task Multi-modal GNN (MTMM-GNN), which jointly learns user and item representations from

multiple modalities such as visual, textual, and auxiliary information. Each modality is modeled through a user-item bipartite graph, and the modality-specific embeddings are fused through MLP layers together with latent factors to form unified representations. Furthermore, MTMM-GNN introduces an auxiliary learning objective that compares the similarities between learned representations and side information features, thereby aligning semantic and behavioral signals across modalities. This joint modeling enables the system to utilize both interaction data and side information more effectively, resulting in improved recommendation accuracy and explainability.

Experiments conducted on three public datasets, including Amazon Clothing, ModCloth, and MovieLens, demonstrated that MTMM-GNN consistently outperforms state-of-the-art baselines in Recall@K, NDCG@K, and Precision@K. The results verify the effectiveness of modeling the relationships between interactions and multi-modal side information, as well as the benefit of explicitly aligning the similarities between user-item representations and item attributes.

However, MTMM-GNN still faces several limitations. The current similarity-based objective assumes that user-item relationships can be well captured by correlations between side information features, while real-world purchase decisions may depend on more complex factors. Even though we apply bias weights for each item to analyze the influence of similarities in each modality, it only can adjust the loss of similarities and not include other types of information that can affect interactions. Moreover, most existing datasets provide only limited types of modalities, typically include visual, textual, and categorical information. In real-world applications, user preferences may also be influenced by other factors that are not included in the single domain. As a result, the model cannot fully capture the richness of user behavior patterns in realistic environments.

In the future, we plan to extend this work by exploring more adaptive and flexible mechanisms for modeling the relationships between interactions and side information. This includes incorporating additional contextual and semantic signals, and developing methods that can deeper capture the user interests. We also aim to enhance the explainability of multi-modal RSs by investigating how different modalities influence user decisions in a more transparent and explainable way.

Chapter 4

LLM-augmented Counterfactual Attention for Reliable Recommendations

4.1 Introduction

In the previous chapter, we investigated how multi-modal side information, such as visual, textual, and auxiliary features, can be leveraged to improve the quality of user and item representations in RSs. The proposed MTMM-GNN framework demonstrated that aligning behavioral representations learned from interactions with semantic features extracted from side information can enhance both accuracy and explainability. However, in practice, the types of side information available in public datasets still remain limited, and the information they capture often fail to reflect the full complexity of users' real-world decision processes. For example, product descriptions or images may represent surface-level attributes, but user preferences are also influenced by contextual, emotional, and cognitive factors that are not included in the original dataset. Therefore, developing methods that can expand the semantic space of items and reason about how enriched information influences user choices has become an important research direction in the next generation of RSs.

Recently, GNNs have emerged as powerful tools for learning from user-item in-

teractions due to their ability to capture high-order connectivity and collaborative signals across users and items [49]. Through neighborhood aggregation, GNN models can propagate information among connected nodes [96]. Meanwhile, knowledge graph (KG)-enhanced recommenders [56, 80] extend this framework by connecting items to rich, typed entities—such as image, brand, category, thereby enabling information propagation through item-entity relations. These models inject explicit semantic knowledge into item representations and have been shown to improve both accuracy and explainability. However, public datasets typically contain only a few pre-defined entity types, making the constructed KGs relatively shallow. Consequently, current KG-based approaches struggle to represent the diverse and fine-grained attributes that drive user preferences in realistic settings.

In parallel, large language models (LLMs) have demonstrated strong capabilities in understanding and generating structured semantic information from unstructured text. Recent studies have explored using LLMs to enrich item attributes, such as extracting latent characteristics, summarizing customer reviews, or generating explainable features that better describe product qualities [97, 98]. By integrating these enriched representations into recommendation pipelines, LLMs can help fill the semantic gap between user behavior and available side information.

4.1.1 Challenge

Despite the success of GNN-based and KG-enhanced RSs, two core challenges remain when incorporating LLM-augmented semantics into graph-based recommendation frameworks. First, existing KGs are still limited by the predefined entities and relations within available datasets. This restriction prevents models from capturing fine-grained and context-dependent semantics that influence user decisions in reality. For example, emotional tone, cultural background, or thematic factors often implicit in item descriptions, cannot be directly modeled within existing KG structures. Expanding the entity space through LLMs introduces new opportunities but also requires careful integration to preserve semantic coherence and avoid noise. Second, most current attention-based GNNs learn global importance weights over entities based solely on graph topology, less effort on user-specific differences in how entities contribute to preference forma-

tion. Such non-personalized attention fails to model the diversity of user motivations and limits the explainability of recommendation results.

The challenges addressed in this chapter are summarized as follows:

- To enhance the representational power of KGs, it is necessary to incorporate additional entities and relations beyond those available in existing datasets. The main challenge lies in how to utilize LLMs to generate new, reliable, and semantically consistent entities that can be integrated with user-item interaction graphs.
- Traditional attention mechanisms capture only correlation-based relationships and are shared across users. A new approach is needed to model the causal influence of each entity on individual users' decisions, enabling personalized and explainable recommendations.

4.1.2 Contribution

To overcome these challenges, we propose a novel framework named LLM Augmentation Enhanced Counterfactual Attention for Graph-based RSs (LACAT). LACAT combines the semantic generative capability of large language models with the structural learning power of GNNs to achieve both performance improvement and explainability enhancement. Specifically, we first employ a survey-guided LLM augmentation strategy to expand the item-entity KG. The LLM generates new semantic attributes and relations for each item, enriching the side information space and providing a more comprehensive understanding of items from multiple perspectives. Next, we introduce a counterfactual attention mechanism that quantifies the causal influence of each entity on a user's preference. Unlike traditional attention, this mechanism explicitly compares factual and counterfactual conditions-measuring how removing or altering an entity changes the recommendation outcome. In this way, LACAT can learn personalized entity importance for each user, leading to more explainable and accurate predictions.

By integrating LLM-based augmentation and counterfactual attention into a unified GNN architecture, LACAT provides an effective way to bridge the gap between semantic richness and behavioral reasoning in RSs. We evaluate our model on two public datasets-Amazon Movies and TV and MovieLens, and demonstrate that LACAT

significantly outperforms several state-of-the-art baselines in terms of Recall@K and NDCG@K. Also, qualitative analysis shows that our counterfactual attention module can provide intuitive explanations at the entity level, showing which augmented entities most strongly influenced a given recommendation.

To summarize, the main contributions of this chapter are as follows:

- We introduce a survey-guided LLM augmentation strategy that enriches the entities to the item-entity KG, thus creating richer connections between user-item interactions and side information.
- We propose a novel method LACAT, that jointly learns user, item, and entity representations by different GNNs and assigns personalized, counterfactual-based attention weights to each entity, delivering both high accuracy and clear, entity-level explanations.
- We conduct experiments on two public datasets, and the results of our proposed model outperform several state-of-the-art models in Recall@K and NDCG@K.

Section 4.2 gives a brief overview of related works. Section 4.3 we introduce some preliminaries in this chapter. Section 4.4 presents our proposed method, and experimental results are illustrated in Section 4.5. Finally, this chapter is concluded in Section 4.6.

4.2 Related Work

In this section, we introduce some related works on multi-modal recommendations and GNN-based recommendations.

4.2.1 Multi-modal Recommendation

The methods of the multi-modal RS are mainly based on combining multimedia models with CF [54]. Matrix factorization (MF) is an effective CF-based method whose inner product can be replaced with a neural architecture that learns a non-linear interaction function from interactions [82]. With the success of deep learning models on

multimedia data, analyzing side information can enrich the representations and improve performance.

On most e-commerce platforms, side information usually includes visual, textual, and auxiliary data. Analyzing multi-modal information not only improves accuracy but also strengthens explainability. Gao et al. [99] surveyed causal inference in RSs. Since the real world is driven by causality, not just correlation, this paper analyzes the relationships between different factors and interactions, which can enhance the explainability of recommendations. Recently, LLMs such as ChatGPT, which encode knowledge from diverse domains, and enriched the modalities of the RS. Ma et al. [97] proposed a framework that combines collaborative signals and LLMs to generate comprehensive and meaningful explanations based on interaction records. Their work leverages the generative capability of LLMs to enhance the explainability of recommendations. Zhang et al. [57] proposed a joint representation learning framework that used each type of side information to learn user and item representations, and these representations are integrated to make predictions. Truong et al. [87] proposed a framework that combines visual information and textual information. This model extracts image features together with user representations to generate reviews and then uses both modalities for prediction. Conceicao et al. [88] proposed a multi-modal framework based on deep learning, pre-trained models, and auto-encoders (AE). Textual and visual features from pre-trained models are fed into AE to produce low-dimensional sparse representations, which are used to rank the top-N relevant items for each user. Guan et al. [58] proposed specific models for different types of side information. Visual and textual features are processed by stacked AEs; an attention layer assigns different weights to various side information for each user before integrating the latent representations. Chen et al. [100] proposed an attribute-aware collaborative filtering model that analyzes the relationships between interactions and side information to provide personalized and explainable recommendations.

