

*Master of Science (Research) in Computing Sciences*

*February 2026*

# *Advanced Recommendation Systems for Overcoming Cold Start and Data Sparsity Challenges*

---

*Yuanming Huang*

School of Computer Science

Faculty of Engineering and Information Technology

University of Technology Sydney

NSW - 2007, Australia



## CERTIFICATE OF ORIGINAL AUTHORSHIP

I, *Yuanming Huang*, declare that this thesis is submitted in fulfilment of the requirements for the award of Master of Science (Research) in Computing Sciences, in the *School of Computer Science, Faculty of Engineering and Information Technology* at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research was supported by an Australian Government Research Training Program (RTP) Scholarship [doi.org/10.82133/C42F-K220](https://doi.org/10.82133/C42F-K220).

Production Note:

SIGNATURE: Signature removed prior to publication.

Yuanming Huang

DATE: 1<sup>st</sup> February, 2026

PLACE: Sydney, Australia



## DEDICATION

*To my parents, whose unwavering love and belief in me made this journey possible.*



## ACKNOWLEDGMENTS

I would like to express my sincere gratitude to all those who have supported me throughout my Master’s research journey.

I wish to thank my parents, Dr. Peng Zhou and Jie Huang, for their steadfast support throughout my study. Their encouragement, patience, and financial assistance allowed me to devote myself to research and complete this thesis.

I am deeply grateful to my supervisory panel. I would like to thank my principal supervisor, Associate Professor Guangquan Zhang, for introducing me to the research direction of recommender systems and for his guidance and support during my candidature. I also sincerely thank my co-supervisors, Distinguished Professor Jie Lu, for her valuable advice, constructive feedback on paper writing, and continuous encouragement, which greatly improved the quality and clarity of this work.

Special thanks go to my co-supervisors, Dr. Keqiuyin Li. My research progress benefited substantially from her mentorship and generosity. In particular, when I felt stuck or overwhelmed, she patiently helped me troubleshoot experiments and locate issues in my code—even the mistakes I could not see myself. Her support and guidance played an important role in keeping my research on track.

I would also like to acknowledge my colleagues and friends at the Australian Artificial Intelligence Institute (AAIL). I am thankful for the supportive research environment, the helpful discussions, and the companionship that made my time at UTS both productive and memorable.

Finally, I would like to thank everyone who has contributed to this thesis in different ways.



## PUBLICATIONS

1. Huang, Y., Lu, J., Li, K., & Zhang, G. (2025b). Learning a wavelet neural filter with mamba for sequential recommendation. *2025 IEEE Symposium on Computational Intelligence in Image, Signal Processing and Synthetic Media (CISM)*, 1–7, Trondheim, Norway, March 17-20, 2025.
2. Huang, Y., Lu, J., Li, K., & Zhang, G. (2025a). Enhancing graph-based recommendations via fuzzy clustering and gated multi-view fusion. *2025 IEEE International Conference on Fuzzy Systems (FUZZ)*, 1–4, Reims, France, July 6-9, 2025.
3. Huang, Y., Lu, J., Li, K., & Zhang, G. (2026). Modality-adversarial generative enhancement for missing-modality multimodal recommendation. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*. Under review.



## TABLE OF CONTENTS

|                                                               |             |
|---------------------------------------------------------------|-------------|
| <b>Certificate of Original Authorship</b>                     | <b>i</b>    |
| <b>Dedication</b>                                             | <b>iii</b>  |
| <b>Acknowledgments</b>                                        | <b>v</b>    |
| <b>Publications</b>                                           | <b>vii</b>  |
| <b>Table of Contents</b>                                      | <b>ix</b>   |
| <b>List of Figures</b>                                        | <b>xi</b>   |
| <b>List of Tables</b>                                         | <b>xiii</b> |
| <b>Abstract</b>                                               | <b>xv</b>   |
| <b>1 Introduction</b>                                         | <b>1</b>    |
| 1.1 Background . . . . .                                      | 1           |
| 1.2 Problem Statement . . . . .                               | 2           |
| 1.3 Research Questions . . . . .                              | 3           |
| 1.4 Research Objectives . . . . .                             | 4           |
| 1.5 Expected Outcomes . . . . .                               | 4           |
| 1.6 Research Significance . . . . .                           | 6           |
| 1.7 Thesis Structure . . . . .                                | 7           |
| <b>2 Literature Review</b>                                    | <b>9</b>    |
| 2.1 Category of Recommendation Tasks . . . . .                | 9           |
| 2.2 Categories of General Recommendation Algorithms . . . . . | 12          |
| 2.2.1 Content-based Filtering . . . . .                       | 12          |
| 2.2.2 Knowledge-based Filtering . . . . .                     | 14          |

## TABLE OF CONTENTS

---

|          |                                                                          |           |
|----------|--------------------------------------------------------------------------|-----------|
| 2.2.3    | Collaborative Filtering . . . . .                                        | 15        |
| 2.2.4    | Knowledge Graph-Based Recommendation . . . . .                           | 19        |
| 2.2.5    | LLM-based Recommendation . . . . .                                       | 21        |
| 2.2.6    | Multimodal Recommendation . . . . .                                      | 22        |
| 2.3      | Chapter Summary . . . . .                                                | 24        |
| <b>3</b> | <b>Wavelet-Enhanced State-Space Models for Sequential Recommendation</b> | <b>25</b> |
| 3.1      | Introduction . . . . .                                                   | 25        |
| 3.2      | Preliminaries . . . . .                                                  | 27        |
| 3.2.1    | Problem Statement . . . . .                                              | 27        |
| 3.2.2    | Fourier Transform vs Wavelet Transform . . . . .                         | 28        |
| 3.3      | Methodology . . . . .                                                    | 28        |
| 3.3.1    | Embedding Layer . . . . .                                                | 29        |
| 3.3.2    | Wavelet Mamba Encoder . . . . .                                          | 30        |
| 3.3.3    | Prediction Layer and Loss Function . . . . .                             | 33        |
| 3.4      | Experiment . . . . .                                                     | 33        |
| 3.4.1    | Experimental Setup . . . . .                                             | 33        |
| 3.4.2    | Experimental Results . . . . .                                           | 35        |
| 3.4.3    | Additional Studies . . . . .                                             | 37        |
| 3.5      | Chapter Summary . . . . .                                                | 40        |
| <b>4</b> | <b>Graph-Based Recommendation with Sparse Interactions</b>               | <b>41</b> |
| 4.1      | Introduction . . . . .                                                   | 41        |
| 4.2      | Methodology . . . . .                                                    | 43        |
| 4.2.1    | Preliminaries . . . . .                                                  | 44        |
| 4.2.2    | Proposed Approach . . . . .                                              | 45        |
| 4.3      | Experiment . . . . .                                                     | 48        |
| 4.3.1    | Experimental Setup . . . . .                                             | 48        |
| 4.3.2    | Experimental Results . . . . .                                           | 50        |
| 4.3.3    | Additional Studies . . . . .                                             | 52        |
| 4.4      | Chapter Summary . . . . .                                                | 54        |
| <b>5</b> | <b>Robust Multimodal Recommendation with Incomplete Item Content</b>     | <b>57</b> |
| 5.1      | Introduction . . . . .                                                   | 57        |
| 5.2      | Preliminaries . . . . .                                                  | 61        |
| 5.2.1    | Problem Statement . . . . .                                              | 61        |

|          |                                                                   |           |
|----------|-------------------------------------------------------------------|-----------|
| 5.3      | Methodology . . . . .                                             | 62        |
| 5.3.1    | Modality Representation and Preference-Aware Filtering Module . . | 62        |
| 5.3.2    | Missing modality Completion Module . . . . .                      | 64        |
| 5.3.3    | Multimodal Recommendation Module . . . . .                        | 67        |
| 5.3.4    | Optimisation . . . . .                                            | 69        |
| 5.4      | Experiment . . . . .                                              | 69        |
| 5.4.1    | Datasets & Baselines . . . . .                                    | 71        |
| 5.4.2    | Experimental Settings . . . . .                                   | 73        |
| 5.4.3    | Evaluation Protocols . . . . .                                    | 73        |
| 5.4.4    | Implementation . . . . .                                          | 74        |
| 5.4.5    | Overall Performance . . . . .                                     | 74        |
| 5.4.6    | Ablation Study . . . . .                                          | 78        |
| 5.4.7    | Hyperparameter Sensitivity . . . . .                              | 79        |
| 5.4.8    | Representation Visualisation . . . . .                            | 81        |
| 5.4.9    | Statistical Significance . . . . .                                | 82        |
| 5.5      | Chapter Summary . . . . .                                         | 83        |
| <b>6</b> | <b>Conclusion</b>                                                 | <b>85</b> |
| 6.1      | Contributions to the Field . . . . .                              | 85        |
| 6.2      | Recommendations for Future Research . . . . .                     | 87        |
| 6.3      | Closing Remark . . . . .                                          | 88        |
|          | <b>Bibliography</b>                                               | <b>91</b> |

## LIST OF FIGURES

| FIGURE                                          | Page |
|-------------------------------------------------|------|
| 1.1 Overview of the thesis framework. . . . .   | 5    |
| 3.1 Overall Framework of Wavelet Mamba. . . . . | 29   |
| 3.2 Case study on ML-1M. . . . .                | 38   |
| 3.3 Sensitivity Study on ML-1M. . . . .         | 39   |

## LIST OF FIGURES

---

|     |                                                                                                        |    |
|-----|--------------------------------------------------------------------------------------------------------|----|
| 4.1 | Overall framework of FuzzGCN. . . . .                                                                  | 44 |
| 4.2 | Sensitivity Study of the layers on Games. . . . .                                                      | 53 |
| 4.3 | Sensitivity Study of the clusters on Games. . . . .                                                    | 54 |
| 5.1 | Illustration of challenges in multimodal recommendation. . . . .                                       | 59 |
| 5.2 | Empirical evidence of missing modality impact and cross-modal imputation bias. . . . .                 | 59 |
| 5.3 | The overall framework of MAGERec. . . . .                                                              | 63 |
| 5.4 | Missing modality robustness on Clothing. . . . .                                                       | 78 |
| 5.5 | Effect of $\alpha$ , $\gamma$ , $\lambda_1$ and $\lambda_2$ on Baby . . . . .                          | 80 |
| 5.6 | Effect of $\alpha$ , $\gamma$ , $\lambda_1$ and $\lambda_2$ on Sports . . . . .                        | 80 |
| 5.7 | Distribution of item modality representations on missing modality setting. . . . .                     | 81 |
| 5.8 | Distribution of item modality representations on missing modality and item cold start setting. . . . . | 82 |
| 5.9 | $T$ -test between the best baseline and MAGERec. . . . .                                               | 83 |

## LIST OF TABLES

| TABLE                                                                    | Page |
|--------------------------------------------------------------------------|------|
| 3.1 Statistics of the Experimental Data after Processing. . . . .        | 34   |
| 3.2 Performance Comparison Across Datasets. . . . .                      | 36   |
| 3.3 Ablation Study on ML-1M. . . . .                                     | 37   |
| 4.1 Processed Dataset Statistics. . . . .                                | 48   |
| 4.2 Overall Performance Comparison Across Datasets. . . . .              | 51   |
| 4.3 Ablation Study on Arts. . . . .                                      | 53   |
| 5.1 Statistics of the Experimental Data after Processing. . . . .        | 71   |
| 5.2 Summary of baselines. . . . .                                        | 72   |
| 5.3 Performance on Missing Modality Setting . . . . .                    | 76   |
| 5.4 Performance on Missing Modality and Item Cold Start Setting. . . . . | 77   |
| 5.5 Ablation Study on Missing Modality Setting. . . . .                  | 79   |



## ABSTRACT

Recommender systems alleviate information overload by delivering personalised content, yet their effectiveness in real platforms is frequently constrained by imperfect signals, including sparse user–item interactions, item cold-start, and non-stationary user behaviour. Moreover, as modern recommender systems increasingly rely on item-side content, missing or degraded modalities further distort representation learning and fusion. This thesis studies robust recommendations under these practical imperfections through three interconnected contributions that progress from temporal modelling to structural representation enhancement, and finally to missing-modality robustness.

Non-stationary user behaviour is examined through the lens of sequential recommendation with long interaction histories. It introduces a state-space-based sequential architecture augmented with a wavelet frequency rescaling mechanism, enabling frequency-aware adaptation to dynamic user behaviours while maintaining efficient modelling of long-range dependencies.

Interaction sparsity is investigated from an item representation perspective in graph-based recommendation. It proposes a graph learning framework that injects global relational priors via fuzzy clustering and strengthens item representations beyond local neighbourhood aggregation, improving robustness when collaborative signals are limited or unevenly distributed.

Item cold-start is further studied in multimodal recommendation, where item content modalities may be incomplete, inducing a completeness shift between complete and incomplete items and entangling modality-invariant and modality-specific information. It proposes a modality-adversarial generative enhancement framework that reconstructs missing modalities from observed ones, aligns complete–incomplete distributions through domain-adversarial learning, and disentangles invariant versus specific signals via mutual-information regularisation to support more reliable multimodal fusion.

Taken together, these three studies demonstrate that robust recommendations in realistic settings benefit from jointly addressing non-stationary temporal dynamics in user behaviour, globally informed structural modelling under interaction sparsity, and distribution-aware generation and fusion under missing modalities. By connecting frequency-aware sequential modelling, fuzzy-structured graph representation learning, and completeness-shift-aware multimodal enhancement, this thesis provides a coherent framework for building recommender systems that remain dependable when data are sparse, behaviours evolve, and content is incomplete.



## INTRODUCTION

## 1.1 Background

In the modern era, characterised by an unprecedented surge in digital information and a vast increase in Internet users, individuals face the overwhelming challenge of information overload (Bawden & Robinson, 2020). While search engines such as Google and Bing provide a foundational means of accessing relevant content, they often lack the capacity to prioritise and personalise information based on individual user preferences (Chiravirakul & Payne, 2014). This limitation underscores the critical demand for advanced recommender systems that can efficiently tailor content to user needs and enhance the relevance of information delivery.

Recommender systems are advanced software tools designed to suggest the most relevant items—such as products or services—to users, whether individuals or businesses, by predicting a user’s potential interest in each item. This prediction leverages information about the items themselves, user profiles, and the interactions between them (Lu et al., 2015). These systems effectively mitigate information overload by filtering vast, dynamic datasets to align with user preferences and past interactions. Through personalised recommendations, they enhance online experiences and drive engagement, proving essential in domains such as e-commerce (C. Gao et al., 2023) and e-learning (Wan & Niu, 2020), where they support user decision-making and foster deeper interaction in both personal and professional contexts. Nowadays, recommender systems have become integral to major internet platforms, including Amazon, YouTube, Netflix, Yahoo, and Facebook, where they

drive user engagement and satisfaction by estimating the utility of items and predicting their relevance for individual users.

Over the years, recommender systems have evolved from simple content-based filtering and collaborative filtering approaches, which recommend items based on user or item similarities, to more sophisticated techniques. Matrix factorisation methods, for example, became popular as a foundational approach in collaborative filtering by capturing latent factors in user-item interactions, which enhanced recommendation accuracy (Koren et al., 2009). As recommendation needs grew more complex, various models emerged, combining multiple approaches to address limitations such as cold-start, data sparsity and personalisation. With the advent of deep learning, recommender systems were further transformed, allowing for complex, non-linear modelling of user and item interactions and enabling the creation of highly personalised recommendations. Deep learning-based systems, such as neural collaborative filtering(X. Wang et al., 2019) and graph neural networks(X. He et al., 2020), are now standard baselines in recommendation tasks due to their ability to capture intricate patterns in user behaviour.

Recently, the emergence of large language model (LLM)-based recommender systems has further expanded the capabilities of these systems. By leveraging vast linguistic knowledge and context understanding, LLM-based recommender systems can interpret nuanced user queries and generate contextual recommendations, opening new possibilities in natural language interactions and content discovery (Y. Gao et al., 2023). Additionally, cross-domain recommendation has become an area of focus, addressing the need to transfer knowledge from one domain to another to improve recommendation accuracy and mitigate data sparsity issues, especially in scenarios where user data in the target domain is limited (S. Zhao & Sahebi, 2024). Cross-domain and LLM-based approaches thus represent the latest advancements in the field, enhancing both the flexibility and scope of modern recommender systems. Together, these foundational and advanced algorithms form the basis of modern recommender systems and will be discussed in detail in Section 2.2.

## 1.2 Problem Statement

Recommender systems are a critical component of modern online platforms, yet their real-world reliability is increasingly challenged by imperfect and heterogeneous signals (Q. Guo et al., 2022; Shambour & Lu, 2012). In practice, user preferences evolve over time, interaction logs are often sparse and long-tailed, and item-side content can be incomplete or degraded (Q. Zhang et al., 2023). However, many existing recommendation frameworks

are developed under simplifying assumptions—such as relatively stationary behaviours, sufficiently dense feedback, or complete auxiliary information—which creates a gap between model design and deployment reality (R. Wu et al., 2024). This gap can lead to unstable ranking quality, biased representations, and poor generalisation to new users, new items, or shifting environments. Therefore, the central problem motivating this research is to develop principled recommendation learning paradigms that remain effective and robust under three practical imperfections: non-stationary long-range behavioural dynamics, sparsity-driven representation fragility, and incomplete multimodal information.

To frame the scope of this study, we highlight three key challenges aligned with the research questions and objectives:

- **Non-stationary long-range behaviour modelling:** User interests are not static; they drift with time, context, and external factors, while historical interactions may span long horizons. Existing sequential recommenders may struggle to simultaneously capture long-range dependencies and adapt to evolving behavioural patterns, often leading to either short-term myopia or inefficient modelling when histories grow long.
- **Representation learning under sparsity:** Real-world interaction data are typically sparse and imbalanced, where most users and items have limited feedback. Under such conditions, learned representations can become fragile and biased, and neighbourhood-based inference may be unreliable due to weak or noisy relational evidence. This challenge is further amplified for new or long-tail items, where insufficient interactions hinder effective personalisation.
- **Robust multimodal learning with incomplete information:** Multimodal recommendation promises richer item understanding, yet in realistic platforms, item content can be missing, partial, or low-quality. Incomplete observations can induce mismatches between items with complete versus incomplete content and make multimodal fusion unreliable, resulting in performance degradation and reduced robustness at inference time.

## 1.3 Research Questions

Our goal is to develop advanced recommender systems that effectively address key limitations in existing models. Specifically, we formulate the following research questions that our study aims to answer:

- **RQ1:** How to learn non-stationary user behaviour and long-range dependencies in sequential recommendation?
- **RQ2:** How to enrich item representations in graph-based recommendation under sparsity?
- **RQ3:** How to improve the robustness of multimodal recommendation when modalities are incomplete?

## 1.4 Research Objectives

To answer the research questions raised above, we set the following research objectives:

- **RO1:** Develop an approach that improves sequential recommendation by modelling non-stationary user behaviours and long-range dependencies (aims to answer RQ1).

User preferences often evolve over time, and long interaction histories contain both short-term intents and long-term patterns. Robust sequential models are therefore required to capture long-range dependencies while remaining sensitive to time-varying behavioural changes.

- **RO2:** Develop an approach that enhances item representations in graph-based recommendation under sparse user-item interactions (aims to answer RQ2).

Data sparsity limits the reliability of locally aggregated collaborative signals, which can result in weak or biased item representations. Strengthening item embeddings with more informative relational structure is essential for stable recommendation performance when feedback is limited or unevenly distributed.

- **RO3:** Develop an approach that ensures the robustness of multimodal recommendation when item content is incomplete (aims to answer RQ3).

In real-world platforms, item modalities such as text or images may be missing or degraded, causing incomplete items to differ systematically from complete ones and making multimodal fusion unreliable. A robust multimodal recommender should remain dependable under such incompleteness and avoid performance degradation.

## 1.5 Expected Outcomes

This thesis is expected to deliver the following outcomes:

- A sequential recommendation framework that models time-varying user behaviours and captures long-range dependencies, enabling more reliable preference modelling for users with long and evolving interaction histories.
- A graph-based recommendation framework that enhances item representations under sparse user-item interactions by exploiting richer relational structure, improving recommendation robustness when collaborative signals are limited or unevenly distributed.
- A multimodal recommendation framework that remains robust when item content modalities are incomplete or degraded, supporting dependable representation learning and fusion for incomplete items in real-world settings.

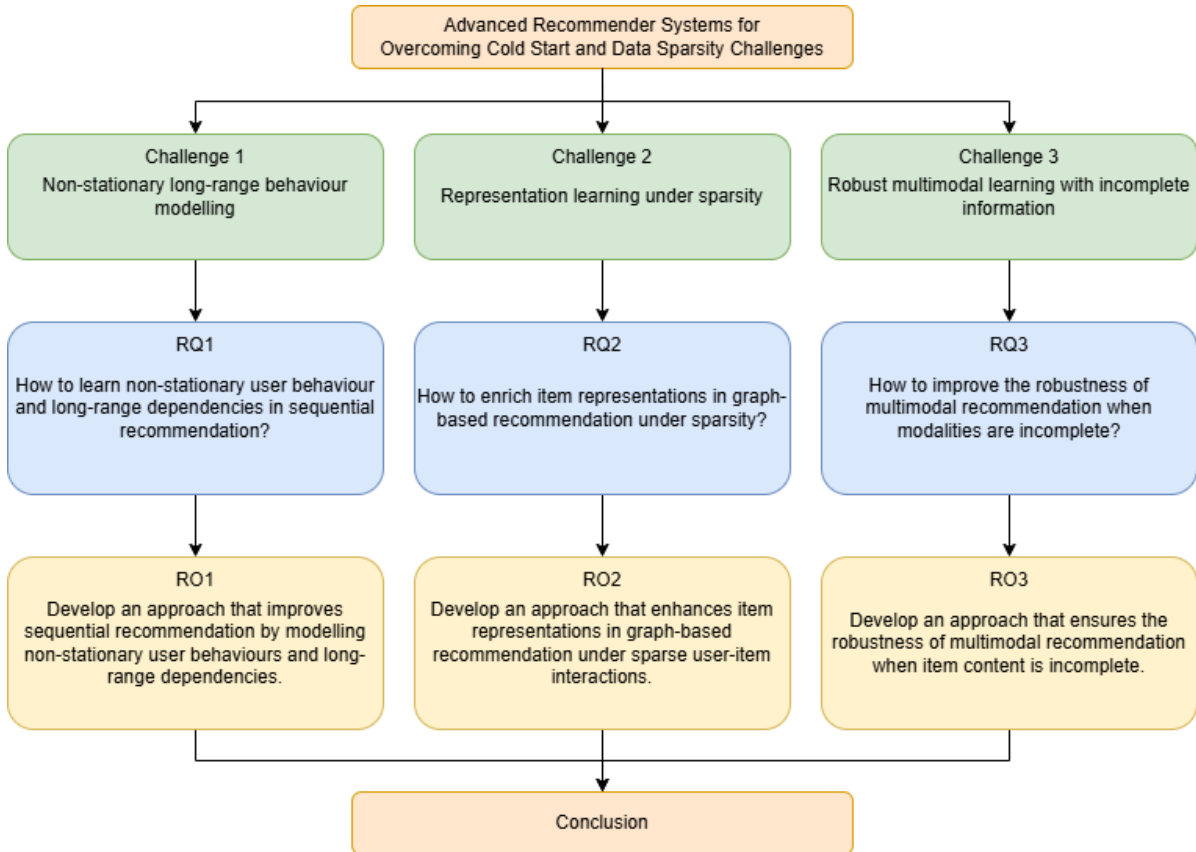


Figure 1.1: Overview of the thesis framework.

Figure 1.1 illustrates the overall thesis framework, which links the identified practical challenges to the corresponding research questions and objectives. Specifically, the thesis is organised around three aligned research streams: sequential recommendation, graph-based

recommendation, and multimodal recommendation, each targeting a distinct real-world imperfection and collectively contributing to the thesis outcomes.

## 1.6 Research Significance

- **Theoretical Significance:** In a theoretical view, the research significance can be articulated in the following three aspects: (i) This thesis enriches the theoretical analysis of sequential recommendation by characterising how user preferences evolve over long horizons under non-stationary behaviours, and by clarifying the role of multi-scale temporal structure in modelling dynamic preference signals. (ii) It advances representation learning theory for sparse recommendation by formalising the value of global relational priors in complementing local neighbourhood evidence, thereby explaining when and why sparsity induces biased or unstable item representations and how such fragility can be mitigated in a principled manner. (iii) This thesis broadens the theoretical foundations of multimodal recommendation with incomplete information by framing missing-modality scenarios as a joint problem of partial observability and distribution mismatch, which elucidates the principles required for reliable multimodal integration and generalisation in realistic settings and motivates further research on robustness in recommender systems.
- **Practical Significance:** From a practical perspective, the findings of this thesis can contribute to real-world recommender systems in the following three aspects: (i) The findings of this thesis benefit real-world recommender systems where user interaction histories are long and continuously evolving. By improving the capability to model dynamic preferences over extended periods, the proposed line of research supports more stable and timely personalisation in applications such as e-commerce, streaming media, and online platforms that operate under rapidly changing user interests. (ii) This thesis contributes to practical recommendation scenarios where user-item interactions are sparse. By strengthening item representations when feedback signals are limited, the outcomes of this research can improve recommendation quality for new or infrequently interacted items, helping platforms maintain reliable service quality even with constrained behavioural data. (iii) The research has practical value for multimodal recommendation in realistic environments where item-side content can be missing, incomplete, or degraded. By enhancing robustness under partial information, the results improve system reliability and user experience, and reduce

the risk of performance deterioration caused by imperfect or unavailable modalities across diverse deployment settings.

## 1.7 Thesis Structure

To present our work in a clear and coherent manner, the chapters of this thesis are organised as follows:

- **Chapter 2:** This chapter provides a comprehensive literature review of recommender systems, covering the category of recommendation tasks. It further reviews representative paradigms in sequential recommendation, graph-based collaborative filtering, and multimodal recommendation. Thereby establishing the theoretical and practical foundation for the subsequent chapters.
- **Chapter 3:** A sequential recommendation framework is developed in this chapter. It investigates how to capture long-range dependencies while adapting to evolving behavioural patterns, and reports empirical studies to demonstrate the effectiveness and efficiency of the proposed solution on benchmark sequential recommendation datasets.
- **Chapter 4:** A graph-based recommendation framework is presented in this chapter. It focuses on graph-based recommendation under sparse user-item interactions and item cold-start. It studies how to strengthen item representations when neighbourhood signals are weak or biased, and validates the proposed approach through systematic experiments and analyses, highlighting improvements in recommendation quality in sparse settings.
- **Chapter 5:** Modality missingness in multimodal recommendation is explored in this chapter. It investigates multimodal recommendation in realistic scenarios where item-side content is incomplete, missing, or degraded. It formulates the robustness issues induced by partial observability and distribution mismatch, and presents experimental results and diagnostic analyses to show how reliable multimodal integration can be achieved under missing-modality conditions.
- **Chapter 6:** This chapter concludes the thesis and discusses future research directions.



## LITERATURE REVIEW

In this chapter, we review the foundations and recent advances of recommender systems to establish the context for the research questions addressed in this thesis. We first introduce major categories of recommendation tasks, clarifying the differences between general recommendation, sequential recommendation, and click-through rate prediction. We then summarise representative algorithmic paradigms for general recommendation, including content-based filtering, knowledge-based filtering, and collaborative filtering, with an emphasis on matrix factorisation and graph-based learning that underpin modern recommender systems. Building on this foundation, we further review emerging directions that are closely related to the thesis objectives, including knowledge graph-based recommendation, LLM-based recommendation, and multimodal recommendation. Finally, we focus on multimodal recommendation under modality missingness, categorising existing solutions into preprocessing-based completion and model-level strategy design, and highlighting the remaining gaps that motivate the methods developed in the subsequent chapters.

### 2.1 Category of Recommendation Tasks

As these foundational and advanced methods evolved, recommender systems have branched into specialised tasks, each addressing distinct aspects of user engagement. The following sections will explore key types of recommendation tasks, beginning with general recommendations, and then examining sequential recommendations and click-through rate

predictions.

**General Recommendation** Unlike specialised recommendation tasks, which may focus on specific user actions or domain-specific behaviours, general recommendation aims to provide users with relevant items based on a broad analysis of historical user-item interactions. The main function of the user-item interactions is:

$$f : U \times I \rightarrow D, \quad (2.1)$$

where a single item  $i \in I$  and a specific user  $u \in U$ .  $D$  is the final recommendation list. For a given user  $u \in U$  and an item  $i \in I$ , the function  $f(u, i)$  computes the utility or relevance of item  $i$  for user  $u$ . The output  $D$  is a ranked list of items, ordered by their predicted utility, providing the user with a tailored set of recommendations based on items they have not yet interacted with (Q. Zhang et al., 2021).

In general recommendation, the utility  $f(u, i)$  often relies on factors such as past user interactions, explicit user ratings, or inferred preferences. Recommendation algorithms find an item for the user by maximising  $f(u, i)$ :

$$\forall u \in U, \arg \max_{i \in I} f(u, i), \quad (2.2)$$

$\arg \max$  identifies the item with the highest utility that has not yet been consumed by the user. By ranking items based on this utility score, the recommender system generates  $D$  that aims to maximise user satisfaction and engagement (Lu et al., 2020).

General recommendation, therefore, serves a wide range of applications, from e-commerce product suggestions to personalised content delivery on streaming platforms, offering broad solutions aimed at enhancing user satisfaction and engagement across diverse domains.

**Sequential Recommendation** Sequential recommendation aims to predict a user's next interaction based on their past interaction history, capturing both short-term and long-term behavioural patterns. Unlike general recommendation tasks, which provide a broad analysis of user-item preferences, sequential recommendation leverages the temporal order of user interactions to deliver contextually relevant predictions. In this setting, each user has a chronologically ordered interaction sequence:  $S^u = [i_1^u, i_2^u, \dots, i_n^u]$ , where  $i_t^u$  denotes the  $t$ -th item that user  $u$  interacted with, and  $n$  is the sequence length.

The primary objective of sequential recommendation is to predict the next item in the user's interaction sequence, specifically targeting the  $(n + 1)$ -th item, denoted  $i_{n+1}^u$ . The

system generates a Top- $k$  recommendation list,  $\mathcal{L}_u = [i_{n+1}^{u,1}, i_{n+1}^{u,2}, \dots, i_{n+1}^{u,k}]$ , by selecting items with the highest predicted probabilities:

$$\mathcal{L}_u = \text{Top-}k(P(i | S^u)) \quad \forall i \in I, \quad (2.3)$$

where  $P(i | S^u)$  represents the probability of user  $u$  interacting with item  $i$  given their interaction history  $S^u$  (C. Liu et al., 2024; Shin et al., 2024).

Traditional approaches in sequential recommendation include methods like Markov chains and sequential pattern mining, which model item-to-item transitions to predict future interactions (R. He & McAuley, 2016a; Rendle et al., 2010; Yap et al., 2012). With advances in machine learning, latent representation models were introduced, capturing user preferences through learned embeddings. A notable example is the Gated Feedback Frequency model (GFF) (Hidasi & Tikk, 2016). Subsequently, the rise of deep learning brought models based on convolutional and recurrent neural networks, such as Caser (Tang & Wang, 2018) (CNN-based) and GRU4Rec (Hidasi, Karatzoglou, et al., 2016) (RNN-based), which significantly enhanced the modelling of sequential patterns.

Recently, transformer-based models like SASRec (Kang & McAuley, 2018) and BERT4Rec (Sun et al., 2019) have further advanced sequential recommendation by applying self-attention mechanisms to capture both local and global dependencies within user sequences. However, deep learning models often face challenges with stability during training, known as the "overshooting problem." To address this, models like FMLPRec (K. Zhou et al., 2022) introduced Fourier frequency filters to control noise, while BsaRec (Shin et al., 2024) improved on this by better integrating low- and high-frequency information. More recently, Mamba4Rec (C. Liu et al., 2024) has utilised state space models to enhance computational efficiency, combining rapid inference with accuracy improvements for long user sequences.

While sequential recommender systems are designed to capture temporal dynamics in user behaviour, they differ from general recommendation in both task formulation and modelling assumptions. General recommendation usually focuses on estimating user-item relevance from historical interactions, whereas sequential recommendation explicitly considers the chronological order of user actions to predict future preferences. Therefore, this section provides a concise overview of representative sequential recommendation models, while the task-specific motivation and technical development are further discussed in Chapter 3.

**Click-Through Rate Prediction** Click-Through Rate Prediction, or CTR Prediction, is a recommendation task focused on estimating the likelihood that a user  $u$  will click on a

specific item  $i$ . Unlike general and sequential recommendation tasks, which primarily rank items based on utility or sequential patterns, CTR Prediction models the probability of a user interacting with an item as a binary classification problem. The objective of CTR prediction is to determine  $P(y = 1 \mid u, i)$ , where  $y = 1$  denotes a positive interaction, such as a click. The CTR Prediction function can be defined as:

$$f_{\text{CTR}} : U \times I \rightarrow [0, 1], \quad (2.4)$$

where  $f_{\text{CTR}}$  outputs the predicted probability that user  $u$  will click on item  $i$ , given their attributes and past interactions. Items are then ranked based on these probabilities, and the system presents a list of items with the highest predicted CTR values.

CTR prediction is commonly applied in online advertising, where accurately predicting click behaviour directly impacts revenue. For example, platforms like Facebook and Google use CTR Prediction models to select and display ads most likely to attract user clicks, optimising ad relevance and user engagement (Pan et al., 2023; X. Wang, 2020).

Models used for CTR prediction often include Support Vector Machines (SVM) (Hearst et al., 1998), logistic regression (Hosmer et al., 2013), and deep learning-based approaches (Feng et al., 2019; Pi et al., 2019; Qin et al., 2021), which differ from typical recommendation algorithms. Therefore, the algorithm discussions in the following sections will primarily focus on methods for general recommendation tasks.

## 2.2 Categories of General Recommendation Algorithms

### 2.2.1 Content-based Filtering

Content-based filtering is one of the earliest and most classical approaches to recommendation. As the name suggests, content-based filtering relies on the attributes or content of items to make recommendations. Each item is characterised by a set of features that define its properties—such as genre, author, or keywords for a book or movie. The system recommends items to a user by analysing the features of items the user has previously interacted with and identifying similar items based on these features (Q. Zhang et al., 2021).

Item representation plays a critical role in determining the relevance of recommendations. Typically, each item is represented as a vector of features that describe its content, allowing the system to compute the similarity between items. A classic approach to item representation is the vector space model, where items are represented in a high-dimensional space based on attributes or keywords. One of the most widely used techniques within this

model is Term Frequency-Inverse Document Frequency (TF-IDF), which assigns weights to terms in an item's description to reflect their importance relative to other items (Jones, 1972). By using TF-IDF, the system can identify significant keywords associated with each item, enabling effective comparison between items based on shared or related terms.