4.2.2 GNN-based Recommendation

Recently, GNNs have been widely adopted in RSs because they provide powerful representation learning for graph-structured data [49]. In most recommendation scenarios,

data can be modeled as graphs.

For our task, most existing approaches [94] apply GNNs directly to the original user-item bipartite graph. Wang et al. [64] integrated user-item interactions into the embedding process by propagating embeddings along high-order connections, thus encoding collaborative filtering signals into the representations. Inspired by the ideas of multi-modalities, Wei et al. [48] proposed a multi-modal graph convolution framework that constructs one user-item bipartite graph for each type of side information. Within every modality-specific graph, user and item nodes aggregate modality-specific features to enrich their vectors. Wang et al. [81] extended this design by adding a graph of user relationships. Node features from each modality are first encoded, and then a Top-K method connects users with similar purchase histories, further enhancing the learned representations. Wang et al. [80] proposed a framework that unifies users, items, and entities into a heterogeneous KG and employs relation-aware attention to propagate signals along high-order relations, enabling the model to capture semantic relations and refine the representations. Sun et al. [56] proposed a heterogeneous KG by combining visual and textual features of entities. Then, cross-modal, relation-aware attention is applied to jointly propagate structural, visual, and textual signals along high-order paths, further enhancing user-item representations with complementary multi-modal semantics and explainability. Wang et al. [101] proposed an entity-enhanced light KG convolutional network that captures high-order connections among users, items, and entities in a heterogeneous KG. This framework assigns relation-aware attention when aggregating neighbors and adds an entity-preference task to align user-entity signals with interactions, which can improve the accuracy and explainability.

4.3 Preliminaries

Let $\mathcal{U}$ and $\mathcal{I}$ denote the sets of users and items, respectively. The historical interactions between users and items are represented by $\mathcal{P} = \{p_{ui} \mid \mathbf{u} \in \mathcal{U}, \mathbf{i} \in \mathcal{I}\}$, where $p_{ui} = 1$ indicates that user $\mathbf{u}$ has interacted with item $\mathbf{i}$. The objective of a RS is to learn a prediction function $f(\mathbf{u}, \mathbf{i})$ that estimates the likelihood of a future interaction between user $\mathbf{u}$ and item $\mathbf{i}$ based on these historical records.

Each item $\mathbf{i} \in \mathcal{I}$ is associated with additional entities that provide semantic infor-

mation beyond direct interactions. We denote the set of entities included in the original dataset as $\mathcal{E}' = \{e_1, e_2, \dots, e_{|\mathcal{E}'|}\}$, where each entity represents a semantic concept related to the item, such as brand, category, or descriptive attribute. Recent advancements in large language models (LLMs) enable the automatic generation of new entities and relations that extend the original KG, resulting in an LLM-augmented entity set $\mathcal{E} = \mathcal{E}' \cup \mathcal{E}_{LLM}$. These additional entities enrich the semantic space of items by providing contextual and fine-grained knowledge that may not exist in the original dataset.

To model the relationships among users, items, and entities, we adopt a graph-based formulation. The graph is defined as $\mathcal{G} = (\mathcal{V}, \mathcal{R})$, where $\mathcal{V} = \mathcal{U} \cup \mathcal{I} \cup \mathcal{E}$ is the set of nodes, and $\mathcal{R} = \mathcal{P} \cup \mathcal{Q}$ is the set of edges, with $\mathcal{P}$ representing user-item interactions and $\mathcal{Q} = \{q_{ie} \mid i \in \mathcal{I}, e \in \mathcal{E}\}$ representing item-entity relations. This heterogeneous graph structure allows information to propagate from users to items through both direct interactions and enriched semantic connections, thereby enabling the model to capture complex and interpretable relationships between user preferences and item-level knowledge.

Finally, the objective of the model is to learn low-dimensional embeddings for users, items, and entities that jointly encode collaborative patterns from historical interactions and semantic knowledge from the augmented entity space. The learned embeddings are used to predict unobserved user-item interactions and to provide personalized, explainable recommendations.

4.4 Proposed Method

4.4.1 Overview

In this section, we elaborate on our framework, which consists of six components: a survey-guided LLM data augmentation module, an item-entity KG structure, a user-item interaction graph structure, GNNs for modeling item-entity and user-item interactions, an entity-aware attention module, and a counterfactual reasoning mechanism. The overall framework is illustrated in Figure 4.1.

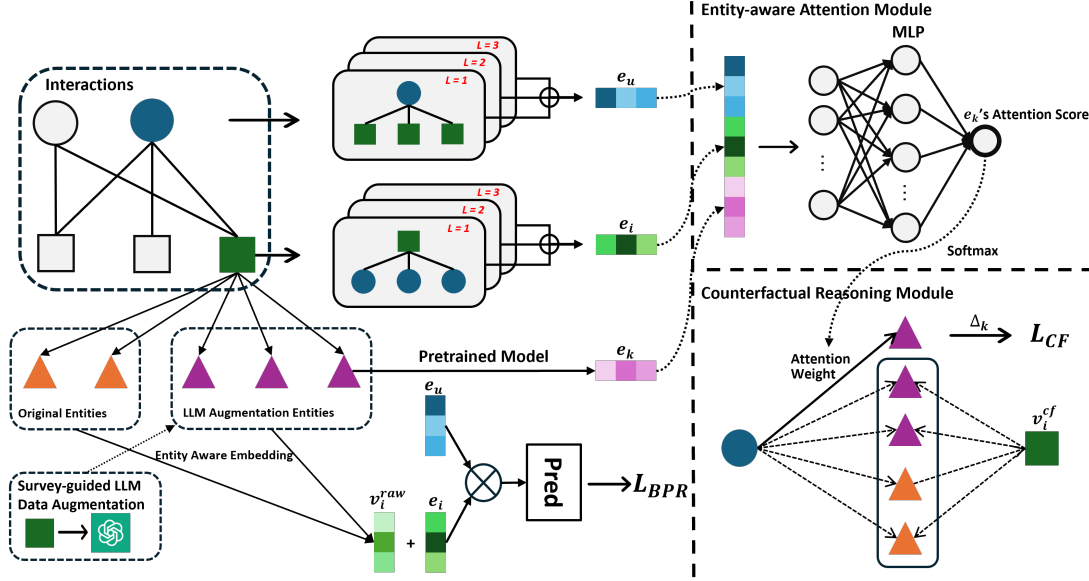


Figure 4.1: The overall framework of our proposed model. It consists of a survey-guided LLM data augmentation module, a KG structure of items and entities, a graph structure of user-item interactions, GNNs for each graph structure, an entity-aware attention module, and a counterfactual reasoning mechanism to generate the representations of users, items, and entities and provide both accurate recommendations and explanations.

4.4.2 Survey-Guided LLM Data Augmentation

Most public datasets, such as Amazon, typically attach only a few kinds of side information to each item. Although these entities are useful for understanding the user preferences, they are still too shallow to explore the reasons behind the interactions. To enrich the characters of items without manually selecting the features, we leverage recent advances in LLMs. Although LLMs contain rich knowledge of the item, directly using the item’s information to augment the related entities without limitations may also cause data redundancy and noisy data. To address this problem, we propose a survey-guided augmentation pipeline that automatically expands the KG with valuable, typed entities.

Concretely, we first instruct the LLM to act as a domain expert and create a concise survey that covers the factors that may affect the user preferences but are missing from the raw metadata. For example, in the Amazon Movies dataset, the generated survey

typically asks questions regarding the related entities (e.g., lead actors, supporting cast, soundtrack composer, etc.). Then, we feed each item’s existing basic information (e.g., title and description), along with the generated survey, into the LLM, instructing it to answer the survey questions. The responses provided by the LLM will become the new and enriched entities connected to the items. Finally, we combine the new entities with the original side information (e.g., image, brand, and category) to construct the KG. Specifically, we connect each item node to the generated entities based on relations defined by the survey questions. Figure 4.2 shows an example of the data augmentation generated by LLM.

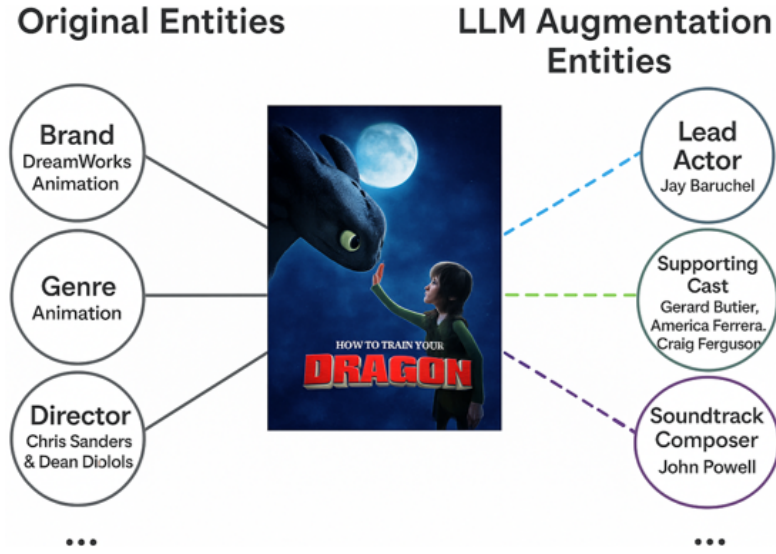


Figure 4.2: Example of the data augmentation for the item from the Amazon Movies and TV dataset by ChatGPT.

4.4.3 Item-entity KG Structure and Training

After generating the enriched entities by the LLM, we then construct the KG between items and entities based on the relations. This enriched graph structure provides a more detailed semantic representation of items, enabling better user preference modeling and more effective recommendation outcomes. Concretely, we represent the KG as a set of triples (h, r, t) , where h and t indicate the head and tail entities, respectively, and r de-

notes the relationship. Entities include both original side information (e.g., brand, category) and additional entities generated by our LLM-based survey augmentation method. Relations correspond to explicit attributes derived from the survey questions. Thus, the KG structure can significantly enrich the semantic information of items.

To effectively model the relations within this enriched KG, we adopt the TransR [102] model, which projects entities from the entity space into relation-specific spaces. Formally, for each triple $(\mathbf{h}, \mathbf{r}, \mathbf{t})$, TransR [102] learns entity embeddings $\mathbf{e}_h, \mathbf{e}_t \in \mathbb{R}^d$, relation embeddings $\mathbf{r} \in \mathbb{R}^{d_r}$, and relation-specific projection matrices $\mathbf{M}_r \in \mathbb{R}^{d_r \times d}$. Given pretrained embeddings $\mathbf{e}_{input}^{pre}$ from a general embedding model (e.g., GloVe [66] and ResNet [65] pretrained models), we initialize the entity embeddings $\mathbf{e}_{input}$ from the pretrained models with trainable transformation matrix $\mathbf{W}_{pretrain}$ rather than using random initialization:

$$\mathbf{e}_{input} = \mathbf{W}_{pretrain} \mathbf{e}_{input}^{pre}. \quad (4.1)$$

The entities are first projected into relation-specific spaces as follows:

$$\mathbf{e}_{h,r} = \mathbf{M}_r \mathbf{e}_h, \quad \mathbf{e}_{t,r} = \mathbf{M}_r \mathbf{e}_t. \quad (4.2)$$

The distance score of a triple is then computed as:

$$\mathbf{f}_r(\mathbf{h}, \mathbf{t}) = \|\mathbf{e}_{h,r} + \mathbf{r} - \mathbf{e}_{t,r}\|_2^2. \quad (4.3)$$

To train the TransR [102] model, we optimize the model with a margin-based ranking loss:

$$\mathcal{L}_{KG} = \sum_{(h,r,t) \in \mathcal{T}} \sum_{(h',r,t') \in \mathcal{T}'} \max(0, \gamma + \mathbf{f}_r(\mathbf{h}, \mathbf{t}) - \mathbf{f}_r(\mathbf{h}', \mathbf{t}')), \quad (4.4)$$

where $\mathcal{T}$ is the set of positive triples, $\mathcal{T}'$ is the set of negative triples, and γ denotes the margin hyperparameter. Training TransR [102] using pretrained entity embeddings helps accelerate the training process and improve representation quality by leveraging prior knowledge captured by pretrained models.

4.4.4 User-Item Interaction Graph Structure

To capture user preferences, we further construct a user-item interaction graph, where nodes represent users and items, and edges represent interactions.

Item embeddings $\mathbf{e}_i$ in this graph are computed by aggregating embeddings of their associated entities from the KG, thus generating the item representations. Specifically, the embedding of an item i is derived by aggregating the embeddings of its related entities:

$$\mathbf{e}_i = \sum_{e \in \mathcal{E}_i} \mathbf{e}_e, \quad (4.5)$$

where $\mathcal{E}_i$ denotes the set of entities associated with item i , and each $\mathbf{e}_e$ is the entity embeddings learned from the TransR [102] model.

User embeddings $\mathbf{e}_u \in \mathbb{R}^d$ are randomly initialized and share the same dimension as item embeddings $\mathbf{e}_i$. To further learn the embeddings of users $\mathbf{e}_u$ and items $\mathbf{e}_i$ from the interactions, we employ LightGCN [70], which is a simplified and efficient graph convolutional network for recommendation tasks. It aggregates the embeddings of neighboring nodes through simple averaging, defined by the following aggregation rule:

$$\begin{aligned} \mathbf{e}_u^{(k+1)} &= \sum_{i \in \mathcal{N}_u} \frac{1}{\sqrt{|\mathcal{N}_u| |\mathcal{N}_i|}} \mathbf{e}_i^{(k)}, \\ \mathbf{e}_i^{(k+1)} &= \sum_{u \in \mathcal{N}_i} \frac{1}{\sqrt{|\mathcal{N}_i| |\mathcal{N}_u|}} \mathbf{e}_u^{(k)}, \end{aligned} \quad (4.6)$$

where $\mathcal{N}_u$ and $\mathcal{N}_i$ represent the sets of items connected to user u and users connected to item i respectively.

After L layers of GNNs we can get the updated embeddings of $\mathbf{e}_u$ and updated embeddings of $\mathbf{e}_i$ as:

$$\mathbf{e}_u = \frac{1}{L+1} \sum_{k=0}^L \mathbf{e}_u^{(k)}, \quad \mathbf{e}_i = \frac{1}{L+1} \sum_{k=0}^L \mathbf{e}_i^{(k)}. \quad (4.7)$$

From the GNN model, user embeddings can capture collaborative signals from the interactions, which represent user preference patterns. Correspondingly, item embeddings can capture both information from entities and collaborative information from user interactions.

4.4.5 Counterfactual Attention Mechanism

Although the learned embeddings for users $\mathbf{e}_u$ and items $\mathbf{e}_i$ contain collaborative and semantic information effectively, they still lack user preferences analysis. Specifically,

current embeddings do not explicitly represent each user's preference towards particular entities associated with items. To deeper understand the reasons behind user preferences and improve the accuracy of the model predictions and explainability, we propose a counterfactual attention approach.

Concretely, for each user-item pair (u, i) and every entity k linked to the item, we feed the concatenation of three vectors into a lightweight multi-layer perceptron (MLP): the user embedding e_u , the item embedding e_i , and the entity embedding e_k to generate the attention score of e_k as:

$$\mathbf{att}_{uik} = \text{MLP}([e_u || e_i || e_k]), \quad (4.8)$$

where $[\cdot || \cdot]$ represents the vector concatenation and $\mathbf{att}_{uik}$ represents the entity k 's attention score. Since e_u contains the user's collaborative preferences learned from interactions, e_i provides the item's global collaborative information captured by GNN, and e_k contains the semantic information of entity k obtained from the KG. Analyzing these three embeddings together enables the MLP to assess how important entity k is for recommending item i to user u . In addition, the personalized attention scores not only improve recommendation accuracy but also offer explainability by indicating how each entity influences the user-item prediction.

After generating the attention scores for each entity, we then calculate the final attention weight of the k 's entity over all entities α_{uik} by the Softmax method as:

$$\alpha_{uik} = \frac{\exp(\mathbf{att}_{uik})}{\sum_{j \in \mathcal{E}_i} \exp(\mathbf{att}_{uij})}. \quad (4.9)$$

Finally, we obtain a user-specific item representation by summing all entities with the specific attention weight as:

$$\mathbf{v}_i^{raw}(u) = \sum_{k \in \mathcal{E}_i} \alpha_{uik} e_k. \quad (4.10)$$

Since the embedding $\mathbf{v}_i^{raw}$ only focuses on entities, it lacks the global collaborative information contained in e_i . As a result, if an item only has a few attributes or missing key entities, $\mathbf{v}_i^{raw}$ can become unstable and cannot reflect the user-item interaction patterns. To retain the global collaborative signal learned from GNN, we inject a skip-connection from the initial embedding of the item e_i and generate the final embeddings

as:

$$\mathbf{v}_i(u) = (1 - \beta) \mathbf{v}_i^{raw}(u) + \beta \mathbf{e}_i, \quad (4.11)$$

where $\mathbf{v}_i(u)$ is the final item embedding. Embedding $\mathbf{e}_i$ contains high-order user-item information, and the hyperparameter β determines how much of the global signal should be incorporated into the final embedding.

To make the learned attention weights truly reflect how much each entity contributes to a recommendation, we propose a counterfactual reasoning method. Concretely, for each training triple $(\mathbf{u}, \mathbf{i}, \mathbf{k})$, we remove entity $\mathbf{k}$ from the item, and keep all other entities unchanged as:

$$\mathbf{v}_i^{cf}(u) = \sum_{j \neq k} \frac{\alpha_{uij}}{\sum_{l \neq k} \alpha_{uil}} \mathbf{e}_j. \quad (4.12)$$

Then we recompute the predicted score and calculate the difference between the full score and the counterfactual score, which indicates the real influence of entity $\mathbf{k}$ as:

$$\Delta_k = \mathbf{e}_u^\top \mathbf{v}_i(u) - \mathbf{e}_u^\top \mathbf{v}_i^{cf}(u). \quad (4.13)$$

A large Δ_k means that entity $\mathbf{k}$ has a high impact on user $\mathbf{u}$'s preferences; otherwise, a small Δ_k means that entity $\mathbf{k}$ is irrelevant for this particular recommendation.