Recently, methods like Word2Vec have been introduced to improve item representation in content-based filtering. Unlike traditional TF-IDF, Word2Vec captures semantic relationships between words by mapping them into dense, low-dimensional vectors. This approach enables the system to detect deeper connections between items based on contextual similarity rather than just keyword overlap. For example, Word2Vec can help identify that "thriller" and "mystery" movies may be similar in user preference, even if these specific terms don't overlap directly (Mikolov et al., 2013).

User profiling is another essential component in content-based filtering, as it captures the preferences and interests of each individual user to generate personalised recommendations. In a content-based system, a user profile is typically created by analysing the features of items the user has previously interacted with, such as items they have viewed, rated, or purchased. This profile represents the user's preferences as a vector, often using the same feature space as the items themselves. For example, if the system is recommending movies, the user profile may include a weighted vector of preferred genres, directors, or keywords based on past interactions.

To construct a user profile, the system aggregates information from the features of the items associated with that user. A common technique is to average the feature vectors of all items a user has interacted with, weighted by the level of interaction or explicit feedback, such as ratings. This approach results in a user preference vector that reflects the user's most significant interests.

The similarity between a user's profile and an item's feature vector is then calculated—often using cosine similarity or other distance metrics—to determine the relevance of the item to the user. Cosine similarity measures the cosine of the angle between the user and item vectors, with values closer to 1 indicating higher similarity. The cosine similarity between two vectors  $\mathbf{A}$  and  $\mathbf{B}$  is defined as:

$$\text{sim}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \cdot \sqrt{\sum_{i=1}^n B_i^2}}, \quad (2.5)$$

where  $\mathbf{A} \cdot \mathbf{B}$  is the dot product of vectors  $\mathbf{A}$  and  $\mathbf{B}$ , and  $\|\mathbf{A}\|$  and  $\|\mathbf{B}\|$  denote the magnitudes of  $\mathbf{A}$  and  $\mathbf{B}$ , respectively. Cosine similarity provides a normalised measure of similarity, which is particularly effective in high-dimensional spaces where vector magnitudes may vary

significantly (Steck et al., 2024).

Common methods for calculating similarity-based recommendations include K-Nearest Neighbours (KNN), which identifies the  $k$  most similar items to a user profile by comparing vectors (G. Guo et al., 2003), and matrix factorisation techniques that reduce the dimensionality of feature vectors for more efficient similarity computation (Koren et al., 2009). Other distance metrics, such as Euclidean distance and Pearson correlation, can also be used depending on the data characteristics and recommendation goals.

Content-based filtering offers several benefits, particularly its ability to provide personalised recommendations to new users with limited interaction history, helping to address the "cold start" problem (Shambour & Lu, 2012). Because recommendations are based on the content and features of items, the system does not rely on other users' data, ensuring privacy and independence in the recommendation process. Additionally, content-based filtering systems are straightforward to implement and interpret, as recommendations are directly tied to user preferences and item attributes. This transparency enables users and developers to understand why specific items are recommended.

However, content-based filtering also has limitations. One of the primary drawbacks is the lack of diversity in recommendations, as the system is content-dependent and tends to recommend items similar to those a user has already interacted with, potentially leading to a narrow range of suggestions and limiting exploration (Balabanović & Shoham, 1997). Furthermore, content-based methods require comprehensive and well-structured item metadata, which may not always be available, especially in domains with complex or ambiguous item characteristics. Another limitation is the challenge of capturing user preferences in more abstract or non-content-based aspects, such as novelty or popularity, as content-based filtering focuses solely on item attributes rather than broader user-item interactions across the system.

Overall, while content-based filtering remains a foundational technique for recommender systems, its limitations have led to the development of more advanced methods, such as knowledge-based filtering and collaborative filtering, to provide a more comprehensive and diverse recommendation experience.

### **2.2.2 Knowledge-based Filtering**

Knowledge-based filtering is a recommendation approach that uses explicit knowledge about items, users, and their interactions to generate recommendations (Burke, 2002). Knowledge-based filtering systems are built on predefined rules and domain knowledge that map user requirements to item features. This approach is particularly useful in do-

mains where user preferences are complex or need to be highly customised, such as in recommending financial products, real estate, or specialised services.

One common technique within knowledge-based filtering is case-based reasoning (CBR), which makes recommendations by finding items that are similar to cases that have satisfied previous users with similar requirements (Aamodt & Plaza, 1994). In a case-based reasoning system, user queries are matched to a database of past cases or scenarios, each representing a unique combination of user needs and item attributes. For instance, in a travel recommender system, a user looking for a “beach vacation with family-friendly activities” might be recommended destinations based on similar cases of past users with comparable preferences. Case-based reasoning provides an intuitive way to handle complex, preference-based queries without requiring large amounts of historical user interaction data.

Knowledge-based filtering, including case-based reasoning, is particularly advantageous in situations where user needs are specific and interaction data is limited. However, this approach requires a comprehensive, well-structured knowledge base and detailed information about user needs, making it labour-intensive to maintain and less adaptable to dynamic environments. Consequently, research on knowledge-based filtering has become less common in modern recommender systems, as newer methods have evolved to address a wider range of recommendation challenges in the dynamic world nowadays.

### 2.2.3 Collaborative Filtering

**Traditional Models** Memory-based collaborative filtering is one of the earliest and most widely used techniques in recommender systems. This approach relies directly on the user-item interaction matrix to make recommendations based on observed similarities between users or items. Memory-based collaborative filtering can be categorised into two main types: user-based and item-based collaborative filtering.

In user-based collaborative filtering, the system identifies users with similar preferences by analysing their interaction histories. For example, if user  $u$  has rated or interacted with a set of items similarly to another user  $v$ , then items that  $v$  has enjoyed are likely to be recommended to  $u$ . The similarity between users can be calculated using metrics like cosine similarity (Steck et al., 2024), and Pearson correlation (Pearson & Galton, 1895), with cosine similarity being one of the most common methods. The user-based approach is intuitive and performs well in environments with sufficient user overlap.

In item-based collaborative filtering, the system recommends items that are similar to those the user has already interacted with. Rather than focusing on user similarities, item-based collaborative filtering looks at similarities between items based on user interactions.

For instance, if two items have been frequently rated or viewed by the same group of users, they are considered similar, and one can be recommended to users who liked the other. This approach is often more scalable than user-based filtering, especially in systems with large user bases, as the number of items tends to be smaller and more stable than the number of users over time.

Memory-based collaborative filtering is straightforward to implement and does not require complex model training, making it highly interpretable and effective for smaller datasets. However, it suffers from several limitations, including cold-start problems, data sparsity and scalability issues. The approach's reliance on direct similarity calculations can lead to poor performance when user-item interaction data is sparse. Additionally, memory-based models may struggle to capture complex relationships, such as non-linear user-item interactions, limiting their effectiveness in highly dynamic environments.

**Matrix Factorization-based Models** Matrix factorisation (MF) has become one of the most influential and widely adopted techniques in collaborative filtering for recommender systems, primarily due to its ability to handle large, sparse datasets. The basic idea behind matrix factorisation is to decompose the user-item interaction matrix into lower-dimensional representations of users and items (Koren et al., 2009). In this approach, each user and item is represented by latent vectors in a shared feature space, capturing hidden relationships between users and items that are not explicitly observable in the original interaction data.

A prominent matrix factorisation method is Singular Value Decomposition (SVD), which factorises the user-item interaction matrix  $R$  into three matrices:

$$R \approx U\Sigma V^T, \quad (2.6)$$

where  $U$  and  $V$  are matrices of latent factors for users and items, respectively, and  $\Sigma$  is a diagonal matrix containing singular values that scale the latent factors. SVD was popularised by the Netflix Prize competition, where it demonstrated significant performance improvements over traditional memory-based collaborative filtering methods (Takács et al., 2008). The factorised matrices capture latent features that implicitly define user preferences and item characteristics, allowing for more accurate predictions even in cases of sparse data.

Another influential method in this category is Non-negative Matrix Factorisation (NMF), which imposes non-negativity constraints on the latent factors. NMF is particularly useful in scenarios where interpretability of the latent factors is essential, as the non-negative constraints ensure that the latent features have a clearer, additive representation (Belyadi &

Haghighat, 2021). This technique has been effective in applications where understanding user preferences in positive-only terms (e.g., likes or positive ratings) is crucial.

A widely adopted extension of matrix factorisation is Probabilistic Matrix Factorisation (PMF), which treats user-item interactions as probabilistic events. PMF introduces a probabilistic framework to MF, allowing for regularisation that reduces the risk of overfitting and enhances the robustness of the model, particularly for sparse datasets (Salakhutdinov & Mnih, 2007). Building on this, Bayesian Personalised Ranking (BPR) is a pairwise ranking approach developed to improve recommendation ranking rather than just predict ratings. BPR optimises for the relative order of item preferences rather than absolute ratings, which is often more effective in real-world recommendation scenarios (Rendle et al., 2009).

Matrix factorisation techniques, despite their advantages, have limitations. They assume linear relationships between users and items, which can oversimplify complex interactions. Moreover, traditional MF approaches do not consider temporal dynamics or sequential patterns in user behaviour. To address these limitations, hybrid models and deep learning extensions have been developed, which incorporate additional contextual information or non-linear interactions. Nevertheless, matrix factorisation remains a foundational approach in collaborative filtering, providing interpretable and computationally efficient solutions for recommender systems.

**Graph-based Models** In recent years, graph-based models have become an increasingly popular approach in collaborative filtering, especially for recommendation tasks that require capturing complex user-item relationships beyond simple interactions. By representing users and items as nodes in a graph and their interactions as edges, graph-based models can better capture higher-order connectivity and exploit user-item structural information. Graph Neural Networks (GNNs) have been particularly successful in this domain, with models such as NGCF (Neural Graph Collaborative Filtering), LightGCN, and SimGCL demonstrating state-of-the-art performance in recommendation.

Neural Graph Collaborative Filtering (NGCF) was one of the pioneering models to introduce GNNs in collaborative filtering. NGCF extends traditional collaborative filtering by leveraging GNN layers to propagate embeddings across multiple layers in the user-item interaction graph, thereby capturing high-order connectivity between users and items (X. Wang et al., 2019). This model generates enriched user and item representations by iteratively aggregating information from neighbours, allowing it to learn more expressive embeddings that capture latent relationships between users and items.

The core of NGCF's propagation mechanism is represented by the following equation:

$$e_u^{(k+1)} = e_u^{(k)} + \sum_{i \in \mathcal{N}(u)} \frac{1}{\sqrt{|\mathcal{N}(u)| \cdot |\mathcal{N}(i)|}} \sigma \left( W^{(k)} e_i^{(k)} + b^{(k)} \right), \quad (2.7)$$

where  $e_u^{(k+1)}$  is the embedding of user  $u$  at the  $(k+1)$ -th layer,  $\mathcal{N}(u)$  denotes the set of items interacted with by user  $u$ ,  $W^{(k)}$  and  $b^{(k)}$  are the learnable weight matrix and bias for layer  $k$ , and  $\sigma$  is the activation function (typically ReLU). This formula shows that each user's embedding is updated by aggregating information from its neighbours in the graph, normalised by node degrees to balance contributions from different nodes.

By stacking multiple layers, NGCF captures high-order connectivity in the graph, enriching the user and item representations with information from multiple hops. This iterative aggregation process enables NGCF to learn latent relationships between users and items, enhancing recommendation accuracy through deeper interaction modelling across the user-item graph.

Building on the success of NGCF, LightGCN (Light Graph Convolutional Network) was introduced to simplify the model architecture while retaining the benefits of GNN-based embedding propagation. LightGCN removes non-linear activation functions and feature transformation layers, resulting in a more lightweight and efficient model. This simplification not only reduces computational complexity but also enhances performance by focusing solely on the message-passing mechanism, which aggregates information from multi-hop neighbours. The core of LightGCN's approach lies in its propagation rule, which updates the embeddings of users and items through each layer as follows:

$$e_u^{(k+1)} = \sum_{i \in \mathcal{N}(u)} \frac{1}{\sqrt{|\mathcal{N}(u)| \cdot |\mathcal{N}(i)|}} e_i^{(k)}, \quad (2.8)$$

where  $e_u^{(k+1)}$  denotes the embedding of user  $u$  at the  $(k+1)$ -th layer, and  $\mathcal{N}(u)$  represents the neighbours of user  $u$  (i.e., the items interacted with by  $u$ ). This equation shows that LightGCN aggregates embeddings from neighbouring nodes, normalised by the degree of each node, to ensure balanced information propagation. By iteratively applying this aggregation, LightGCN captures multi-hop neighbour information, enhancing the embedding representations of users and items.

LightGCN has become a widely adopted baseline due to its efficiency and strong performance across various recommendation benchmarks (X. He et al., 2020). The model's streamlined design—focused exclusively on message-passing—makes it computationally less intensive and easier to scale, particularly in large datasets where traditional GNNs may struggle.

More recently, SimGCL (Simple Graph Contrastive Learning) has introduced contrastive learning techniques into the graph-based recommendation domain. Unlike traditional graph-based models that focus solely on collaborative signals, SimGCL applies contrastive learning to maximise agreement between different augmentations of the user-item graph (J. Yu et al., 2022). This approach helps to regularise the embeddings, improving robustness to noise and sparsity in the interaction data. By learning more distinct and well-separated embeddings, SimGCL addresses challenges related to data sparsity and provides more diverse recommendations.

The main objective of SimGCL is to enhance representation learning by pulling together embeddings of similar items and pushing apart embeddings of dissimilar items. The contrastive learning objective in SimGCL can be defined as follows:

$$\mathcal{L}_{\text{SimGCL}} = - \sum_{(u,v) \in \mathcal{P}} \log \frac{\exp(\text{sim}(e_u, e_v)/\tau)}{\sum_{(u,v') \in \mathcal{N}} \exp(\text{sim}(e_u, e_{v'})/\tau)}, \quad (2.9)$$

where  $\mathcal{L}_{\text{SimGCL}}$  is the contrastive loss,  $\mathcal{P}$  denotes the set of positive pairs (i.e., similar items or items connected by a user interaction), and  $\mathcal{N}$  represents the set of negative samples. The function  $\text{sim}(e_u, e_v)$  calculates the similarity between embeddings  $e_u$  and  $e_v$ , typically using cosine similarity. The temperature parameter  $\tau$  controls the concentration level of the distribution.

In SimGCL, augmentations of the graph are generated by randomly masking or perturbing edges in the user-item interaction graph. By training the model to maximise agreement between the representations of these augmented views, SimGCL learns embeddings that are more robust to noise and able to capture diverse recommendation patterns, thus mitigating data sparsity issues and enhancing recommendation diversity.

Graph-based models, particularly those based on GNNs, offer several advantages in recommendation tasks, such as their ability to capture high-order interactions and learn more robust user-item representations. However, these models can be computationally expensive, especially on large graphs, and may require careful tuning of hyperparameters and graph augmentations. Despite these challenges, graph-based models have proven to be effective in various recommendation scenarios and continue to be a significant area of research and development in collaborative filtering.

#### 2.2.4 Knowledge Graph-Based Recommendation

Knowledge graph-based recommender systems leverage structured information about entities and their relationships to enhance recommendation accuracy and provide meaningful

explanations. A knowledge graph represents items, users, and associated attributes as nodes, with edges denoting relationships. This structure allows KG-RS to capture semantic connections between entities that may not be apparent in traditional recommender systems. For instance, in a movie recommendation setting, a KG can incorporate relations such as "directed by," "acted in," or "similar genre," enabling the recommender system to make suggestions closely aligned with user preferences (Q. Guo et al., 2022).

Methods in KG-RS can be broadly divided into embedding-based, connection-based, and propagation-based approaches. Embedding-based methods learn vector representations of entities and relationships within the KG, effectively projecting them into a continuous vector space. Models like TransE and TransR are often used in this category, transforming entities and relations into low-dimensional vectors to facilitate similarity computations between users and items (Bordes et al., 2013; Y. Lin et al., 2015). Some embedding-based models follow a two-stage learning approach, where embeddings are first pre-trained using KG embedding algorithms and then fine-tuned with user-item interaction data, leading to improved recommendation performance.

Connection-based methods focus on the structural paths within the KG to model the connections between entities. Techniques such as meta-paths and meta-graphs define composite relations between items, users, and attributes, providing clear relational pathways in the KG that can be intuitively understood. For example, a meta-path like "user–watched–actor–acted in–movie" can highlight shared actors between movies a user has previously watched and potential recommendations, improving both the interpretability and relevance of recommendations.

Propagation-based methods use graph neural networks (GNNs) to aggregate information iteratively from neighbouring nodes in the KG, enabling the model to capture high-order connections between entities. This approach allows for a more comprehensive understanding of user preferences by propagating features across the graph. However, these models can be computationally demanding due to the need for neighbour sampling and attention mechanisms to handle the extensive relationships and nodes in large KGs (C. Zhao et al., 2019). Some propagation-based models incorporate attention layers to dynamically weigh the relevance of neighbours, refining the focus on significant relationships within the graph.

Knowledge graph-based recommendation offers distinct advantages. By integrating rich, structured information from the KG, these methods often achieve enhanced recommendation accuracy, better handling of cold-start and data sparsity issues, and improved explainability of recommendations. However, they also present challenges, including increased computational complexity and reliance on high-quality KG data, as propagation

across large graphs can be resource-intensive, and accurate, up-to-date KG data is critical for effective performance (Q. Guo et al., 2022).