We then calculate the counterfactual loss $\mathcal{L}_{CF}$ as:

$$\mathcal{L}_{CF} = \left(\alpha_{uik} - \frac{\Delta_k}{\sum_{j \in \mathcal{E}_i} \Delta_j} \right)^2, \quad (4.14)$$

where $\frac{\Delta_k}{\sum_{j \in \mathcal{E}_i} \Delta_j}$ represents the impact caused by removing entity $\mathbf{k}$. Minimizing the squared error can let the attention weight α_{uik} track the entity's actual influence.

4.4.6 BPR Loss

After we generate the final embeddings of users and items, we then optimize the model parameters using Bayesian Personalized Ranking (BPR) loss [72], which encourages the model to rank positive interactions higher than negative ones. Formally, the BPR loss is defined as:

$$\mathcal{L}_{BPR} = - \sum_{(u, i, i') \in \mathcal{S}} \ln \sigma(\mathbf{e}_u^\top \mathbf{v}_i(u) - \mathbf{e}_u^\top \mathbf{v}_{i'}(u)), \quad (4.15)$$

where $(\mathbf{u}, \mathbf{i}, \mathbf{i}')$ represents a user $\mathbf{u}$, a positive item $\mathbf{i}$, and a negative item $\mathbf{i}'$. $\mathcal{S}$ represents a set of positive and negative interactions:

$$\mathcal{S} = \{(\mathbf{u}, \mathbf{i}, \mathbf{i}') | (\mathbf{u}, \mathbf{i}) \in \mathcal{P}, (\mathbf{u}, \mathbf{i}') \notin \mathcal{P}\}. \quad (4.16)$$

4.4.7 Combined Model Training

Finally, we calculate the final loss $\mathcal{L}$ which combines ranking accuracy and explainability to train our model as:

$$\mathcal{L} = \mathcal{L}_{BPR} + \lambda \mathcal{L}_{CF}, \quad (4.17)$$

where the hyperparameter λ can balance the two goals of our framework. This joint objective allows the model to maintain high recommendation accuracy and assign weights that can analyze and explain how each entity affects the preference of the user.

4.5 Experiment

In this section, we conduct experiments on our model and state-of-the-art methods. We also conduct studies on our proposed model and answer the following questions:

- **RQ1:** How does our model perform compared with the state-of-the-art methods on the real-world datasets?
- **RQ2:** How do different components (data augmentation, entity-aware attention mechanism, skip-connection from the global collaborative signal) of our model affect the performance?
- **RQ3:** How do different designs (different LLM models for data augmentation) of our model affect the performance?

4.5.1 Datasets

Amazon Movies and TV. This dataset is publicly available and commonly used for recommendation tasks. This dataset contains reviews of the products that were sold on the Amazon Movies and TV category. Besides interaction records between users and

items, this dataset also contains some basic types of side information such as image URLs, descriptions, brands, and categories. We use the ResNet50 model [65] and the GloVe model [66] to extract the visual and textual features.

MovieLens. This public dataset is also commonly used for recommendation tasks. This dataset contains ratings of the movies from the MovieLens website. Besides interaction records between users and movies, this dataset also contains some basic types of side information, such as titles and genres. We use the GloVe model [66] to extract the textual features.

4.5.2 Baselines

To demonstrate the performance, we compared our proposed method with the following methods:

VBPR [73]. A benchmark framework in multimedia recommendation. It combines visual side information into a CF framework. We follow the original setting and concatenate all multi-modal features with the item’s latent vector when making predictions.

XRec [97]. An LLM-enhanced GNN framework, which combines the collaborative signals and LLM to generate recommendations and explanations from the interaction records.

MMGCN [48]. A GCN-based multi-modal framework. For each modality, it builds a user-item bipartite graph, propagates features with GCN, and then fuses the modality-specific representations through combination layers to make the predictions.

KGAT [80]. A KG-based framework, which unifies users, items, and item-related entities into a heterogeneous KG. And then applies relation-aware attention to pass messages along high-order paths, enabling richer semantics to refine user and item representations.

MKGAT [56]. A KG-based framework, which combines visual and textual features of entities into the heterogeneous graph of users, items, and item-related entities. It consists of a cross-modal, relation-aware attention mechanism that jointly propagates structural, visual, and textual signals, further enhancing accuracy and explainability.

LKGCN [101]. A lightweight entity-enhanced KG convolutional network. It assigns relation-specific attention when aggregating neighbors and adds an auxiliary user-

entity preference task, thereby capturing high-order user-item-entity connections to improve both accuracy and explainability.

4.5.3 Evaluation Metrics

We use the commonly used evaluation methods in recommendation works: Recall@K and Normalized Discounted Cumulative Gain (NDCG@K) to evaluate the performance of our proposed model, where @K denotes the top-K candidate items of each user. Specifically, we set $K = 5, 10$, and calculated the average scores of the above metrics for all users during the testing phase.

4.5.4 Parameter Settings

To train our proposed model, we randomly initialize the parameters and optimize the model with Adam [74]. We searched the learning rate in $\{0.01, 0.001, 0.0001\}$ and the regularization weight in $\{0.001, 0.0001, 0.00001\}$. We show the results of the learning rate 0.001, and regularization weight 0.0001, and we unified the dimension to 64 for all models.

4.5.5 Performances Comparison (RQ1)

The comparative results are summarized in Table 4.1. Note that “Improv.” denotes the improvement of LACAT’s result compared with the best results of the baseline models. From the table, we have the following observations:

All GNN-based models outperformed the CF-based model VBPR in all metrics. It demonstrates that graph convolutional networks can better learn the characters from the interactions because most of the interactions can be structured into graphs. KG-based models outperformed the GCN-based model MMGCN, because KG can unify the users, items, and entities into a heterogeneous graph, which contains edges of both interactions and relationships between entities, thus enriching the representations.

LACAT outperforms the other GNN-based baseline models in all metrics, which can verify the effectiveness of our model. The reasons are: firstly, we apply a survey-guided LLM data augmentation, which can enrich the related entities under guidance. Although

Table 4.1: Performance comparison between our model and baseline models over two datasets.

Amazon Movies and TV				
Model	Recall@5	Recall@10	NDCG@5	NDCG@10
VBPR	0.1029	0.2017	0.0971	0.1123
XRec	0.2276	0.3272	0.1991	0.2101
MMGCN	0.2225	0.3220	0.1989	0.2093
KGAT	0.2352	0.3315	0.2126	0.2281
MKGAT	0.2415	0.3370	0.2192	0.2304
LKGCN	0.2508	0.3414	0.2224	0.2350
LACAT	0.2884	0.3959	0.2576	0.2711
Improv.	14.99%	15.97%	15.83%	15.37%
MovieLens				
Model	Recall@5	Recall@10	NDCG@5	NDCG@10
VBPR	0.0721	0.1325	0.0618	0.0802
XRec	0.1426	0.2502	0.1403	0.1611
MMGCN	0.1419	0.2488	0.1382	0.1599
KGAT	0.1620	0.2597	0.1498	0.1695
MKGAT	0.1673	0.2701	0.1522	0.1712
LKGCN	0.1704	0.2732	0.1580	0.1765
LACAT	0.1992	0.3217	0.1802	0.2019
Improv.	16.90%	17.75%	14.05%	14.39%

XRec also takes advantage of LLMs, it may also generate redundant or noisy information without guidance. Compared with other baseline models, our model contains more valuable entities related to user preferences on specific domains. Secondly, our model contains a specific counterfactual-based entity-aware attention module. Instead of only learning the attention weights of entities from the graph topology, our model provides a more reasonable attention mechanism that can truly reflect each entity’s importance, thus improving both accuracy and explainability. Lastly, our model contains a specific

combination layer that leverages the entity-aware embedding and the global collaborative information with the skip-connection to make the predictions. This enables our model to take advantages from both interaction and entity aspects.

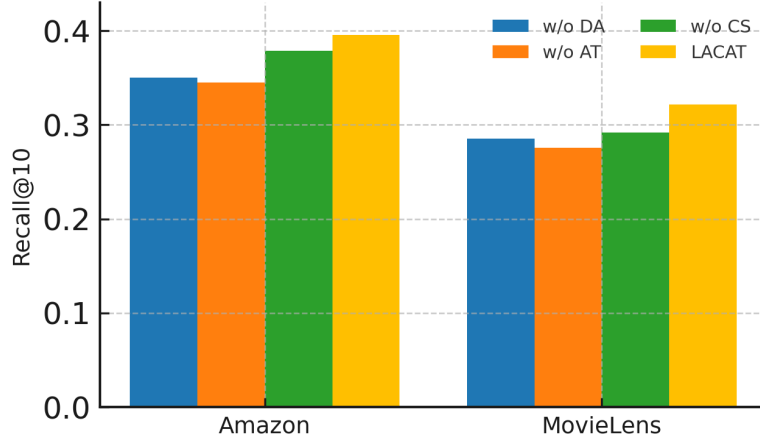


Figure 4.3: Ablation study on Recall@10 for Amazon and MovieLens.