Overall, KG-based methods have significantly advanced the field of recommender systems by enabling the incorporation of structured knowledge. These approaches open new possibilities for delivering highly relevant, contextually enriched, and interpretable recommendations across diverse domains, from e-commerce to social media.

### 2.2.5 LLM-based Recommendation

Large Language Models (LLMs) have recently gained significant attention in the field of recommender systems due to their advanced natural language understanding and generation capabilities. Traditional recommendation models, while effective at user-item interaction, often struggle to incorporate complex textual data and understand diverse user intents across different scenarios. LLMs, with their ability to process vast amounts of text and model intricate patterns in language, offer promising enhancements to recommender systems.

One primary approach for integrating LLMs in recommender systems is through prompting, where LLMs generate recommendations based on carefully designed prompts. For example, systems like Chat-Rec(Y. Gao et al., 2023) leverage ChatGPT to refine recommendation lists provided by traditional models, thereby enhancing both accuracy and explainability. Another approach involves using task-specific prompts that allow users to specify preferences in natural language, improving recommendation personalisation by catering to explicit user intents. This is evident in systems like P5(Geng et al., 2022), which use personalised prompt frameworks to handle various recommendation formats, enabling the model to generalise across multiple recommendation tasks with minimal retraining.

Fine-tuning is another powerful method for aligning LLMs with recommendation tasks. Fine-tuning allows the model to learn from task-specific data, enhancing its ability to predict relevant items by adjusting model parameters based on user-item interactions. Methods like TALLRec(Bao et al., 2023) and PALR(Yang et al., 2023) adopt fine-tuning strategies to improve performance on sequential recommendations, while LoRA(Hu et al., 2021) fine-tuning reduces computational overhead, making it feasible to deploy LLMs on resource-constrained devices.

Beyond direct recommendations, LLMs also contribute significantly to explanation generation and conversational recommendations. LLMs like ChatGPT can generate human-like explanations for their suggestions, as demonstrated in Chat-Rec, where LLMs provide clear reasoning for their recommendations, helping to increase user trust and satisfaction. Furthermore, conversational recommender systems, such as UniCRS, employ LLMs to

maintain dynamic user interactions, allowing users to refine their preferences through dialogue and receive updated recommendations accordingly.

Pre-training and prompt tuning are additional strategies explored in LLM-based recommender systems. Pre-training on extensive datasets with diverse content allows LLMs to generalise across different recommendation domains, from movies to e-commerce products. In particular, prompt tuning methods that adjust task-specific prompts enable LLMs to handle a variety of recommendation subtasks, such as top-k recommendation and rating prediction, without extensive retraining. This flexibility makes LLMs highly adaptable, allowing them to serve as powerful engines for next-generation recommender systems.

Overall, LLMs present a robust framework for addressing the evolving needs of recommendation tasks. By enabling deep contextual understanding, LLMs enhance recommendation accuracy, improve personalisation, and facilitate more engaging, interactive user experiences. The rapid advances in LLM capabilities suggest a promising future for LLM-based recommender systems, particularly as methods for integrating, fine-tuning, and prompting these models continue to mature.

### **2.2.6 Multimodal Recommendation**

Multimodal recommendation mitigates data sparsity and cold start in collaborative filtering (CF) by leveraging rich item-side content to strengthen item representations. Early work such as VBPR (R. He & McAuley, 2016b) adopts a straightforward strategy, augmenting user & item ID embeddings with visual and textual features to model preferences. With the advent of graph neural networks, GCNs have become central in CF-based recommenders. LightGCN (X. He et al., 2020), in particular, offers a clean and effective propagation design, removing feature transformations and nonlinear activations to focus on pure neighbourhood aggregation over the user-item graph. Building on this backbone, MMGCN (Wei et al., 2019) performs modality-aware message passing on the user-item graph, while DualGNN (Q. Wang et al., 2023) maintains two modality-specific GNNs and fuses them with adaptive weights to learn multimedia-aware representations. From BM3 (X. Zhou, Zhou, et al., 2023) onward, contrastive learning objectives have been used to inject self-supervision. Representative methods such as LATTICE (J. Zhang et al., 2021), FREEDOM (X. Zhou & Shen, 2023), PGL (P. Yu et al., 2025), GUME (G. Lin et al., 2024), and SMORE (Ong & Khong, 2025), etc (X. Guo et al., 2025), further strengthen multimodal modelling by exploiting richer item-side structure in conjunction with collaborative signals. Despite these advances, most approaches assume complete modality availability; coping with partially complete or modality missingness remains underexplored and is the focus of this work.

### 2.2.6.1 Modality Missingness in Multimodal Recommendation

Multimodal learning with missing modalities has attracted increasing attention, as modality absence is common in real-world multimodal systems (Ma et al., 2021). Given the importance of rich modality signals, simply discarding items with missing modalities wastes valuable supervision and may further exacerbate data sparsity, ultimately harming recommendation performance. Consequently, existing studies on missing modality-aware multimodal recommender systems (MMARec) generally fall into two complementary directions: *data preprocessing* and *missing-modality strategy design* (R. Wu et al., 2024).

**Preprocessing-Based Completion** Preprocessing-based completion attempts to construct complete inputs before training by imputing missing modalities, such as mean, zero filling or neighbourhood-based  $k$ -nn over observed item features. Although these approaches are simple and efficient, the resulting surrogate features can be biased when missingness is severe or structured. To better exploit collaborative signals, several graph-aware methods further propagate information on the user-item graph or content-derived item-item graphs to refine the filled features (Malitesta et al., 2024). However, because such imputation is often performed outside the recommender and then held fixed, the filled representations may not adapt to the evolving embedding space during training, leading to distribution mismatch. Moreover, repeatedly reusing imputed features can amplify hubness and over-smoothing effects, which may limit robustness under high missing ratios.

**Strategy Design** Strategy design develops model-level mechanisms to enhance robustness under incomplete modalities. This line of work can be broadly categorised into *representation learning* and *generation-based approaches*. Representation learning methods aim to learn modality-robust, modality-invariant, or modality-adaptive representations that remain effective even when one or more modalities are absent. MILK (Bai et al., 2024) pursues invariant preference modelling via cross-modality alignment and cyclic mixup. SiBraR (Ganhör et al., 2024) adopts a single-branch, weight-sharing encoder to project different available modalities into a shared embedding space. I<sup>3</sup>-MRec (H. Chen et al., 2025) further combines invariant risk minimisation across modality environments with an information-bottleneck-based fusion to learn compact and missing-aware representations. HIRE (Z. Lin et al., 2025) leverages unified heterogeneous hypergraph representation learning to capture higher-order cross-modal relations. MMAM (J. Chen et al., 2025) employs an adversarial generator-discriminator architecture to learn transferable, modality-robust features under uncertain missing-modality patterns.

In contrast, generation-based methods explicitly reconstruct missing modalities from observed ones, allowing the model to recover informative cross-modal cues and preserve item discriminability under incomplete inputs. CI2MG (Z. Lin et al., 2023) proposes a contrastive intra- and inter-modality generation framework that completes missing modalities by leveraging hypergraph-based cross-modal structure learning. DGMRec (Kim et al., 2025) disentangles modality-general and modality-specific representations and then synthesises missing views through cross-modal generators guided by collaborative signals. MoDiCF (J. Li et al., 2025) models modality-specific distributions with a diffusion-based generation module to produce high-quality substitutes for missing modalities.

In summary, preprocessing-based methods rely on static imputation that may introduce biased or hubness-amplified features, while strategy-design methods either pursue modality-robust representation learning or reconstruct missing views with dedicated generators. However, the distribution gap between complete and incomplete items is still insufficiently addressed in ranking-oriented MMRec settings. Motivated by this limitation, this work formulates missing-modality recommendation as a domain-gap problem and introduces an adversarial learning based alignment-and-completion framework under the modality missingness setting.

## 2.3 Chapter Summary

This chapter reviewed the key concepts and representative approaches in recommender systems. We first distinguished major recommendation tasks and clarified their objectives and evaluation targets. We then surveyed core families of general recommendation algorithms, ranging from content- and knowledge-driven methods to collaborative filtering, and further discussed matrix factorisation and graph-based models that constitute mainstream backbones in modern recommenders. In addition, we reviewed knowledge graph-based and LLM-based recommendation as complementary directions for incorporating structured knowledge and richer user intent. Finally, we summarised multimodal recommendation and conducted a focused review on modality missingness, organising existing methods into preprocessing-based completion and strategy-design approaches (including representation learning and generation-based solutions). The review highlights that real-world recommendation increasingly require robustness to non-ideal conditions—such as evolving behaviours, sparse interactions, and incomplete item content—which motivates the problem formulation and methodological developments presented in the following chapters.

## WAVELET-ENHANCED STATE-SPACE MODELS FOR SEQUENTIAL RECOMMENDATION

### 3.1 Introduction

As discussed in Chapter 2, sequential recommendation has become a central paradigm for modelling users' evolving preferences from interaction histories. Many existing sequential models rely on recurrent or attention-based architectures to capture temporal dependencies, yet they often become computationally expensive for long sequences or less effective when user behaviour exhibits time-varying and non-stationary patterns. In realistic platforms, interaction streams frequently contain multiple behavioural rhythms and shifts, where short-term intents and long-term preferences coexist and change over time. This motivates the need for a sequential framework that remains efficient for long histories while better characterising non-stationary dynamics. Accordingly, this chapter investigates frequency-aware sequential modelling and introduces a state-space-based architecture augmented with a wavelet frequency rescaling mechanism, aiming to capture long-range dependencies under evolving behavioural patterns.

Recommender systems tackle information overload by navigating large amounts of data and aligning them with user interactions (Lu et al., 2015; M. Mao et al., 2017; Q. Zhang, Lu, et al., 2019). Widely used in many applications and online services (Lu, 2003; Lu et al., 2010, 2013; M. Mao et al., 2019; Schafer et al., 1999), these systems aim to deliver personalised suggestions. However, as user preferences evolve, static models often fall short of capturing

the temporal dynamics of user behaviour. Sequential recommendation addresses this challenge by analysing the order and timing of user interactions, allowing future preferences to be predicted based on historical sequences (Tang & Wang, 2018).

The main goal of sequential recommendation is to predict the next item a user is likely to interact with, by considering both recent interactions and long-term behaviour (S. Wang et al., 2019). Sequential recommendation systems deliver more accurate, context-aware predictions by modelling the temporal dependencies in a user's interaction history. This makes them particularly effective in real-world environments, where user interests change over time due to trends, personal shifts, and external factors.

Traditional approaches, such as Markov chain models and sequential pattern mining, primarily focus on modelling item-to-item transitions (R. He & McAuley, 2016a; Rendle et al., 2010; Yap et al., 2012). Following these, latent representation models emerged (Lu et al., 2018a), learning latent representations of users and items to predict the next interaction, with models like GFF (Hidasi & Tikk, 2016) as notable examples. With the rise of deep neural networks, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) gained popularity. For instance, Tang et al. (Tang & Wang, 2018) introduced Caser, a CNN-based model, while many RNN-based models (Hidasi & Karatzoglou, 2018; Hidasi, Karatzoglou, et al., 2016) have been developed following the introduction of the gated recurrent unit (GRU). Transformer-based models, such as SASRec (Kang & McAuley, 2018) and BERT4Rec (Sun et al., 2019), have demonstrated the effectiveness of the self-attention mechanism in sequential recommendation tasks. However, deep learning-based recommendation models often suffer from the overshooting problem, where models become unstable during training. To address this, FMLPRec (K. Zhou et al., 2022) introduced a filter-enhanced MLP model that utilises a Fourier frequency filter to attenuate noise in the frequency domain. Building on this, BSARec (Shin et al., 2024) further improved FMLPRec by introducing an inductive bias that more effectively integrates both low- and high-frequency information. Recently, with advances in state space models, Mamba4Rec (C. Liu et al., 2024) leveraged the power of the Mamba architecture to tackle inference inefficiency, enhancing both speed and accuracy in sequential recommendation systems.

Despite the significant progress made in sequential recommendation, challenges remain in effectively handling long user-item interaction sequences, which often result in model inefficiencies and reduced accuracy when capturing long-term dependencies (S. Wang et al., 2019). Additionally, many models struggle to manage non-stationary data, where user preferences may change dynamically over time, further complicating the prediction task. While attention-based models with frequency filters offer one potential solution, to the best

of our knowledge, few works have explored the potential of using wavelet transforms, to address these challenges. Existing approaches, such as the Fourier transform-based methods in (Shin et al., 2024; K. Zhou et al., 2022), only focus on the frequency domain, but overlook time-domain localisation, which is critical for handling transient or non-stationary signals. In contrast, the wavelet transform provides a multi-resolution analysis, enabling both time and frequency domain insights. This motivates our exploration of integrating a wavelet frequency rescaler to better capture non-stationary user behaviour patterns in sequential recommendation tasks.

To address these challenges, we propose a Mamba-based model enhanced by a wavelet frequency rescaler. Specifically, our model is the first to integrate a discrete wavelet transform (DWT) to capture both low- and high-frequency patterns in user interactions, allowing it to better handle non-stationary behaviour. The combination of state-space models and wavelet frequency rescaling improves the model’s ability to efficiently process long interaction histories and deliver more accurate sequential recommendations. The contributions of this chapter are as follows:

- We propose the first integration of wavelet frequency rescaling with Mamba to better capture both high- and low-frequency patterns in non-stationary user interactions, enabling more accurate sequential recommendations.
- We demonstrate the effectiveness of the state-space model in overcoming computational inefficiencies common in self-attention-based approaches.
- Experiments on two public datasets demonstrate that our proposed model surpasses seven baseline methods, validating the effectiveness of the model for sequential recommendation.

The code for our model will be made available upon acceptance at GitHub repository.

## 3.2 Preliminaries

### 3.2.1 Problem Statement

Under the setting of sequential recommendation with implicit feedback, given a user’s previous interaction history, the goal is to predict the user’s next interaction with an item. Let  $U$  and  $V$  denote the sets of users and items, respectively, where  $u \in U$  is a user and  $v \in V$  is

an item. Each user has an interaction sequence in chronological order:  $S^u = [v_1^u, v_2^u, \dots, v_n^u]$ , where  $v_t^u$  represents the  $t$ -th item that user  $u$  interacted with, and  $n$  is the sequence length.

The task is to recommend a top- $k$  list of items the user is likely to interact with next, specifically predicting the item at the  $(n+1)$ -th step, denoted  $v_{n+1}^u$  for user  $u$ . The recommendation list  $\mathcal{L}_u = [v_{n+1}^{u,1}, v_{n+1}^{u,2}, \dots, v_{n+1}^{u,k}]$  is formed by selecting the top- $k$  items with the highest predicted probabilities:

$$\mathcal{L}_u = \text{Top-}k(P(v | S^u)) \quad \forall v \in V, \quad (3.1)$$

where  $P(v | S^u)$  represents the probability of user  $u$  interacting with item  $v$ , given their interaction history  $S^u$ . The objective is to maximise the accuracy of these recommendations, ensuring that the predicted top- $k$  list closely matches the user's actual future interactions.

### 3.2.2 Fourier Transform vs Wavelet Transform

The Fourier transform effectively analyses stationary signals by decomposing them into sinusoidal components to reveal frequency characteristics. However, it suffers from critical limitations: operating solely in the frequency domain makes it difficult to determine when specific frequencies occur. Furthermore, the Fourier transform struggles with signals containing abrupt changes or transients, such as sharp edges or spikes. This limitation leads to issues like the Gibbs effect, where representing discontinuities requires numerous sinusoidal components, resulting in slow convergence. Its efficiency also decreases for non-stationary signals, where patterns evolve over time. While the short-time Fourier transform attempts to address this with fixed windowing, selecting an optimal window size remains challenging.

In contrast, the wavelet transform provides multi-resolution analysis in both time and frequency domains, making it ideal for sequential recommendation tasks with non-stationary user interactions, thanks to its dual nature. The DWT used in our model dynamically adjusts resolution based on data needs, effectively capturing evolving user behaviour driven by trends and preferences.

## 3.3 Methodology

In this section, we outline the framework of our proposed sequential recommendation model. As shown in Figure 3.1, the methodology is structured around several core components that enhance the model's capability to capture complex temporal dynamics and user-item interactions.