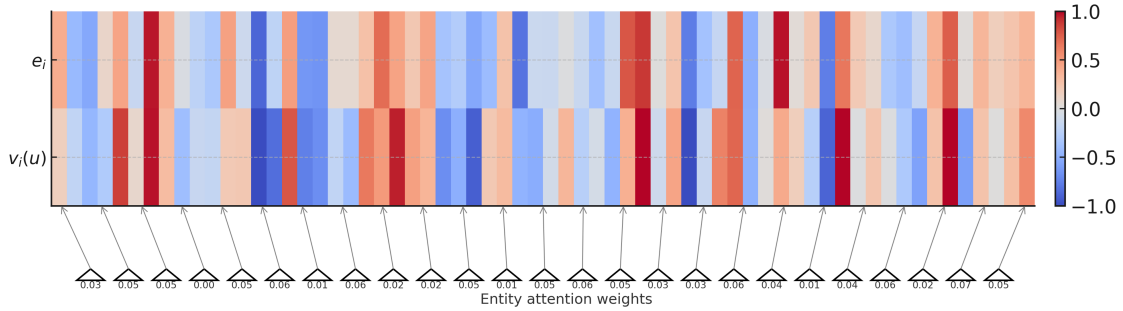


Figure 4.4: Visualization of the initial item embedding, the user-specific item embedding, and the attention weights over entities.

4.5.6 Ablation Study (RQ2)

To demonstrate the effectiveness of each components in our model, we conduct an ablation study by removing individual components, including the data augmentation module (w/o DA), the counterfactual-based entity-aware attention mechanism (w/o AT), and skip-connection from the global collaborative signal e_i (w/o CS). The results are shown

in Figure 4.3. It is evident that each individual component adds a positive contribution to the model performance.

Furthermore, Figure 4.4 shows a visualization of item initial embedding $\mathbf{e}_i$, the user-specific item embedding $\mathbf{v}_i(u)$, and the attention weight of each entity, clearly showing how these weights improve accuracy and make the recommendations explainable.

4.5.7 Effect of Different LLM models (RQ3)

In our main experiments, we use GPT-4o for the LLM data augmentation. To evaluate the effect of different LLMs, we compare the results with GPT-3.5-turbo, GPT-4, and GPT-4-turbo. The results are shown in Figure 4.5. We find that GPT-4o achieves the best performance, indicating that stronger LLMs can provide better entity augmentation, which leads to better recommendation results.

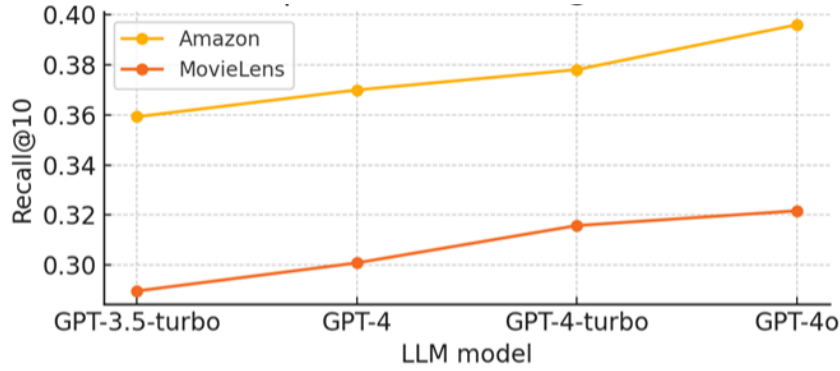


Figure 4.5: Performances comparison on Recall@10 of different LLM models.

4.6 Conclusion

In this chapter, we proposed a novel framework named LLM Augmentation Enhanced Counterfactual Attention for Graph-based RSs (LACAT). LACAT was designed to address the challenges of semantic enrichment and personalized causal reasoning in KG-enhanced recommendation. Unlike traditional graph-based recommenders that rely on a limited number of predefined entities, LACAT integrates a survey-guided large language model (LLM) to expand the entity space with richer and fine-grained semantic

information. This augmentation enables the model to represent complex contextual factors—such as user interests, thematic nuances, and latent attributes—that are often missing in conventional datasets.

Based on the augmented entity set, we constructed a heterogeneous user-item-entity graph and applied GNNs to jointly learn representations of users, items, and entities. To further improve explainability and personalization, we introduced a counterfactual attention mechanism that quantifies the causal influence of each entity on user preferences. Through this personalized reasoning process, the model not only captures which entities are most relevant to individual users but also explains why specific entities contribute to the recommendation results. In this way, LACAT achieves both high accuracy and strong explainability.

We evaluated our framework on two publicly accessible datasets—Amazon and MovieLens, and experimental results demonstrated that LACAT significantly outperforms several state-of-the-art baselines in Recall@K and NDCG@K. These improvements verify the effectiveness of combining LLM-based semantic augmentation with counterfactual reasoning for enhancing both the precision and explainability of recommendations.

Although LACAT effectively bridges the gap between behavioral and semantic spaces, there remain several limitations to explore in future work. First, the diversity of generated entities depends on the quality and domain coverage of the LLM, which may introduce redundant or noisy attributes. Second, while the counterfactual attention module enhances explainability, its computational cost may increase when dealing with very large graphs. In future research, we plan to improve the efficiency of entity selection and develop adaptive filtering strategies to balance semantic richness with model scalability. Furthermore, we intend to extend the framework to multi-domain settings, where causal reasoning over LLM-augmented knowledge could further improve the robustness and generalization of RSs.

Chapter 5

Summary

5.1 Summary of Contributions

In this thesis, we have investigated how to enhance the accuracy and explainability of recommender systems (RSs) in realistic scenarios where user-item interactions are sparse and side information is heterogeneous. We focused on improving the representation quality of users and items by incorporating multi-modal data, relational structures, and large language model (LLM)-driven knowledge augmentation. To achieve this goal, we systematically addressed three major challenges that hinder existing recommendation frameworks: (1) modeling the relationships between interactions and side information in multi-modal environments, (2) aligning semantic and behavioral spaces for modality-aware learning, and (3) reasoning about personalized causal influence through enriched semantic entities.

In Chapter 2, we addressed the first challenge by exploring the deeper relationships between user-item interactions and item side information in multi-modal recommendation. Existing models primarily relied on independent feature extraction for each modality, followed by simple feature concatenation or averaging. These methods failed to capture how different types of side information jointly influence user preferences. To overcome this limitation, we proposed a Multi-modal Relational Side Information Graph Attention Network (MRSI-GAT). MRSI-GAT constructs a specific graph for each modality that includes both user-item interactions and item-item similarity rela-

tions based on side information. Through graph attention mechanisms, our model adaptively captures the importance of multi-hop neighbors and integrates relational signals across modalities. Experiments on public datasets demonstrated that MRSI-GAT effectively enhances representation learning and improves recommendation performance under sparse data conditions.

In Chapter 3, we addressed the second challenge by aligning behavioral representations learned from interactions with semantic features extracted from multiple modalities. Although existing multi-modal recommender systems can combine side information, they often ignore the semantic consistency between learned embeddings and raw side features, leading to shallow feature fusion. To handle this issue, we proposed a Multi-task Multi-modal Graph Neural Network (MTMM-GNN). MTMM-GNN constructs modality-specific user-item graphs to model interaction patterns within each modality and introduces an auxiliary task that aligns the similarity of learned embeddings with that of original side features. Through multi-task optimization, the model jointly learns modality-specific interaction patterns and semantic relationships. Experimental results on Amazon, ModCloth, and MovieLens datasets confirmed that MTMM-GNN achieves superior performance in terms of Recall@K, NDCG@K, and Precision@K, demonstrating its ability to bridge the gap between semantic and behavioral spaces.

In Chapter 4, we addressed the third challenge by incorporating enriched semantic knowledge generated by large language models into a graph-based recommendation framework. Traditional knowledge graph (KG)-enhanced recommenders are limited by the small number of predefined entities and relations, which restricts their ability to represent complex user preferences. To overcome this limitation, we proposed a novel framework, LLM Augmentation Enhanced Counterfactual Attention for Graph-based Recommender Systems (LACAT). LACAT enriches the entity space by augmenting the original KG with LLM-generated entities and relations, enabling a richer representation of item semantics. Furthermore, a counterfactual attention mechanism is introduced to quantify the causal influence of each entity on user preferences, achieving personalized and interpretable recommendations. Extensive experiments on Amazon and MovieLens datasets validated that LACAT not only improves accuracy but also enhances explainability by revealing entity-level reasoning behind recommendation results.