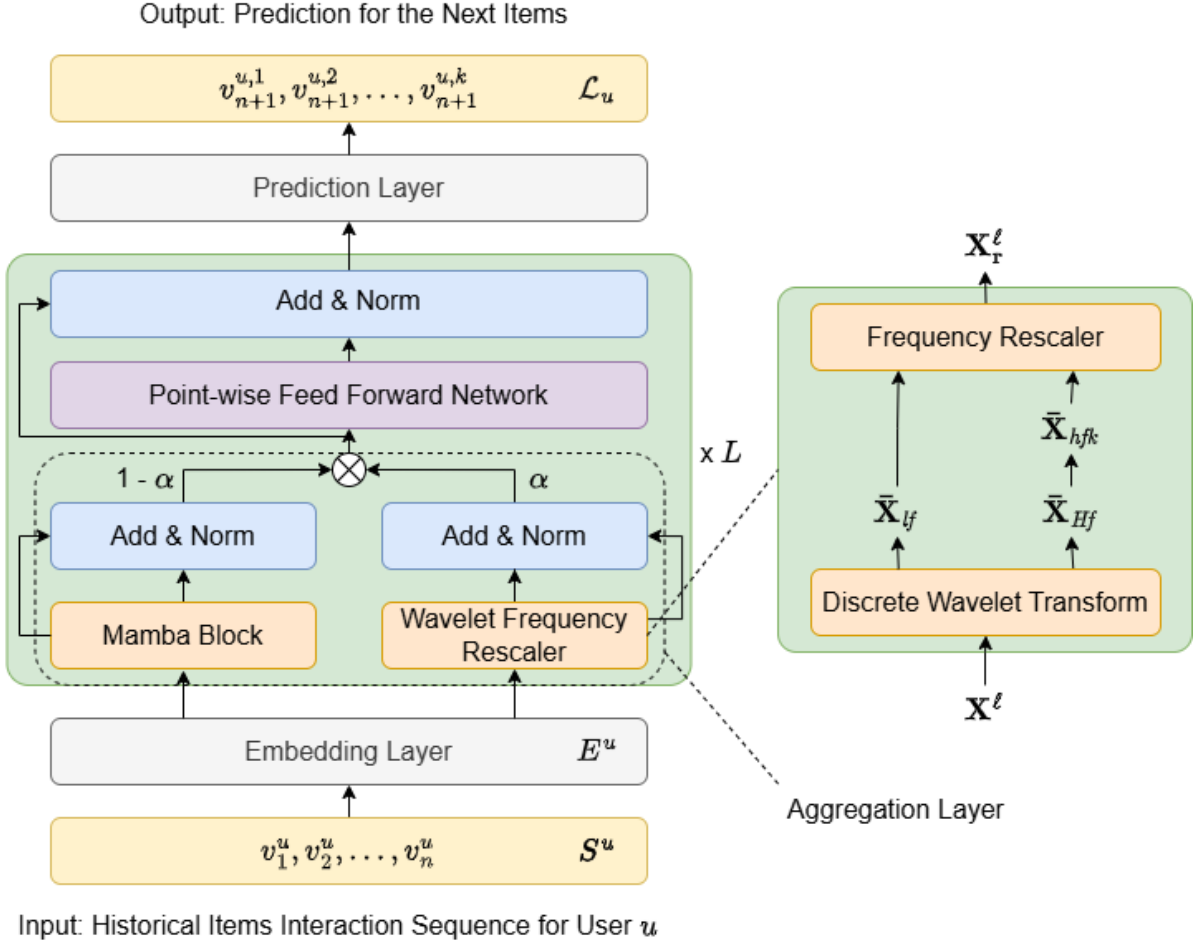


Figure 3.1: Overall Framework of Wavelet Mamba.

### 3.3.1 Embedding Layer

Given the interaction sequence  $S^u$  and a maximum sequence length  $N$ , we construct an item embedding matrix  $M_I \in \mathbb{R}^{|V| \times D}$ , where  $|V|$  is the size of the item set and  $D$  is the latent dimension. The embedding representation of the sequence is defined as  $E^u \in \mathbb{R}^{t \times D}$ , where  $t$  denotes the current time step. As indicated by (C. Liu et al., 2024), state space models process inputs through recurrent modules, which differentiates them from Transformer-based models by not requiring positional embeddings in the embedding layer. Since  $M_I$  is initialised randomly, we apply embedding dropout and layer normalisation to stabilise training while reducing the risk of overfitting:

$$E^u = \text{LayerNorm}(\text{Dropout}(M_I[S^u])). \quad (3.2)$$

During the training phase, the sequence is padded to a fixed length  $L$ , ensuring consistency across samples. We then combine multiple samples into mini-batches. The resulting

input data for the model is represented as  $\mathbf{X}^\ell \in \mathbb{R}^{B \times L \times D}$ , where  $B$  is the batch size,  $\ell$  is the number of layers. The first layer's input  $\mathbf{X}^0$  is initialised with  $E^u$ , ensuring the model starts with the correct representation of the sequence.

### 3.3.2 Wavelet Mamba Encoder

After extracting the embedding features, we propose a wavelet Mamba encoder to further process the input sequence embeddings  $E_u$ . The encoder consists of four main components: a wavelet frequency rescaler, a Mamba block, an aggregation layer, and a point-wise feed-forward network. Specifically, the wavelet frequency rescaler captures time-frequency characteristics, the Mamba block models long-term dependencies, the aggregation layer fuses their outputs, and the point-wise feed-forward network enhances the resulting sequence representation.

#### 3.3.2.1 Aggregation Layer

The aggregation layer integrates the outputs of the wavelet frequency rescaler and the Mamba block, enabling the model to capture both time-frequency information and long-term dependencies within sequential user interaction data.

The combination of the outputs from the wavelet frequency rescaler and the Mamba block follows the structure proposed in BSARec, where a linear combination of the two is used:

$$\hat{\mathbf{X}}^\ell = \alpha \cdot \mathbf{X}_w^\ell + (1 - \alpha) \cdot \mathbf{X}_o^\ell, \quad (3.3)$$

where  $\alpha$  is a hyperparameter that controls the balance between the wavelet frequency rescaler and the Mamba block. This weighted combination allows the model to dynamically balance the capturing of transient, high-frequency patterns and long-term temporal dependencies, thereby improving the model's ability to make accurate sequential recommendations.

**Wavelet Frequency Rescaler:** As shown in Algorithm 1, inspired by BSARec, we incorporate a wavelet frequency rescaler into our model to introduce an attentive inductive bias specifically designed for the sequential recommendation task. Rather than using a Discrete Fourier Transform, we employ the DWT, which is more effective for handling both low-frequency and high-frequency components in user interaction sequences. This approach allows our model to operate efficiently in both the time and frequency domains, making it particularly suitable for non-stationary signals often observed in user behaviours. In this study, a one-level DWT is used to separate the sequence representation into low-frequency

and high-frequency components, corresponding to stable preference trends and short-term behavioural variations, respectively.

---

**Algorithm 1** Wavelet Frequency Rescaler
 

---

- 1: **Input:**  $\mathbf{X}^\ell \in \mathbb{R}^{B \times L \times D}$
  - 2:  $F_{perm} \leftarrow \mathbf{X}^\ell$  according to (3.4) ▷ Permutes dimensions to apply DWT on  $L$
  - 3:  $\tilde{\mathbf{X}}_{lf}, \tilde{\mathbf{X}}_{Hf} \leftarrow \text{DWT}(F_{perm})$
  - 4:  $\tilde{\mathbf{X}}_{lf} \leftarrow \tilde{\mathbf{X}}_{lf}$  according to (3.4) ▷ Permutes back after DWT
  - 5:  $\tilde{\mathbf{X}}_{hfk} \leftarrow \tilde{\mathbf{X}}_{Hf}[-1]$
  - 6:  $\tilde{\mathbf{X}}_{hf} \leftarrow \tilde{\mathbf{X}}_{hfk}$  according to (3.4)
  - 7:  $\mathbf{X}_r^\ell \leftarrow \tilde{\mathbf{X}}_{lf}, \tilde{\mathbf{X}}_{hf}$  according to (3.5)
  - 8:  $\mathbf{X}_r^\ell \leftarrow \text{Padding}(\mathbf{X}_r^\ell)$
  - 9:  $\mathbf{X}_w^\ell \leftarrow \mathbf{X}_r^\ell, \mathbf{X}^\ell$  according to (3.7)
  - 10: **Output:**  $\mathbf{X}_w^\ell \in \mathbb{R}^{B \times L \times D}$
- 

The input,  $\mathbf{X}^\ell \in \mathbb{R}^{B \times L \times D}$ , is permuted along the sequence length dimension to obtain  $F_{perm}$ :

$$F_{perm} = \text{Permute}(\mathbf{X}^\ell, (0, 2, 1)), \quad (3.4)$$

where the permutation (0,2,1) swaps the second and third dimensions of the tensor, resulting in a shape of  $\mathbf{X}^\ell \in \mathbb{R}^{B \times D \times L}$ . This reordering is essential because DWT operates along the last dimension, which represents the sequence length  $L$ .

Once the DWT is applied, the sequence is decomposed into a low-pass component  $\tilde{\mathbf{X}}_{lf}$  that captures the smooth, long-term behaviour of the interaction sequence, and multiple high-pass components  $\tilde{\mathbf{X}}_{Hf}$  that capture the finer, transient changes.

Following the decomposition, the low-pass component  $\tilde{\mathbf{X}}_{lf}$  is retained, and the roughest high-pass component  $\tilde{\mathbf{X}}_{hfk}$  is selected from  $\tilde{\mathbf{X}}_{Hf}$  to match the shape of the low-pass component. We permute these two components back into  $\tilde{\mathbf{X}}_{lf}, \tilde{\mathbf{X}}_{hfk} \in \mathbb{R}^{B \times L \times D}$  before the frequency rescaler. This high-pass component is scaled by a trainable parameter  $\beta$ , enabling the model to control the influence of high-frequency, short-term patterns in the sequence:

$$\mathbf{X}_r^\ell = \tilde{\mathbf{X}}_{lf} + (\beta^2) \cdot \tilde{\mathbf{X}}_{hfk}. \quad (3.5)$$

The resulting combination of these components yields the rescaled sequence representation  $\mathbf{X}_r^\ell$  at  $\ell$  layer.

To maintain consistent  $L$  after applying the DWT, padding is applied where necessary, ensuring that the input and output shapes align for subsequent operations. This step is

crucial in preserving the overall temporal structure of the sequence, allowing the model to process sequences of uniform length across all layers.

Since the wavelet transform directly transforms the sequence from the time domain to the time-frequency domain, there is no need for an inverse wavelet transform. The model can leverage the extracted time-frequency features without reconstructing the original sequence.

To stabilise the model and avoid overfitting, the rescaled sequence undergoes dropout, layer normalisation and a residual connection, ensuring that both time-frequency features and the original sequence information are preserved. To simplify the notation, we define  $\mathcal{N}(\mathbf{X})$  as a combination of layer normalisation, dropout, and a residual connection:

$$\mathcal{N}(\mathbf{X}) = \text{LayerNorm}(\text{Dropout}(\mathbf{X}) + \mathbf{X}_i), \quad (3.6)$$

where  $\mathbf{X}_i$  is the input to the layer. The output is then computed as:

$$\mathbf{X}_w^\ell = \mathcal{N}(\mathbf{X}_r^\ell, \mathbf{X}^\ell). \quad (3.7)$$

**Mamba Block:** We utilise the Mamba block in Mamba4Rec as a core building component of our model to efficiently capture long-term dependencies in user interaction sequences. Layer normalisation, dropout, and a residual connection are applied:

$$\mathbf{X}_o^\ell = \mathcal{N}(\text{MambaBlock}(\mathbf{X}^\ell), \mathbf{X}^\ell). \quad (3.8)$$

The Mamba block leverages input-dependent selective mechanisms and the recurrent nature of state space models to efficiently capture complex temporal dependencies in the sequence data. It offers an effective balance between long-range sequence modelling and computational efficiency, operating in linear time with respect to the sequence length. The structured state matrix allows the Mamba block to outperform traditional self-attention mechanisms, particularly in modelling long user interaction sequences.

### 3.3.2.2 Point-wise Feed-Forward Network

Following the design of Transformer-based models, we utilise a point-wise feed-forward network (PFFN) to introduce nonlinearity and enhance the representation of user actions in the hidden dimension:

$$\tilde{\mathbf{X}}^\ell = \text{PFFN}(\hat{\mathbf{X}}^\ell) = \text{GELU}(\hat{\mathbf{X}}^\ell W_1 + b_1) W_2 + b_2, \quad (3.9)$$

where  $W_1 \in \mathbb{R}^{D \times 4D}$ ,  $W_2 \in \mathbb{R}^{4D \times D}$ ,  $b_1 \in \mathbb{R}^{4D}$ ,  $b_2 \in \mathbb{R}^D$  are learnable parameters. GELU activation is employed as common practice in PFFN. The dropout layer and layer normalisation

are employed to prevent overfitting. We also utilise the residual connection for the smooth propagation of features through multiple layers:

$$\mathbf{X}^{\ell+1} = \mathcal{N}(\tilde{\mathbf{X}}^\ell, \mathbf{X}^\ell). \quad (3.10)$$

### 3.3.3 Prediction Layer and Loss Function

In the final layer of our model, the prediction score for user preference is calculated by measuring the similarity between the sequence output and the target item embeddings:

$$\hat{y}_u = \mathbf{X}_{|S_u|}^L \cdot \mathbf{e}_v^\top, \quad (3.11)$$

where  $\mathbf{X}_{|S_u|}^L \in \mathbb{R}^{1 \times D}$  represents the user sequence output of the  $L$  layer blocks at step  $|S_u|$ , and  $\mathbf{e}_v \in \mathbb{R}^{1 \times D}$  is the embedding for the item  $v$ .

To optimise the model parameters, we employ a cross-entropy loss function for sequential recommendation. This loss is used to differentiate the correct item from the rest in the catalogue:

$$\mathcal{L}_{CE} = - \sum_{u=1}^B \log \frac{\exp(\hat{y}_u)}{\sum_{v=1}^{|V|} \exp(\hat{y}_v)}, \quad (3.12)$$

where  $B$  is the batch size, and  $|V|$  is the total number of items in the item set  $V$ . The softmax function normalises the scores across all candidate items, ensuring the probabilities sum to 1.

## 3.4 Experiment

### 3.4.1 Experimental Setup

#### 3.4.1.1 Datasets

We conduct our experiments on two publicly available benchmark datasets. The MovieLens-1M (ML-1M) and MovieLens-10M (ML-10M) are based on reviews of movies provided by MovieLens<sup>1</sup> (Harper & Konstan, 2015). They have long average interaction lengths, but different scales and sparsity. To focus on more informative data, we follow the common practice (Hidasi, Quadrana, et al., 2016) of filtering unpopular users and items with less than 5 interactions. The statistics of the processed datasets are shown in Table 3.1.

<sup>1</sup><https://grouplens.org/datasets/movielens/>

| Dataset | #Users | #Items | #Interactions | Avg.Length | Density |
|---------|--------|--------|---------------|------------|---------|
| ML-1M   | 6,041  | 3,417  | 999,611       | 165.47     | 4.84%   |
| ML-10M  | 69879  | 10197  | 999,8816      | 143.09     | 1.40%   |

Table 3.1: Statistics of the Experimental Data after Processing.

### 3.4.1.2 Evaluation Metrics

Following widely used evaluation metrics in related work (W. X. Zhao et al., 2020), we adopt the top- $k$  Hit Ratio ( $HR@k$ ), top- $k$  Normalised Discounted Cumulative Gain ( $NDCG@k$ ), and top- $k$  Mean Reciprocal Rank ( $MRR@k$ ) to measure the recommendation’s accuracy, in which  $k$  is set to 5, 10, and 20. A leave-one-out strategy is employed for each user’s item sequence. For each user, all but the last interaction are used to train. The last interaction is "left out" of the training data and used for evaluation, while the second-to-last interaction is used as a validation set. We rank the predictions over the entire item set without negative sampling following the advice in (Krichene & Rendle, 2020).

### 3.4.1.3 Baselines

To demonstrate the effectiveness of our model, we compare our proposed method with seven baseline methods:

- **Caser** (Tang & Wang, 2018) is a CNN-based sequential model that captures the temporal patterns at different levels using horizontal and vertical filters.
- **GRU4Rec** (Hidasi, Karatzoglou, et al., 2016) is an RNN-based sequential recommender that utilises Gated Recurrent Units (GRUs) to model user interaction sequences and capture dependencies over time.
- **SASRec** (Kang & McAuley, 2018) is a unidirectional Transformer-based sequential recommendation model that leverages self-attention mechanisms.
- **BERT4Rec** (Sun et al., 2019) is a bidirectional Transformer-based sequential recommender. It employs masked item prediction and Cloze objective loss to learn user preferences from both left and right contexts in the interaction sequences.
- **FMLPRec** (K. Zhou et al., 2022) is based on a MLP architecture. It incorporates a frequency filter mechanism, which serves as an inductive bias by attenuating noise in the frequency domain.

- **BSARec**(Shin et al., 2024) goes beyond self-attention by leveraging a learnable frequency rescaler to balance between high and low-frequency behavioural patterns.
- **Mamba4Rec**(C. Liu et al., 2024) leverages selective state-space models (SSMs) to efficiently model long-term dependencies, instead of traditional attention mechanisms.

#### 3.4.1.4 Implementation Details

We implement our model using the RecBole framework (L. Xu et al., 2023) in PyTorch, running on a single NVIDIA RTX A5500 GPU with 24GB of memory. The hyperparameter  $\alpha$  is tuned over the values {0.1, 0.3, 0.5, 0.7, 0.9}. The number of layers is set to two. For the DWT, we use the Daubechies 6 wavelet with a decomposition level of 1. The Daubechies 6 wavelet is selected because it provides a practical balance between compact support, smoothness, and local time-frequency representation. This is suitable for sequential recommendation, where user interaction sequences may contain both stable long-term preference trends and short-term behavioural fluctuations. The decomposition level is set to 1 to obtain a simple separation between low-frequency and high-frequency components while avoiding excessive fragmentation of relatively short and sparse recommendation sequences. Deeper decomposition may introduce additional noise and computational cost, whereas a one-level decomposition preserves the main temporal-frequency information in an efficient manner. The Mamba block’s configuration follows the settings from Mamba4Rec, including the state expansion factor, kernel size, and block expansion factor. We set the maximum sequence length to 50 and the hidden dimension size to 64. The model is trained using the Adam optimiser with a learning rate of 0.0005, a training batch size of 2048, and an evaluation batch size of 4096. For the baseline models, we use the recommended hyperparameters from their respective papers.

#### 3.4.2 Experimental Results

The results of our experiments are presented in Table 3.2. Our findings can be summarised as follows:

First, we observe that Transformer-based architectures such as SASRec and BERT4Rec generally outperform CNN and RNN-based models, such as Caser and GRU4Rec, across most metrics. This is likely due to the ability of Transformers to better capture long-range dependencies in sequential data through self-attention mechanisms, which are more effective than convolutional or recurrent operations.

| Datasets | Metric  | Caser  | GRU4Rec | SASRec | BERT4Rec | FMLPRec | BsaRec | Mamba4Rec | Our Model     | Improv. |
|----------|---------|--------|---------|--------|----------|---------|--------|-----------|---------------|---------|
| ML-1M    | HR@5    | 0.0927 | 0.1005  | 0.1374 | 0.1512   | 0.1316  | 0.1944 | 0.2199    | <b>0.2333</b> | +6.09%  |
|          | HR@10   | 0.1556 | 0.1657  | 0.2137 | 0.2346   | 0.2065  | 0.2757 | 0.3121    | <b>0.3268</b> | +4.71%  |
|          | HR@20   | 0.2488 | 0.2664  | 0.3245 | 0.3440   | 0.3137  | 0.3884 | 0.4159    | <b>0.4318</b> | +3.82%  |
|          | NDCG@5  | 0.0592 | 0.0619  | 0.0873 | 0.1021   | 0.0846  | 0.1306 | 0.1523    | <b>0.1620</b> | +6.37%  |
|          | NDCG@10 | 0.0795 | 0.0828  | 0.1116 | 0.1289   | 0.1087  | 0.1568 | 0.1822    | <b>0.1921</b> | +5.43%  |
|          | NDCG@20 | 0.1028 | 0.1081  | 0.1395 | 0.1564   | 0.1356  | 0.1851 | 0.2081    | <b>0.2187</b> | +5.09%  |
|          | MRR@5   | 0.0099 | 0.0487  | 0.0737 | 0.0770   | 0.0688  | 0.1027 | 0.1300    | <b>0.1385</b> | +6.54%  |
|          | MRR@10  | 0.0126 | 0.0569  | 0.0841 | 0.0881   | 0.0798  | 0.1135 | 0.1421    | <b>0.1509</b> | +6.19%  |
|          | MRR@20  | 0.0148 | 0.0635  | 0.0911 | 0.0954   | 0.0871  | 0.1206 | 0.1493    | <b>0.1582</b> | +5.96%  |
|          |         |        |         |        |          |         |        |           |               |         |
| ML-10M   | HR@5    | 0.0399 | 0.0474  | 0.0479 | 0.0933   | 0.0648  | 0.1456 | 0.1823    | <b>0.1952</b> | +7.08%  |
|          | HR@10   | 0.0739 | 0.0811  | 0.0876 | 0.1514   | 0.1088  | 0.2162 | 0.2575    | <b>0.2736</b> | +6.25%  |
|          | HR@20   | 0.1285 | 0.1348  | 0.1536 | 0.2328   | 0.1826  | 0.3071 | 0.3513    | <b>0.3701</b> | +5.35%  |
|          | NDCG@5  | 0.0241 | 0.0299  | 0.0294 | 0.0601   | 0.0399  | 0.0989 | 0.1259    | <b>0.1344</b> | +6.75%  |
|          | NDCG@10 | 0.0350 | 0.0408  | 0.0422 | 0.0788   | 0.0540  | 0.1197 | 0.1501    | <b>0.1597</b> | +6.40%  |
|          | NDCG@20 | 0.0487 | 0.0542  | 0.0587 | 0.0993   | 0.0725  | 0.1445 | 0.1738    | <b>0.1841</b> | +5.93%  |
|          | MRR@5   | 0.0190 | 0.0243  | 0.0234 | 0.0493   | 0.0318  | 0.0836 | 0.1074    | <b>0.1145</b> | +6.61%  |
|          | MRR@10  | 0.0234 | 0.0287  | 0.0286 | 0.0569   | 0.0376  | 0.0929 | 0.1173    | <b>0.1249</b> | +6.48%  |
|          | MRR@20  | 0.0271 | 0.0323  | 0.0331 | 0.0625   | 0.0426  | 0.0991 | 0.1238    | <b>0.1315</b> | +6.22%  |
|          |         |        |         |        |          |         |        |           |               |         |

Table 3.2: Performance Comparison Across Datasets.

Second, models that incorporate frequency-based mechanisms, such as FMLPRec and BSARec, demonstrate superior performance compared to pure Transformer-based models. This highlights the importance of incorporating inductive biases, such as frequency filtering, which can help models focus on informative signals while attenuating noise.

Third, Mamba4Rec, which leverages selective state-space models in place of traditional attention mechanisms, achieves the best performance among the baseline methods. The state-space model’s ability to balance the trade-off between long-term dependencies and computational efficiency allows it to manage long sequences effectively.

Finally, our proposed model surpasses all seven baseline methods, with clear improvements over Mamba4Rec. The integration of the wavelet frequency rescaler enhances the model’s ability to handle non-stationary data by adapting to both high- and low-frequency patterns, thus outperforming the state-space model. This result underscores the potential of combining state-space models with wavelet transforms for improving sequential recommendation performance, particularly for datasets with long interaction histories and complex, dynamic user behaviours.

### 3.4.3 Additional Studies

#### 3.4.3.1 Ablation Studies

| Model                      | HR@20  | NDCG@20 | MRR@20 |
|----------------------------|--------|---------|--------|
| Our Model                  | 0.4318 | 0.2187  | 0.1582 |
| Fourier Frequency Rescaler | 0.4260 | 0.2089  | 0.1474 |
| Self-attention Layer       | 0.3649 | 0.1759  | 0.1226 |
| w/o Frequency Rescaler     | 0.4159 | 0.2081  | 0.1493 |
| w/o Mamba Block            | 0.2725 | 0.1183  | 0.0693 |
| w/o Add & Norm             | 0.3871 | 0.1875  | 0.1309 |

Table 3.3: Ablation Study on ML-1M.

To better understand the contribution of each component in our proposed model, we perform an ablation study on the ML-1M dataset. The results are shown in Table 3.3. Specifically, we analyse the effects of removing and substituting the following components:

- **Wavelet Frequency Rescaler:** Replacing the wavelet frequency rescaler with the Fourier frequency rescaler or removing it altogether results in a notable drop in performance

across all evaluation metrics. This demonstrates the effectiveness of the wavelet frequency rescaler in capturing non-stationary patterns.

- **Mamba Block:** Replacing the Mamba block with standard self-attention layers or removing it altogether leads to a significant reduction in model performance. This highlights the advantage of selective state-space models in handling long-range dependencies more efficiently than traditional attention mechanisms.
- **Add & Norm Operation:** This operation is crucial for stable training and efficient gradient flow, as its removal leads to performance deterioration.

These findings confirm that each component—wavelet frequency rescaler, Mamba block, and Add & Norm—plays a vital role in capturing complex, long-range dependencies and enhancing overall performance.

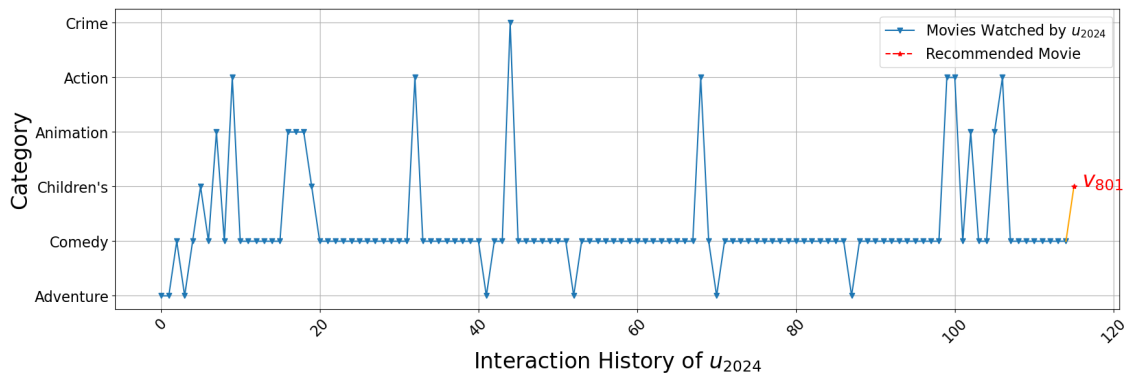


Figure 3.2: Case study on ML-1M for user  $u_{2024}$ , illustrating the model’s ability to recommend items aligned with the user’s preference for comedy while capturing related categories like children’s movies.

### 3.4.3.2 Case Study

In Figure 3.2, we analyse the recommendation outcomes for user  $u_{2024}$  from the ML-1M dataset, who has a long interaction history and a preference for comedy movies. This analysis illustrates how our model effectively captures both long-term preferences and subtle shifts in user behaviour.

Our model recommends item  $v_{801}$ , a children’s movie, which, while not strictly a comedy, shares thematic elements aligning with the user’s historical interests. This recommendation highlights the model’s ability to identify high-frequency patterns in user interactions, en-

abled by the wavelet frequency rescaler. In comparison, baseline models tend to recommend generic comedy movies, missing these nuanced preferences.

Furthermore, the Mamba block allows the model to maintain long-range dependencies efficiently, ensuring the user’s long-standing preference for comedy movies continues to influence recommendations. The combination of these components enables the model to generate personalised recommendations that adapt to both the stable and evolving aspects of user behaviour.

It is worth noting that the proposed model is primarily designed to capture long-term dependencies and moderate non-stationary variations in user behaviour. The wavelet frequency rescaler can help identify short-term fluctuations through high-frequency components, while the Mamba block preserves long-range preference signals. However, when a user’s preference changes drastically within a very short period, the model may still be influenced by historical interactions and may not fully adapt without additional recency-aware or online adaptation mechanisms. Therefore, handling abrupt preference shifts remains a challenging scenario and can be further explored in future work through time-aware evaluation protocols, adaptive forgetting mechanisms, or dynamic preference updating strategies.

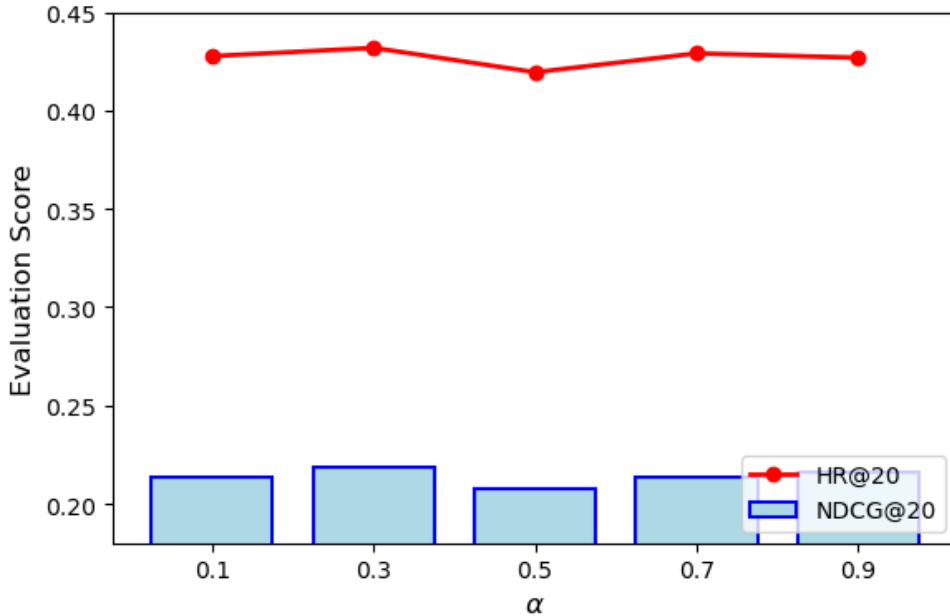


Figure 3.3: Sensitivity study on ML-1M. The horizontal axis represents different values of  $\alpha$ , and the vertical axis denotes the evaluation score of HR@20 and NDCG@20.

### 3.4.3.3 Sensitivity Study

In Figure 3.3, the sensitivity study on ML-1M shows that the model's performance remains stable across different values of  $\alpha$ , with the best performance at  $\alpha = 0.3$ .

## 3.5 Chapter Summary

In this chapter, we addressed the challenges of modelling long-range dependencies and non-stationary behaviour in sequential recommendation by introducing a novel approach that combines a wavelet frequency rescaler with a Mamba-based architecture. The proposed model captures both low- and high-frequency patterns in user interaction sequences, enabling it to model long-term preferences and short-term behavioural variations more efficiently. Experimental results on two benchmark datasets demonstrate that the model consistently outperforms seven baseline methods, including CNN-, RNN-, and Transformer-based models, highlighting the effectiveness of the proposed approach.

While the model shows promising results on datasets with long interaction histories, future work will focus on enhancing its generalisation to datasets with varying sparsity levels and shorter interaction sequences, such as those commonly found in e-commerce and music domains. Further optimisation of the wavelet rescaling technique may also yield additional improvements in recommendation accuracy. In addition, although the proposed model can capture long-term preferences and moderate behavioural variations, abrupt preference shifts within a short period may still require additional recency-aware adaptation or online updating mechanisms. Investigating such sudden preference changes under time-aware evaluation protocols is therefore a promising direction for future work.

## GRAPH-BASED RECOMMENDATION WITH SPARSE INTERACTIONS

### 4.1 Introduction

As introduced in the previous chapter, improving temporal modelling alone is often insufficient when user–item feedback is sparse, since limited observations restrict the reliability of purely sequence-based learning. Graph-based recommendation provides an alternative view by exploiting high-order collaborative signals through user–item connectivity, and many methods adopt neighbourhood aggregation to propagate information across the interaction graph. However, under sparse feedback, locally aggregated signals can be weak or biased, and purely local propagation may struggle to reflect broader item structure beyond immediate neighbourhoods. In practice, item relationships often exhibit global organisation (e.g., latent groups or communities) that is not explicitly captured by local message passing. This chapter, therefore, focuses on enriching item representations in graph-based recommendation under sparsity, and presents a framework that leverages globally informed relational structure to strengthen item embeddings and improve robustness in sparse regimes.

Recommender systems are crucial tools that assist users in obtaining preferred information by analysing their preferences. These systems have become a must in everyday life, spanning domains such as e-commerce, entertainment, education, and healthcare (Lu et al., 2015).

Collaborative filtering, as a foundational technique in recommender systems, operates by leveraging user-item interactions to uncover patterns and generate personalised suggestions. It can be broadly categorised into memory-based and model-based methods.

Model-based approaches employ matrix factorisation (Rendle et al., 2009), to learn latent factors representing users and items. However, traditional collaborative filtering methods often face challenges such as data sparsity and cold-start problems, limiting their performance in real-world applications (Q. Liu et al., 2024a).

Graph-based collaborative filtering has emerged as a powerful approach for recommender systems. By modelling user-item interactions as a bipartite graph, these methods leverage graph convolutional networks (GCNs) to propagate and aggregate collaborative signals across multiple layers (X. Wang et al., 2019). This structure effectively captures high-order user-item relationships, making GCN-based models well-suited for handling sparse interaction data. To improve efficiency, LR-GCCF (Lei et al., 2019) removed non-linearity, streamlining message propagation. LightGCN (X. He et al., 2020) further simplified GCNs by eliminating both non-linear activation functions and feature transformations, establishing itself as a strong baseline for graph-based collaborative filtering. UltraGCN (K. Mao et al., 2021) introduced a constraint loss, approximating infinite-layer propagation to address the over-smoothing issue in deep GCN models. Beyond architectural simplifications, recent models have incorporated contrastive learning to enhance representation quality. Methods like SGL (J. Wu et al., 2021) and SimGCL (J. Yu et al., 2022) introduce self-supervised contrastive loss to refine user and item embeddings, improving robustness and generalisation under sparse conditions.

Despite the significant progress made in the graph-based recommendation, challenges remain in effectively handling data sparsity. Existing GCN-based models rely solely on local neighbourhood aggregation, which may lead to suboptimal item embeddings, especially for items with limited interactions (Y. Zhang et al., 2024). Moreover, the lack of explicit global item structural modelling results in inadequate representations for items with sparse interactions, further limiting recommendation performance.

A way to mitigate these issues is to integrate Fuzzy C Means (FCM) clustering techniques with graph-based collaborative filtering. FCM clustering has been widely explored in graph-based learning tasks, such as community detection and graph partitioning, due to its ability to model overlapping structures and capture soft relationships among nodes (Lu et al., 2024; Song et al., 2020). However, despite its success in general graph learning, FCM clustering remains largely underexplored in graph-based recommender systems. By leveraging soft membership assignments, fuzzy clustering allows items to belong to multiple latent groups, preserving global structural relationships while providing more expressive item representations. Additionally, multi-view learning can further enhance item embeddings by fusing local graph-propagated signals with global cluster-level semantics, thereby

improving recommendation accuracy under sparse conditions.