Through these studies, this thesis contributes a unified perspective on enhancing recommendation systems with multi-modal, graph-based, and LLM-augmented information. By combining representation learning, relational reasoning, and causal interpretability, our work provides practical methodologies for building next-generation recommender systems that are accurate, explainable, and adaptable to data-sparse environments. Although the three proposed methods are evaluated on similar datasets, they are designed to address different challenges in recommender systems. As a result, there is no single method that is universally optimal across all scenarios. Instead, each model demonstrates its strengths under specific conditions, such as relational sparsity, semantic-behavioral inconsistency, or the need for interpretability. Understanding these contextual differences is essential for selecting or integrating appropriate techniques in real-world applications.

5.2 Limitations

Although the three proposed models in this thesis effectively address key challenges in multi-modal, graph-based, and LLM-augmented recommendation, several limitations remain due to the inherent complexity of real-world data and the constraints of current methodologies.

First, the model proposed in Chapter 2 (MRSI-GAT) relies on manually defined similarity thresholds to construct item-item relations within each modality. While this enables the model to capture modality-specific relational structures, manually selecting thresholds may introduce bias and limit generalizability across datasets or domains. Additionally, the number of modalities available in public datasets is typically small, and many real-world user decision factors such as temporal context, emotional attributes, or social cues are not explicitly encoded, restricting the richness of side information that the model can exploit.

Second, the framework in Chapter 3 (MTMM-GNN) focuses on aligning the similarity patterns between learned embeddings and raw side features. However, this alignment is based solely on similarity distributions, which assumes that semantic relationships can be fully captured by pairwise similarity measures. In reality, user preferences may exhibit non-linear dependencies or contextual interactions that similarity-based losses

cannot fully model. Furthermore, although MTMM-GNN reduces inconsistency across modalities, it still depends on the availability and quality of the original side information; when important modalities are missing, the alignment objective may become less informative.

Third, the model introduced in Chapter 4 (LACAT) enriches semantic representations using LLM-generated entities, but the augmentation process may introduce noisy or redundant information. Although counterfactual attention helps handle this issue by assigning personalized causal importance to each entity, the accuracy of causal reasoning still depends on the coherence of the augmented knowledge graph. Moreover, current evaluation settings rely on publicly available datasets with limited entity diversity, which may not fully reflect the complexity of deploying LACAT in large-scale, real-world environments.

Overall, these limitations highlight the challenges of modeling complex semantic relationships, scaling multi-modal graphs, and ensuring the reliability of augmented knowledge. Addressing these issues provides promising opportunities for future research.

Furthermore, an important practical limitation shared across all three proposed methods is that they are developed as independent solutions targeting different aspects of the recommendation pipeline. Although each model solves a specific challenge: multi-modal relational learning, semantic, behavior alignment, and explainable semantic enrichment. How these techniques can be integrated into a unified framework remains unexplored. A deeper investigation on combining the strengths of these approaches is essential for building an end-to-end system that simultaneously leverages multi-modal signals, aligns semantic and behavioral spaces, and provides personalized, causally grounded explanations. This integration challenge motivates part of the future research agenda discussed in the next section.

5.3 Future Work

While this thesis has proposed effective methods for multi-modal and LLM-augmented recommender systems, several research directions remain open for future exploration.

5.3.1 Enhancing the Robustness of LLM-augmented Knowledge

Although LLM-based data augmentation expands the semantic coverage of items, the generated entities may include noisy or redundant information that could reduce model robustness. Future work will focus on developing automatic validation and filtering mechanisms to evaluate the quality and reliability of LLM-generated knowledge before integration into recommendation graphs. Additionally, exploring dynamic update strategies that adapt entity augmentation according to user feedback can further improve consistency between semantic representations and real-world user preferences.

5.3.2 Scalable and Efficient Graph Learning

As the number of modalities and entities increases, the graph structure in multi-modal and knowledge-augmented recommendations becomes increasingly large and complex. Efficient training and inference on such heterogeneous graphs remain challenging. Future work will investigate scalable graph learning techniques—such as sampling-based aggregation, hierarchical message passing, or lightweight attention mechanisms—to reduce computational cost while maintaining recommendation accuracy.

5.3.3 Extending to Cross-domain Recommendations

Another promising direction is to extend our proposed methods to multi-domain scenarios. In many real-world applications, user behavior and item attributes across multiple domains (e.g., books, movies, and e-commerce) and languages. Building cross-domain knowledge graphs with LLM-augmented entities and integrating multiple semantic representations can help further improve generalization and adaptability of recommender systems.

5.3.4 Integrated Framework Combining All Proposed Techniques

While each model presented in this thesis addresses a different challenge in multi-modal and knowledge-augmented recommendation, an important future direction is to explore how these techniques can be integrated into a unified framework. MRSI-GAT focuses

on modeling relational structures within modalities, MTMM-GNN aligns semantic features with user behaviors, and LACAT enhances item semantics and interpretability through LLM-generated entities and personalized reasoning. Combining these capabilities could enable a next-generation recommender system that simultaneously exploits multi-modal relations, maintains semantic-behavioral consistency, and reasons about personalized causal factors behind user decisions.

A possible direction is to design a shared representation space in which relational graphs, modality-specific semantics, and enriched entities are jointly encoded. Such a system could incorporate relational graph structures, apply semantic alignment objectives to ensure consistency across modalities, and integrate the augmented entities and personalized attention for interpretable reasoning. Another direction is to construct a hybrid multi-task framework in which each component contributes complementary training signals: relational, semantic, and causal, thereby improving robustness under sparse interactions. Exploring these integration strategies may lead to more powerful, adaptable, and explainable recommendation models.

In summary, the contributions presented in this thesis provide a comprehensive foundation for advancing personalized, interpretable, and semantically rich recommender systems. We expect that continued research along these directions will further strengthen the connection between human understanding and machine intelligence in the field of recommendation.

Acknowledgment

This thesis encapsulates my academic endeavors at the University of Osaka and a significant chapter of my life spent studying and residing in Japan over these years. The successful completion of this thesis is attributed to the selfless support and hard work of a number of individuals.

First and foremost, I extend my deepest gratitude to my advisor, Professor Takahiro Hara, who has provided me with countless opportunities since the beginning of my journey in this laboratory. He has been a profound motivator for me to start the Ph.D. course and grow as a researcher. It has been an honor to be his Ph.D. student. His guidance and mentorship have supported me throughout my doctoral journey. His insightful feedback and encouragement have been invaluable in shaping my research.

I gratefully acknowledge the help of my associate professor, Associate Professor Yihong Zhang, who has meticulously guided every detail in each step of my research and all the drafts of my papers. Without his kind guidance and valuable advice, it would have been impossible for me to finish my research and this thesis. Also, I am grateful for the precious opportunities for tutorial presentations at international conferences provided by Professor Yihong Zhang.

I would like to express my appreciation to Mr. Mori Kurokawa, Mr. Kei Yonekawa, Dr. Duc Nguyen, Dr. Shuichiro Haruta, Dr. Hao Niu, and Dr. Chihiro Ono from KDDI Research Inc. They always gave me helpful guidance and technological support throughout my research.

In addition, I would like to acknowledge the Interdisciplinary Innovation Human Resource Development Fellowship for its generous academic and financial support throughout my doctoral studies. The fellowship not only enabled me to focus on my research without financial concerns but also provided valuable opportunities for academic train-

ing, interdisciplinary collaboration, and presentation experience. The guidance, feedback, and encouragement from the fellowship committee have been essential to the progress of my research and my growth as a researcher.

I would like to acknowledge my research team, both past and present members, Dr. Qingxin Xia, Mr. Bunki Cao, Mr. Yuan Lyu, Mr. Hanxin Wang, Mr. Yuya Takahashi, Mr. Takuma Kaiho, Ms. Yue Zhao, Ms. Hikaru Nomura, Mr. Jun Murao, Mr. Zhi Li, Mr. Kentaro Shiga, Mr. Rikuto Tsubouchi, Mr. Takuma Yamashita, Mr. Yuma Dose, Mr. Aoran Chen, Mr. Yoshiyuki Maekawa, Mr. Keita Matsunaka, Mr. Ryusei Iwasa, Mr. Jo Masuda, and Mr. Yoshiyuki Sone. Also, it is my pleasure to work with all the smart members of Hara laboratory. All of them have provided me with cooperative and active support throughout my study.

Finally, this thesis would not have taken place without the support of my family and friends who gave me unconditional love and motivation. I give my thanks to all involved.

REFERENCE

- [1] Dietmar Jannach and Markus Zanker. Value and impact of recommender systems. In *Recommender Systems Handbook*, pages 519–546. 2022.
- [2] J. Ben Schafer, Joseph Konstan, and John Riedl. Recommender systems in e-commerce. In *Proceedings of the ACM Conference on Electronic Commerce (EC)*, pages 158–166, 1999.
- [3] Yan Guo, Chengxin Yin, Mingfu Li, Xiaoting Ren, and Ping Liu. Mobile e-commerce recommendation system based on multi-source information fusion for sustainable e-business. *Sustainability*, 10(1):147, 2018.
- [4] Farah Tawfiq Abdul Hussien, Abdul Monem S. Rahma, and Hala Bahjat Abdul Wahab. Recommendation systems for e-commerce systems: an overview. In *Journal of Physics: Conference Series*, volume 1897, page 012024, 2021.
- [5] Jie Lu, Dianshuang Wu, Mingsong Mao, Wei Wang, and Guangquan Zhang. Recommender system application developments: a survey. *Decision Support Systems*, 74:12–32, 2015.
- [6] Ruihua Shao. Improvement of business analysis method of e-commerce system from the perspective of intelligent recommendation systems. *Advances in Multimedia*, 2022:7860718, 2022.
- [7] Jatin Sharma, Kartikay Sharma, Kaustubh Garg, and Avinash Kumar Sharma. Product recommendation systems: a comprehensive review. In *IOP Conference Series: Materials Science and Engineering*, volume 1022, page 012021, 2021.