However, existing methods fail to effectively unify local and global item representations, limiting the ability to leverage complementary information from both perspectives. To address these challenges, we propose a novel graph-based recommendation model, FuzzGCN, that integrates FCM clustering with gated multi-view fusion to enhance item representations, thereby boosting recommendation results. We incorporate FCM clustering, which enables items to belong to multiple latent groups, preserving global structural information while mitigating the limitations of traditional graph-based methods. Next, a gated fusion mechanism is introduced to dynamically integrate graph-based embeddings with FCM-enhanced item representations, ensuring an adaptive balance between local and global information. Finally, we leverage contrastive learning to align these multi-view representations, enforcing consistency between different item perspectives and enhancing the discriminative power of item embeddings. The contributions of this chapter are threefold:

- We propose a graph-based recommendation model that incorporates FCM clustering to enhance global item representations, effectively addressing data sparsity challenges.
- We design a gated fusion mechanism that dynamically integrates local graph-based embeddings with global clustering features, achieving a robust balance between local and global item semantics.
- We leverage contrastive learning to align multi-view embeddings, ensuring that fused representations retain complementary information from both global and local perspectives, thus boosting the model's discriminative power.

## 4.2 Methodology

In this section, we present the framework of FuzzGCN, which leverages GCN to address the challenges of data sparsity and effectively capture intricate user-item interaction patterns. As illustrated in Fig. 4.1, the model integrates several key modules designed to enhance its capacity for modelling complex user-item relationships and improving recommendation accuracy.

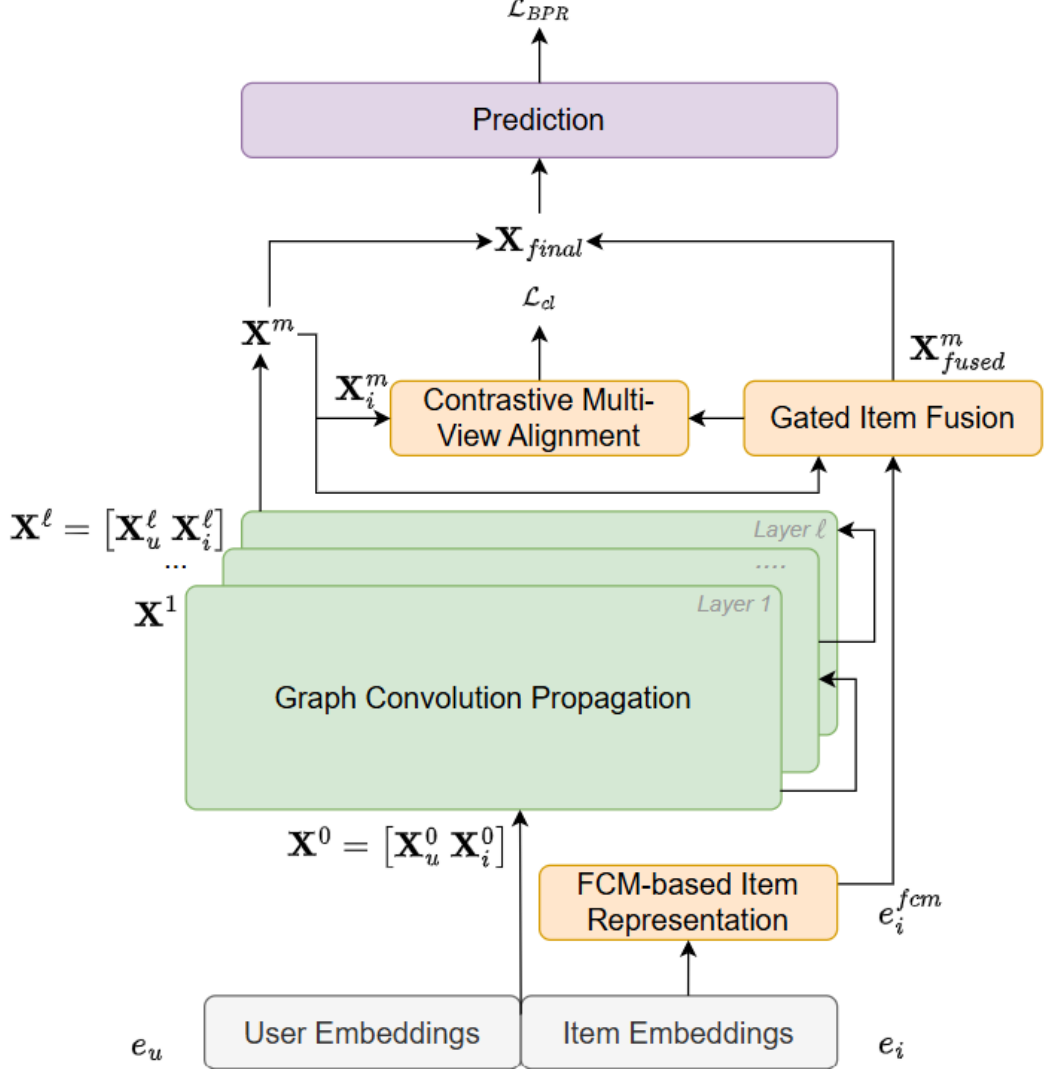


Figure 4.1: Overall framework of FuzzGCN.

## 4.2.1 Preliminaries

### 4.2.1.1 Problem Statement

Under the setting of recommendation with implicit feedback in collaborative filtering, let  $G = (V, E)$  represent a bipartite graph modelling user-item interactions, where the node set  $V$  is partitioned into users  $U$  and items  $I$ , with  $|U| = N_U$  and  $|I| = N_I$ , such that  $N = N_U + N_I$ , where  $N$  represents the total number of nodes in the graph. The edge set  $E$  corresponds to observed interactions between users and items, and the graph structure is encoded using an adjacency matrix  $A \in \mathbb{R}^{N \times N}$ , where  $A_{ij} = 1$  if there is an interaction between user  $i$  and item  $j$ . The diagonal degree matrix  $D$  captures the degree of each node.

Each node (user or item) is initialised with an embedding vector, forming the embedding matrix  $\mathbf{X}^0 \in \mathbb{R}^{N \times T}$ , where  $T$  is the embedding dimension. The goal is to learn refined user and item representations that accurately capture their interactions, enabling accurate ranking of items for each user in the recommendation task.

#### 4.2.1.2 LightGCN

LightGCN is a simplified version of GCN without feature transformation and non-linear activation (X. He et al., 2020). The graph convolutional layer in LightGCN is:

$$\mathbf{X}^{\ell+1} = (\mathbf{D}^{-1/2} \mathbf{A} \mathbf{D}^{-1/2}) \mathbf{X}^{\ell}, \quad (4.1)$$

where  $\mathbf{X}^{\ell}$  is the embedding matrix at layer  $\ell$ ,  $\mathbf{D}^{-1/2} \mathbf{A} \mathbf{D}^{-1/2}$  is the symmetric normalised adjacency matrix used for message propagation.

#### 4.2.1.3 Fuzzy C-Means Clustering

Fuzzy C-Means (FCM) is a soft clustering algorithm that assigns each data point a degree of membership to multiple clusters, rather than a single hard assignment (K. Li et al., 2023b). Given a dataset with feature representations, FCM optimises a clustering objective that minimises the weighted distance between data points and cluster centroids while allowing each point to belong to multiple clusters with different degrees of confidence. Unlike traditional clustering methods such as k-means, which enforce hard assignments, FCM allows for more flexible and nuanced representations, making it particularly useful in scenarios where data points exhibit latent multi-cluster affiliations (K. Li et al., 2023a).

### 4.2.2 Proposed Approach

#### 4.2.2.1 User-Item Embedding Initialisation

User and item embeddings serve as the foundation for learning meaningful representations in the recommendation task. We initialise these embeddings and construct the interaction graph at the very first stage.

**Graph Convolutional Embedding** Given the user-item bipartite graph  $G = (V, E)$ . The initial user and item embeddings are represented as:

$$\mathbf{X}^0 = \begin{bmatrix} \mathbf{X}_u^0 & \mathbf{X}_i^0 \end{bmatrix} \in \mathbb{R}^{N \times T}, \quad (4.2)$$

where  $\mathbf{X}_u^0$  and  $\mathbf{X}_i^0$  denote the initial embedding matrices for users and items, respectively. Each user embedding  $\mathbf{e}_u \in \mathbf{X}_U$  and each item embedding  $\mathbf{e}_i \in \mathbf{X}_I$  are learnable vectors of dimension  $T$ .

**FCM-based Item Representation** While graph convolutional embedding effectively captures local user-item interactions, it may fail to fully capture global item-level semantics, especially under sparse data conditions. To address this, we integrate Fuzzy C-Means (FCM) clustering into our framework.

The FCM algorithm assigns each item to a soft membership vector  $F \in \mathbb{R}^{N \times C}$ .  $C$  is the number of clusters. Every  $f \in F$  in this membership function is computed as:

$$f_{ij} = \frac{1}{\sum_{k=1}^C \left( \frac{\|\mathbf{e}_i - \mathbf{c}_k\|^2}{\|\mathbf{e}_i - \mathbf{c}_j\|^2} \right)^{\frac{1}{m-1}}} \quad (4.3)$$

where  $\mathbf{c}_j$  is the centroid of cluster  $j$ . The resulting soft membership matrix  $F$  ensures that items with similar characteristics are grouped together, even when interactions are sparse.

The learned FCM membership vectors are mapped to the embedding space using a fully connected layer with a ReLU activation:

$$\mathbf{e}_i^{fcm} = \text{ReLU}(\mathbf{F}\mathbf{W}_{fcm}) \quad (4.4)$$

where  $\mathbf{W}_{fcm} \in \mathbb{R}^{C \times T}$  is a learnable weight matrix, and  $\mathbf{e}_i^{fcm}$  represents the FCM-enhanced item embeddings.

#### 4.2.2.2 Global-Local Representation Enhancement

We integrate local collaborative signals captured through graph convolution with global structural information obtained from fuzzy clustering, ensuring that item representations effectively balance both perspectives for improved recommendation performance.

**Graph Convolution Propagation:** To effectively model high-order collaborative filtering signals, we adopt a GCN-based message-passing mechanism similar to LightGCN (X. He et al., 2020). The embedding update at each layer follows equation 4.1.

After  $\ell$  layers of propagation, the representations are obtained via a layer-wise mean aggregation:

$$\mathbf{X}^m = \frac{1}{\ell + 1} \sum_{\ell=0}^{\ell} \mathbf{X}^{\ell}. \quad (4.5)$$

Here, averaging over multiple layers mitigates over-smoothing while capturing higher-order interactions.  $\mathbf{X}^m$  represents the aggregated representation after  $\ell$  layers of propagation.

However, item embeddings matrix  $\mathbf{X}_i^m$  require additional refinement to incorporate global item-level semantics. Specifically, at each propagation step, we integrate information from the FCM-enhanced item embeddings  $e_i^{fcm}$  using a gated fusion mechanism shown below. This dynamic fusion ensures that item embeddings effectively balance local interaction signals and global structural information.

**FCM-based Item Fusion:** We employ a gated fusion mechanism to dynamically combine the FCM-enhanced item embeddings  $e_i^{fcm}$  with the graph-convolutional embeddings  $e_i$  after the final propagation step. First, we compute the fusion weights:

$$g = \sigma \left( \mathbf{W} \cdot \left[ \mathbf{X}_i^m \parallel \mathbf{e}_i^{fcm} \right] \right), \quad (4.6)$$

where  $\sigma(\cdot)$  represents the activation function, for which we choose the sigmoid.  $\mathbf{W} \in \mathbb{R}^{2T \times T}$  is a learnable weight matrix, and  $\left[ \mathbf{X}_i^m \parallel \mathbf{e}_i^{fcm} \right]$  denotes the concatenation of graph-based and FCM-based embeddings.

Then the fused item embeddings are calculated as:

$$\mathbf{X}_{fused}^m = g \odot \mathbf{X}_i^m + (1 - g) \odot \mathbf{e}_i^{fcm}, \quad (4.7)$$

where  $\odot$  denotes element-wise multiplication. This formulation ensures that the fusion process is adaptive to each item's characteristics, allowing the model to dynamically determine the importance of global clustering and local propagation for each item. Following the gated item fusion process, we obtain the final user and item representations,  $\mathbf{X}_{final} = \left[ \mathbf{X}_u^m \mathbf{X}_{fused}^m \right]$ , which encapsulate both local collaborative signals and global structural semantics.

**Contrastive Multi-View Alignment:** To maximise the agreement between the two views of the same item while minimising their similarity to embeddings from different items. We align the final aggregated representation  $\mathbf{X}_i^m$  with the fused representation  $\mathbf{X}_{fused,i}^m$  using an InfoNCE contrastive objective (van den Oord et al., 2019a).

$$\mathcal{L}_{cl} = -\frac{1}{N_I} \sum_{i \in I} \log \frac{\exp \left( \frac{\mathbf{x}_i^m \cdot \mathbf{x}_{fused,i}^m}{\tau} \right)}{\sum_{j \in I} \exp \left( \frac{\mathbf{x}_i^m \cdot \mathbf{x}_{fused,j}^m}{\tau} \right)}, \quad (4.8)$$

where  $\tau$  is the temperature scaling factor to control the distribution's sharpness. The numerator captures the similarity between the two views of the same item, while the denominator accounts for similarities with all items' FCM-based embeddings. This ensures that the two views are effectively aligned, reinforcing complementary strengths while maintaining discriminative power.

### 4.2.2.3 Prediction and Loss Functions

By incorporating graph convolutional propagation and FCM-based item fusion, we obtain the final representations of the user and item  $e_u, e_i^{fused}$ , which encode both local collaborative filtering signals and global structural relationships.

For a given user  $u$ , the predicted preference scores for all items are computed via the dot product:

$$\hat{y}_{u,i} = \mathbf{e}_u \cdot \mathbf{e}_i^{fused^\top}. \quad (4.9)$$

To optimise the ranking of items, we employ the Bayesian Personalised Ranking (BPR) loss, which maximises the difference between positive and negative items:

$$\mathcal{L}_{BPR} = -\frac{1}{|O|} \sum_{(u,i,j) \in O} \log \sigma(\hat{y}_{u,i} - \hat{y}_{u,j}) \quad (4.10)$$

where  $\sigma(\cdot)$  is the sigmoid function,  $\hat{y}_{u,i}$  and  $\hat{y}_{u,j}$  are the prediction scores for positive and negative items.  $O$  represents the set of observed interactions.

To prevent over-fitting, we apply an  $L_2$  penalty on the model. Finally, the total loss function is formulated as:

$$\mathcal{L} = \mathcal{L}_{BPR} + \lambda_1 \mathcal{L}_{cl} + \lambda_2 \|\Theta\|_2^2, \quad (4.11)$$

where  $\lambda_1$  regulates the contrastive loss, while  $\lambda_2$  regulates the L2 regularisation term.

## 4.3 Experiment

### 4.3.1 Experimental Setup

| Dataset | #Users  | #Items | #Interactions | Density |
|---------|---------|--------|---------------|---------|
| Arts    | 45,624  | 21,104 | 396,556       | 0.05%   |
| Food    | 115,144 | 39,688 | 1,025,169     | 0.03%   |
| Games   | 50,677  | 16,897 | 454,529       | 0.06%   |

Table 4.1: Processed Dataset Statistics.

#### 4.3.1.1 Datasets

To evaluate our proposed model, we conduct experiments on three public datasets: Amazon Arts, Crafts and Sewing (Arts), Amazon Grocery and Gourmet Food (Food), and Amazon Video Games (Games). These datasets were selected to represent a diverse range of scenarios

in terms of sparsity, user/item counts, interaction density, and domain, thereby enabling a comprehensive evaluation of FuzzGCN’s robustness and generalisability. All raw datasets are preprocessed to remove users and items with fewer than five interactions. Detailed statistics of the processed datasets are provided in Table 4.1. Each processed dataset is split into training, validation, and testing sets using a 7:1:2 ratio. To ensure that evaluation is performed only on entities with learned representations, cold-start users and items that do not appear in the training set are removed from the validation and testing sets. We adopt a pairwise learning approach during training, consistent with standard practices (Rendle & Freudenthaler, 2014).

#### 4.3.1.2 Evaluation Metrics

Following widely used evaluation metrics (Jadon & Patil, 2025), we adopt  $\text{Recall}@k$  and Normalised Discounted Cumulative Gain ( $\text{NDCG}@k$ ) to measure recommendation accuracy, where  $k$  is set to 10, 20, and 50.

#### 4.3.1.3 Baselines

We compare FuzzGCN with seven baseline models, which include GCN-based models and traditional models.

- **BPR** (Rendle et al., 2009) is a pairwise optimisation-based collaborative filtering method that optimises the ranking of implicit feedback data using a matrix factorisation approach.
- **MultiVAE** (Liang et al., 2018) is a variational autoencoder-based recommendation model that employs a probabilistic latent variable framework to model user-item interactions.
- **EHCF** (C. Chen et al., 2020) is designed to learn from multi-behaviour user feedback without relying on negative sampling.
- **NGCF** (X. Wang et al., 2019) explicitly models high-order collaborative signals by propagating user-item interactions through multiple layers of graph convolutions.
- **LR-GCCF** (Lei et al., 2019) removes non-linear transformations in NGCF, using only linear embedding propagation. It also introduces a residual preference learning mechanism for prediction.

- **LightGCN**(X. He et al., 2020) simplifies NGCF by removing feature transformations and non-linear activations, focusing solely on neighbourhood aggregation to improve efficiency and effectiveness.
- **UltraGCN**(K. Mao et al., 2021) further simplifies LightGCN. It entirely eliminates explicit message passing by directly approximating the infinite-layer limit of graph convolutions through a constraint loss function.