- [8] Dokyun Lee and Kartik Hosanagar. Impact of recommender systems on sales volume and diversity. In *Proceedings of the International Conference on Information Systems (ICIS)*, 2014.
- [9] Renjie Zhou, Samamon Khemmarat, and Lixin Gao. The impact of the YouTube recommendation system on video views. In *Proceedings of the ACM SIGCOMM Conference*, pages 404–410, 2010.
- [10] Abhinandan Das, Mayur Datar, Ashutosh Garg, and Shyamsundar Rajaram. Google news personalization: scalable online collaborative filtering. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 271–280, 2007.
- [11] Dietmar Jannach and Michael Jugovac. Measuring the business value of recommender systems. *ACM Transactions on Management Information Systems*, 10(4):16:1–16:23, 2019.
- [12] M. Benjamin Dias, Dominique Locher, Ming Li, Wael El-Deredy, and Paulo J. G. Lisboa. The value of personalised recommender systems to e-business: a case study. In *Proceedings of the ACM Conference on Recommender Systems (RecSys)*, pages 291–294, 2008.
- [13] Richard D. Lawrence, George S. Almási, Vladimir Kotlyar, Marisa S. Viveros, and Sastry Duri. Personalization of supermarket product recommendations. *Data Mining and Knowledge Discovery*, 5(1–2):11–32, 2001.
- [14] Ye Chen and John F. Canny. Recommending ephemeral items at web scale. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1013–1022, 2011.
- [15] M. Benjamin Dias, Dominique Locher, Ming Li, Wael El-Deredy, and Paulo J. G. Lisboa. The value of personalised recommender systems to e-business: a case study. In *Proceedings of the ACM Conference on Recommender Systems (RecSys)*, pages 291–294, 2008.

- [16] Zeshan Fayyaz, Mahsa Ebrahimian, Dina Nawara, Ahmed Ibrahim, and Rasha Kashef. Recommendation systems: algorithms, challenges, metrics, and business opportunities. *Applied Sciences*, 10(21):7748, 2020.
- [17] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: a survey and new perspectives. *ACM Computing Surveys*, 52(1), 2019.
- [18] Vineeta Anand and Ashish Kumar Maurya. A survey on recommender systems using graph neural networks. *ACM Transactions on Information Systems*, 43(1):1–49, 2025.
- [19] Trong Dinh Thac Do and Longbing Cao. Metadata-dependent infinite Poisson factorization for efficiently modelling sparse and large matrices in recommendation. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5010–5016, 2018.
- [20] Hanxin Wang, Daichi Amagata, Takuya Maekawa, Takahiro Hara, Hao Niu, Kei Yonekawa, and Mori Kurokawa. A DNN-based cross-domain recommender system for alleviating the cold-start problem in e-commerce. *IEEE Open Journal of Industrial Electronics*, 1:194–206, 2020.
- [21] Hanxin Wang, Daichi Amagata, Takuya Maekawa, Takahiro Hara, Hao Niu, Kei Yonekawa, and Mori Kurokawa. Preliminary investigation of alleviating the user cold-start problem in e-commerce with a deep cross-domain recommender system. In *Proceedings of the IEEE International Conference on Emerging Networking and Communication Technologies (ECNCT)*, pages 398–403, 2019.
- [22] Adit Krishnan, Ashish Sharma, Aravind Sankar, and Hari Sundaram. An adversarial approach to improve long-tail performance in neural collaborative filtering. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1491–1494, 2018.
- [23] Zhihong Chen, Rong Xiao, Chenliang Li, Gangfeng Ye, Haochuan Sun, and Hongbo Deng. ESAM: Discriminative domain adaptation with non-displayed

- items to improve long-tail performance. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 579–588, 2020.
- [24] Daniele Malitesta, Giandomenico Cornacchia, Claudio Pomo, Felice Antonio Merra, Tommaso Di Noia, and Eugenio Di Sciascio. Formalizing multimedia recommendation through multimodal deep learning. *ACM Transactions on Recommender Systems*, 3(3):1–33, 2025.
- [25] Ioannis Arapakis, Yashar Moshfeghi, Hideo Joho, Reede Ren, David Hannah, and Joemon M. Jose. Enriching user profiling with affective features for the improvement of a multimodal recommender system. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pages 1–8, 2009.
- [26] Sorabh Gupta, Amit Kumar Bindal, and Devendra Prasad. Multi-modality collaborative recommender systems: an overview of techniques and evaluation metrics. In *Proceedings of the International Conference on Advanced Computing and Communication Technologies (ICACCTech)*, pages 426–433, 2023.
- [27] Jieming Zhu, Xin Zhou, Chuhan Wu, Rui Zhang, and Zhenhua Dong. Multi-modal pretraining and generation for recommendation: a tutorial. In *Companion Proceedings of the ACM Web Conference (WWW Companion)*, pages 1272–1275, 2024.
- [28] Dhanesh Ramachandram and Graham W. Taylor. Deep multimodal learning: a survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108, 2017.
- [29] Peng Xu, Xiatian Zhu, and David A. Clifton. Multimodal learning with transformers: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132, 2023.
- [30] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: a survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018.

- [31] Arthur F. da Costa and Marcelo G. Manzato. Exploiting multimodal interactions in recommender systems with ensemble algorithms. *Information Systems*, 56:120–132, 2016.
- [32] Hong Fang, Jindong Liang, and Leiyuxin Sha. Enhanced multimodal recommendation systems through reviews integration. *Knowledge and Information Systems*, 67(4):3459–3486, 2025.
- [33] Hongyu Zhou, Xin Zhou, Zhiwei Zeng, Lingzi Zhang, and Zhiqi Shen. A comprehensive survey on multimodal recommender systems: taxonomy, evaluation, and future directions. *arXiv preprint arXiv:2302.04473*, 2023.
- [34] Guipeng Xv, Xinyu Li, Ruobing Xie, Chen Lin, Chong Liu, Feng Xia, Zhanhui Kang, and Leyu Lin. Improving multi-modal recommender systems by denoising and aligning multi-modal content and user feedback. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 3645–3656, 2024.
- [35] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 689–696, 2011.
- [36] Zhou Lu. A theory of multimodal learning. *Advances in Neural Information Processing Systems*, 36:57244–57255, 2023.
- [37] Yongheng Mu and Yun Wu. Multimodal movie recommendation system using deep learning. *Mathematics*, 11(4):895, 2023.
- [38] Haimonti Dutta and Anushka Tiwari. Content-based art recommendation using multimodal graph neural networks. In *Proceedings of the IEEE International Conference on Knowledge Graph (ICKG)*, pages 67–74, 2024.
- [39] Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, 38(8):2939–2970, 2022.

- [40] Zhenhua Huang, Xin Xu, Juan Ni, Honghao Zhu, and Cheng Wang. Multimodal representation learning for recommendation in internet of things. *IEEE Internet of Things Journal*, 6(6):10675–10685, 2019.
- [41] Amar Khelloufi, Huansheng Ning, Abdenacer Naouri, Abdelkarim Ben Sada, Attia Qammar, Abdelkader Khalil, Lingfeng Mao, and Sahraoui Dhelim. A multimodal latent-features-based service recommendation system for the social internet of things. *IEEE Transactions on Computational Social Systems*, 11(4):5388–5403, 2024.
- [42] Mahesh Goyani and Neha Chaurasiya. A review of movie recommendation systems: limitations, survey and challenges. *Electronic Letters on Computer Vision and Image Analysis*, 19(3):18–37, 2020.
- [43] S. Rajarajeswari, Sharat Naik, Shagun Srikant, M. K. Sai Prakash, and Prarthana Uday. Movie recommendation system. In *Emerging Research in Computing, Information, Communication and Applications (ERCICA 2018), Volume 1*, pages 329–340. 2019.
- [44] Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Wei Wang, Xiping Hu, Steven Hoi, and Edith Ngai. A survey on multimodal recommender systems: recent advances and future directions. *arXiv preprint arXiv:2502.15711*, 2025.
- [45] Shuo Wang, Jing Yang, and Fanshu Shang. A personalized recommendation model with multimodal preference-based graph attention network. *The Journal of Supercomputing*, 80(15):22020–22048, 2024.
- [46] Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. Multi-modal self-supervised learning for recommendation. In *Proceedings of the ACM Web Conference (WWW)*, pages 790–800, 2023.
- [47] Jianing Li, Chaoqun Yang, Guanhua Ye, and Quoc Viet Hung Nguyen. Graph neural networks with deep mutual learning for designing multi-modal recommendation systems. *Information Sciences*, 654:119815, 2024.

- [48] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pages 1437–1445, 2019.
- [49] Shiwen Wu, Fei Sun, Wentao Zhang, Xu Xie, and Bin Cui. Graph neural networks in recommender systems: a survey. *ACM Computing Surveys*, 55(5), 2022.
- [50] Hao Yan, Chaozhuo Li, Jun Yin, Zhigang Yu, Weihao Han, Mingzheng Li, Zhengxin Zeng, Hao Sun, and Senzhang Wang. When graph meets multimodal: benchmarking and meditating on multimodal attributed graph learning. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 5842–5853, 2025.
- [51] Zhaoliang Wang, Baisong Liu, Weiming Huang, Tingting Hao, Huiqian Zhou, and Yuxin Guo. Leveraging multimodal large language models for multimodal sequential recommendation. *Scientific Reports*, 15(1):28960, 2025.
- [52] Shengzhe Jiao, Yihong Zhang, and Takahiro Hara. Multi-modal relational side information graph attention networks for recommender system. In *Proceedings of the IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1012–1017, 2023.
- [53] Shengzhe Jiao, Yihong Zhang, and Takahiro Hara. Multi-task multi-modal graph neural network for recommender systems. *Knowledge and Information Systems*, pages 1–26, 2025.
- [54] Qidong Liu, Jiayi Hu, Yutian Xiao, Xiangyu Zhao, Jingtong Gao, Wanyu Wang, Qing Li, and Jiliang Tang. Multimodal recommender systems: a survey. *ACM Computing Surveys*, 57(2):1–17, 2024.
- [55] Quoc-Tuan Truong, Aghiles Salah, and Hady Lauw. Multi-modal recommender systems: hands-on exploration. In *Proceedings of the ACM Conference on Recommender Systems (RecSys)*, pages 834–837, 2021.

- [56] Rui Sun, Xuezhi Cao, Yan Zhao, Junchen Wan, Zhou Kun, Fuzheng Zhang, Zhongyuan Wang, and Kai Zheng. Multi-modal knowledge graphs for recommender systems. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1405–1414, 2020.
- [57] Yongfeng Zhang, Qingyao Ai, Xu Chen, and W. Bruce Croft. Joint representation learning for top-N recommendation with heterogeneous information sources. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1449–1458, 2017.
- [58] Yue Guan, Qiang Wei, and Guoqing Chen. Deep learning based personalized recommendation with multi-view information integration. *Decision Support Systems*, 118:58–69, 2019.
- [59] Yunsen Hong, Hui Li, Xiaoli Wang, and Chen Lin. DEAMER: A deep exposure-aware multimodal content-based recommendation system. In *Proceedings of the International Conference on Database Systems for Advanced Applications (DASFAA)*, pages 621–637, 2020.
- [60] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L. Hamilton, and Jure Leskovec. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 974–983, 2018.
- [61] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: a review of methods and applications. *AI Open*, 1:57–81, 2020.
- [62] Rianne van den Berg, Thomas Kipf, and Max Welling. Graph convolutional matrix completion. *arXiv*, abs/1706.02263, 2017.
- [63] Jianing Sun, Yingxue Zhang, Wei Guo, Huifeng Guo, Ruiming Tang, Xiuqiang He, Chen Ma, and Mark Coates. Neighbor interaction aware graph convolution networks for recommendation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1289–1298, 2020.

- [64] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. Neural graph collaborative filtering. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2019.
- [65] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.
- [66] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.
- [67] Baoli Li and Liping Han. Distance weighted cosine similarity measure for text classification. In *Proceedings of the International Conference on Intelligent Data Engineering and Automated Learning (IDEAL)*, pages 611–618, 2013.
- [68] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.
- [69] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pages 1025–1035, 2017.
- [70] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. LightGCN: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 639–648, 2020.
- [71] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. *International Conference on Learning Representations*, 2018.

- [72] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 452–461, 2009.
- [73] Ruining He and Julian McAuley. VBPR: visual Bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 144–150, 2016.
- [74] Diederik P. Kingma and Jimmy Ba. Adam: a method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.
- [75] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.
- [76] Robin Burke. Hybrid recommender systems: survey and experiments. *User Modeling and User-Adapted Interaction*, 12, 2002.
- [77] Yashar Deldjoo, Markus Schedl, Balázs Hidasi, Yinwei Wei, and Xiangnan He. Multimedia recommender systems: algorithms and challenges. In *Recommender Systems Handbook*, pages 973–1014. 2022.
- [78] Congying Yan, Qingwei Xia, Yuhan Hu, and Li Li. Multi-modal knowledge graph recommendation based on self-supervised learning. In *Proceedings of the International Conference on Frontier Technologies of Information and Computer (ICFTIC)*, pages 618–625, 2024.
- [79] Mahima Kansal and Sohit Agarwal. Enhanced multimodal recommendation system for personalized lifestyle recommendations. In *Proceedings of the International Conference on Advancements in Computing Technologies and Artificial Intelligence (COMPUTATIA)*, page 36, 2025.
- [80] Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. KGAT: Knowledge graph attention network for recommendation. In *Proceedings of the*

- ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 950–958, 2019.
- [81] Qifan Wang, Yinwei Wei, Jianhua Yin, Jianlong Wu, Xuemeng Song, and Liqiang Nie. DualGNN: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia*, 25:1074–1084, 2023.
- [82] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 173–182, 2017.
- [83] Chintoo Kumar, C. Ravindranath Chowdary, and Ashok Kumar Meena. Recent trends in recommender systems: a survey. *International Journal of Multimedia Information Retrieval*, 13(4):41, 2024.
- [84] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.
- [85] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
- [86] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. *Learning internal representations by error propagation*, pages 318–362. MIT Press, Cambridge, MA, USA, 1986.
- [87] Quoc-Tuan Truong and Hady Lauw. Multimodal review generation for recommender systems. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 1864–1874, 2019.
- [88] Felipe L. A. Conceição, Flávio L. C. Pádua, Anisio Lacerda, Adriano C. Machado, and Daniel H. Dalip. Multimodal data fusion framework based on autoencoders for top-N recommender systems. *Applied Intelligence*, 49:3267–3282, 2019.

- [89] Lianjie Long, Faliang Huang, Yunfei Yin, and Youquan Xu. Multi-task learning for collaborative filtering. *International Journal of Machine Learning and Cybernetics*, pages 1–14, 2022.
- [90] Yichao Lu, Ruihai Dong, and Barry Smyth. Why I like it: multi-task learning for recommendation and explanation. In *Proceedings of the ACM Conference on Recommender Systems (RecSys)*, pages 4–12, 2018.
- [91] Chen Gao, Xiangnan He, Dahua Gan, Xiangning Chen, Fuli Feng, Yong Li, Tat-Seng Chua, and Depeng Jin. Neural multi-task recommendation from multi-behavior data. In *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, pages 1554–1557, 2019.
- [92] Guy Hadash, Oren Sar-Shalom, and Rita Osadchy. Rank and rate: multi-task learning for recommender systems. In *Proceedings of the ACM Conference on Recommender Systems (RecSys)*, pages 451–454, 2018.
- [93] Chintoo Kumar, Showkat Ahmad Dar, Charishma Batchu, Kundan Sai Singamaneni, and Venkata Lakshmi Alekhya Panguluri. Peer collaboration in DBLP using graph convolutional network. *SN Computer Science*, 6(1):1–7, 2025.
- [94] Ziteng Wu, Chengyun Song, Yunqing Chen, and Lingxuan Li. A review of recommendation system research based on bipartite graph. *MATEC Web of Conferences*, 336:05010, 2021.
- [95] Christopher Bishop. *Pattern Recognition and Machine Learning*. 2006.
- [96] Chong Li, Kunyang Jia, Dan Shen, C. J. Richard Shi, and Hongxia Yang. Hierarchical representation learning for bipartite graphs. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2873–2879, 2019.
- [97] Qiyao Ma, Xubin Ren, and Chao Huang. XRec: large language models for explainable recommendation. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 391–402, 2024.

- [98] Alejo Lopez-Avila and Jinhua Du. A survey on large language models in multi-modal recommender systems. *arXiv preprint arXiv:2505.09777*, 2025.
- [99] Chen Gao, Yu Zheng, Wenjie Wang, Fuli Feng, Xiangnan He, and Yong Li. Causal inference in recommender systems: a survey and future directions. *ACM Transactions on Information Systems*, 42(4):1–32, 2024.
- [100] Tong Chen, Hongzhi Yin, Guanhua Ye, Zi Huang, Yang Wang, and Meng Wang. Try this instead: personalized and interpretable substitute recommendation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 891–900, 2020.
- [101] Qinqin Wang, Elias Tragos, Neil Hurley, Barry Smyth, Aonghus Lawlor, and Ruihai Dong. Entity-enhanced graph convolutional network for accurate and explainable recommendation. In *Proceedings of the ACM Conference on User Modeling, Adaptation and Personalization (UMAP)*, pages 79–88, 2022.
- [102] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 2181–2187, 2015.