#### 4.3.1.4 Implementation and Hyper-parameter Details

For a fair comparison, we integrate FuzzGCN into ImRec (X. Zhou, Lin, et al., 2023) using PyTorch. All experiments run on a single NVIDIA RTX A5500 GPU with 24GB of memory with fixed embedding size 64, following previous works (X. He et al., 2020; X. Wang et al., 2019). We initialise the embedding parameters with the Xavier initialisation (Glorot & Bengio, 2010) and we choose Adam as the optimiser (Kingma & Ba, 2015). Number of graph convolution layers  $\ell$  is selected from  $\{2, 3, 4\}$ ; the regularisation weight  $\lambda$  is chosen from  $\{1e^{-2}, 1e^{-3}, 1e^{-4}, 1e^{-5}\}$ ; and the number of clusters  $C$  is selected from  $\{16, 32, 64\}$ . The weight for the contrastive loss is fixed at 0.01, the temperature  $\tau$  is fixed at 0.2. For a fair comparison, the baseline models are tuned according to the settings provided in their respective papers.

#### 4.3.2 Experimental Results

Table 4.2 provides a comprehensive comparison of FuzzGCN against various baselines. Among non-graph-based methods, EHCF outperforms BPR, likely due to its ability to model user-item interactions without relying on negative sampling. However, MultiVAE exhibits a performance decline as dataset sparsity increases, suggesting that its variational autoencoder framework struggles in low-density interaction settings.

In contrast, graph-based models consistently outperform traditional matrix factorisation approaches. Notably, GCNs that incorporate non-linear activation functions (e.g., NGCF) perform worse than their simplified counterparts, such as LR-GCCF, LightGCN, and UltraGCN, reinforcing prior findings that removing non-linearity can enhance information propagation. Furthermore, LightGCN surpasses LR-GCCF, demonstrating the advantages of eliminating feature transformations in GCN-based collaborative filtering.

Interestingly, UltraGCN achieves superior performance over LightGCN on two datasets, suggesting that its additional constraint loss and item-item regularisation contribute to improved ranking quality. However, an unexpected observation is that LightGCN outperforms UltraGCN on the Arts dataset, indicating that the benefits of UltraGCN’s constraints may

| Datasets | Metric    | BPR    | MultivAE | EHCF   | NGCF   | LR-GCCF | LightGCN      | UltraGCN      | FuzzGCN       | Improv. |
|----------|-----------|--------|----------|--------|--------|---------|---------------|---------------|---------------|---------|
| Arts     | Recall@10 | 0.0201 | 0.0171   | 0.0202 | 0.0205 | 0.0221  | <u>0.0231</u> | 0.0217        | <b>0.0239</b> | +4.71%  |
|          | Recall@20 | 0.0327 | 0.0268   | 0.0319 | 0.0342 | 0.0365  | <u>0.0371</u> | 0.0346        | <b>0.0396</b> | +3.82%  |
|          | Recall@50 | 0.0589 | 0.0503   | 0.0567 | 0.0623 | 0.0636  | <u>0.0663</u> | 0.0646        | <b>0.0702</b> | +3.82%  |
|          | NDCG@10   | 0.0137 | 0.0113   | 0.0136 | 0.0142 | 0.0151  | <u>0.0156</u> | 0.0156        | <b>0.0165</b> | +5.43%  |
|          | NDCG@20   | 0.0177 | 0.0145   | 0.0175 | 0.0186 | 0.0197  | <u>0.0201</u> | 0.0193        | <b>0.0216</b> | +5.09%  |
|          | NDCG@50   | 0.0245 | 0.0205   | 0.0240 | 0.0260 | 0.0268  | <u>0.0277</u> | 0.0269        | <b>0.0295</b> | +5.09%  |
| Food     | Recall@10 | 0.0138 | 0.0133   | 0.0158 | 0.0158 | 0.0172  | 0.0184        | <u>0.0187</u> | <b>0.0199</b> | +6.42%  |
|          | Recall@20 | 0.0222 | 0.0208   | 0.0243 | 0.0254 | 0.0277  | 0.0286        | <u>0.0287</u> | <b>0.0312</b> | +8.71%  |
|          | Recall@50 | 0.0390 | 0.0374   | 0.0416 | 0.0456 | 0.0478  | 0.0497        | <u>0.0501</u> | <b>0.0531</b> | +5.99%  |
|          | NDCG@10   | 0.0097 | 0.0092   | 0.0111 | 0.0102 | 0.0120  | 0.0125        | <u>0.0129</u> | <b>0.0135</b> | +4.65%  |
|          | NDCG@20   | 0.0124 | 0.0116   | 0.0137 | 0.0132 | 0.0154  | 0.0157        | <u>0.0161</u> | <b>0.0171</b> | +6.21%  |
|          | NDCG@50   | 0.0167 | 0.0159   | 0.0182 | 0.0185 | 0.0206  | 0.0211        | <u>0.0216</u> | <b>0.0227</b> | +5.09%  |
| Games    | Recall@10 | 0.0210 | 0.0238   | 0.0278 | 0.0254 | 0.0259  | 0.0275        | <u>0.0290</u> | <b>0.0303</b> | +4.48%  |
|          | Recall@20 | 0.0369 | 0.0376   | 0.0445 | 0.0425 | 0.0446  | 0.0461        | <u>0.0480</u> | <b>0.0494</b> | +2.92%  |
|          | Recall@50 | 0.0699 | 0.0718   | 0.0772 | 0.0825 | 0.0824  | 0.0841        | <u>0.0879</u> | <b>0.0891</b> | +1.36%  |
|          | NDCG@10   | 0.0135 | 0.0154   | 0.0175 | 0.0166 | 0.0171  | 0.0175        | <u>0.0191</u> | <b>0.0194</b> | +1.57%  |
|          | NDCG@20   | 0.0183 | 0.0196   | 0.0227 | 0.0217 | 0.0228  | 0.0231        | <u>0.0250</u> | <b>0.0253</b> | +1.20%  |
|          | NDCG@50   | 0.0265 | 0.0280   | 0.0308 | 0.0314 | 0.0320  | 0.0326        | <u>0.0348</u> | <b>0.0351</b> | +0.86%  |

Table 4.2: Overall Performance Comparison Across Datasets. The top-performing results for each dataset and metric are highlighted in boldface, while the second-best results are underlined.

vary across domains. This result highlights the nuanced impact of regularisation strategies in GCN-based recommendation models.

FuzzGCN consistently outperforms all baselines across all datasets and evaluation metrics. Notably, our approach achieves the highest Recall and NDCG scores, demonstrating the effectiveness of incorporating FCM-based item clustering and gated fusion mechanisms in enhancing item representations under sparse conditions. Specifically, the FCM clustering module captures global structure among items by assigning soft cluster memberships, allowing items with limited interactions to be better represented. Meanwhile, the gated fusion module dynamically integrates these global embeddings with graph-based representations, balancing local connectivity and global semantics. This interplay enables FuzzGCN to generate more expressive and robust item embeddings, leading to superior ranking performance even in sparse datasets. In practical terms, higher Recall means that more relevant items are included in users' top recommendation lists, while higher NDCG indicates that these relevant items are placed closer to the top ranks. Such improvements are important in sparse Amazon datasets, where many users and items have limited interaction records and local neighbourhood signals may be weak or unreliable. Therefore, the gains achieved by FuzzGCN suggest that the proposed model can deliver more useful and stable recommendations for long-tail or weakly connected items, rather than only improving numerical benchmark results.

### 4.3.3 Additional Studies

#### 4.3.3.1 Ablation Studies

To better understand the contribution of each key component in FuzzGCN, we conduct an ablation study on Arts by removing FCM-based item clustering, gated item fusion, and contrastive loss, respectively. The results are reported in Table 4.3.

- **FCM-based item Clustering:** Removing the FCM-based item clustering effectively reduces FuzzGCN to LightGCN, eliminating the ability to incorporate global item relationships. This highlights the crucial role of global clustering in enhancing item representations, particularly in sparse interaction scenarios.
- **Gated Item Fusion:** The removal of gated item fusion results in a slight decrease in performance. Although the performance drop is smaller compared to removing FCM, without the gating mechanism, the fusion ratio between FCM-enhanced and GCN-

based embeddings is fixed, requiring manual tuning of the weight. This rigid fusion scheme is less flexible and may not optimally adjust to different items and users.

- **Contrastive Multi-View Alignment:** Removing the contrastive loss leads to the most significant performance decline. This demonstrates that contrastive learning plays a crucial role in aligning the multi-view embeddings, ensuring that the FCM-enhanced and GCN-propagated representations reinforce each other rather than diverging.

| Model                                | Recall@20     | NDCG@20       |
|--------------------------------------|---------------|---------------|
| FuzzGCN                              | <b>0.0396</b> | <b>0.0216</b> |
| w/o FCM                              | 0.0376        | 0.0201        |
| w/o Gated Item Fusion                | 0.0394        | 0.0213        |
| w/o Contrastive Multi-View Alignment | 0.0306        | 0.0171        |

Table 4.3: Ablation Study on Arts.

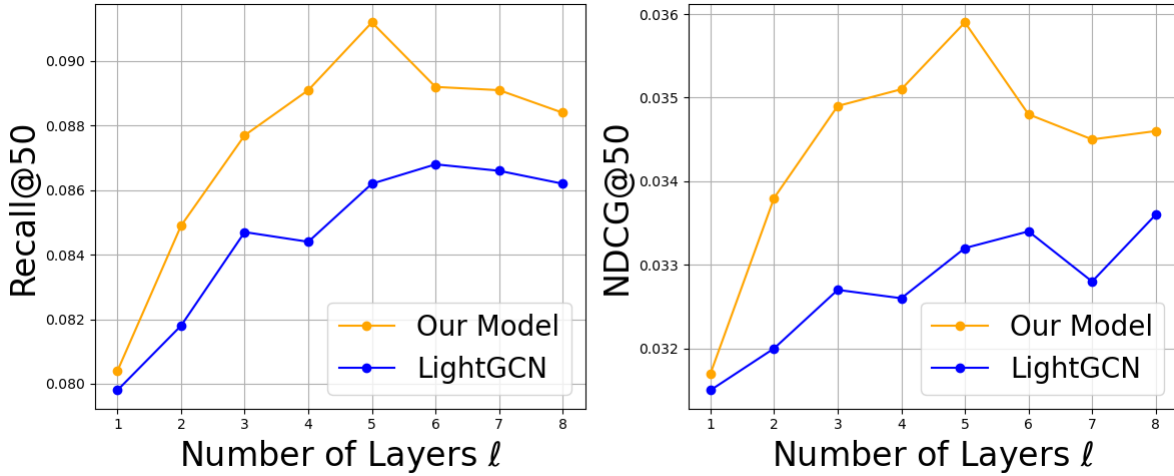


Figure 4.2: Effect of the number of layers  $\ell$  on FuzzGCN and LightGCN, evaluated on Games.

#### 4.3.3.2 Sensitivity Study

We analyse the impact of key hyperparameters on FuzzGCN’s performance, focusing on the number of graph convolutional layers  $\ell$  and FCM clusters  $C$ .

- **Effect of Layers:** As shown in Fig. 4.2, performance improves with increasing  $\ell$ , peaking at  $\ell = 6$ , before degrading due to over-smoothing. Compared to LightGCN, which

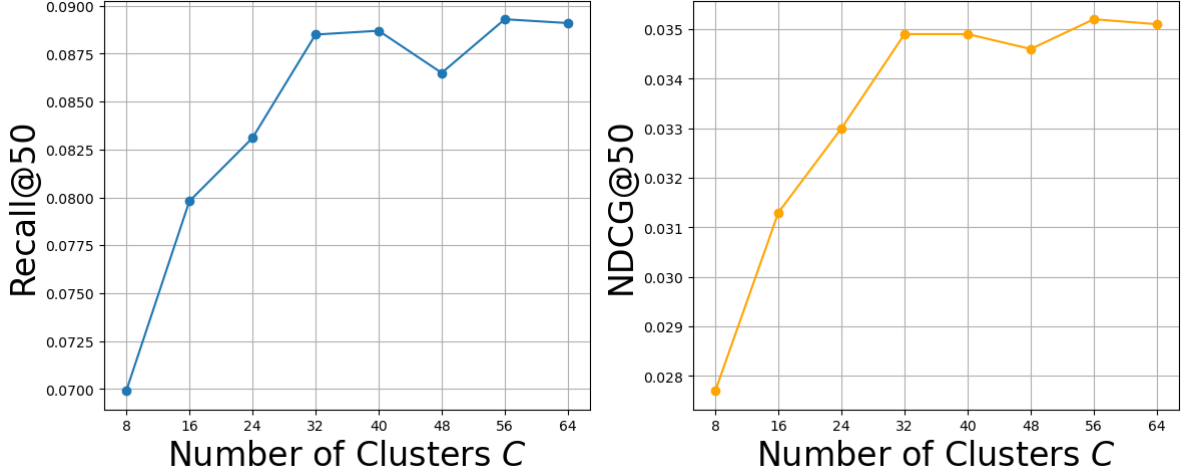


Figure 4.3: Effect of the number of clusters  $C$  on FuzzGCN, evaluated on Games.

peaks at  $\ell = 6$ , FuzzGCN benefits from deeper propagation, as fuzzy clustering mitigates early over-smoothing by incorporating global item semantics.

- **Effect of Clusters:** Fig. 4.3 shows that performance increases with more clusters, reaching an optimal  $C = 56$ , where global item structures are best captured. Beyond this, excessive clustering introduces noise and reduces generalisation, leading to slight performance degradation. Although  $C$  is fixed during each training run, it is treated as a tunable hyperparameter and analysed through the sensitivity study. The results indicate that the choice of  $C$  affects the balance between capturing global item structures and avoiding noisy or overly fragmented clusters. However, the current method does not adaptively determine the number of clusters during training. Future work may explore adaptive cluster selection strategies, such as validation-guided search, Bayesian optimisation, or split-and-merge fuzzy clustering, to automatically adjust  $C$  according to dataset characteristics.

## 4.4 Chapter Summary

In this chapter, we introduced FuzzGCN, an enhanced graph-based recommendation model that integrates Fuzzy C-Means (FCM) clustering and a gated multi-view fusion mechanism to address the challenges of data sparsity in collaborative filtering. By incorporating global item clustering, FuzzGCN enriches item representations beyond raw interaction signals, while the gated fusion mechanism adaptively balances the contribution of local and global

information. Furthermore, we employ contrastive learning to align the multi-view embeddings, reinforcing their consistency and improving overall representation quality. Extensive experiments demonstrate that FuzzGCN consistently outperforms competitive baselines, validating its effectiveness in sparse recommendation settings. While our current framework focuses on static user-item interactions, extending FuzzGCN to sequential recommendation tasks could enhance its applicability to time-sensitive and dynamic recommendation scenarios. Additionally, exploring more advanced fuzzy set formulations may further refine the model's adaptability to diverse user behaviours. A potential limitation of FuzzGCN is that FCM clustering may become computationally expensive when the number of items is extremely large. This is because the clustering cost increases with the number of items, clusters, embedding dimension, and optimisation iterations. In large-scale applications, this issue can be mitigated through mini-batch FCM, approximate or sampling-based clustering, offline pre-clustering, and periodic cluster updates rather than frequent re-clustering. Future work may also explore adaptive cluster selection to balance recommendation accuracy and computational efficiency.



## ROBUST MULTIMODAL RECOMMENDATION WITH INCOMPLETE ITEM CONTENT

### 5.1 Introduction

As discussed in Chapters 2, both sequential and graph-based recommendations largely rely on interaction signals, and performance can degrade when feedback is sparse or behaviour evolves over time. Modern recommender systems increasingly incorporate multimodal item content (e.g., text and images) to enrich representations and alleviate sparsity and cold-start; nevertheless, real-world content is often incomplete. Missing modalities not only reduce the available semantic signals but can also induce systematic differences between complete and incomplete items, making multimodal fusion unreliable and leading to biased representations. Moreover, modality-specific noise and entangled signals further complicate the learning of dependable multimodal features. Motivated by these challenges, this chapter investigates robust multimodal recommendation under incomplete modalities and develops a framework that maintains reliable representation learning and fusion when item content is partially unavailable.

Recommender systems have been widely adopted as a means of filtering content and delivering personalised suggestions based on users' preferences (C. Gao et al., 2024; S. Wu et al., 2022). Among different paradigms, collaborative filtering (CF) remains one of the most popular approaches (Lu et al., 2018b). Despite its success, CF suffers from well-known limitations such as data sparsity and the cold-start problem (Q. Zhang et al., 2023). CF

fundamentally relies on user–item interactions; however, in real-world scenarios, these interactions are inherently sparse. For example, in an online shopping platform, a single user can only interact with a tiny fraction of all available items, and each item is exposed to only a limited subset of users. This scarcity of interactions leads to the data sparsity problem. The cold-start problem emerges as an extreme case of sparsity: when a new user or item is introduced into the system with little or no interaction history, the recommender lacks sufficient information to provide accurate suggestions. These two long-standing challenges have motivated extensive research on enhanced recommendation settings and models (Zeng et al., 2025; Q. Zhang, Zhang, et al., 2019), among which multimodal recommendation has become one of the most popular and promising directions (Pomo et al., 2025).

Multimodal recommendation has emerged as a promising direction for alleviating data sparsity and cold-start issues by incorporating auxiliary information from multiple data modalities, such as textual, visual, and audio signals (Q. Liu et al., 2024b). The central idea is to emulate the human ability to integrate diverse sensory inputs in order to form a more complete understanding of items and user preferences. Multimodal data not only characterises item properties and user intent, but also implicitly reflects user–item relationships. Different modalities provide complementary views of item attributes, thereby enriching the available information and enabling more accurate preference modelling (H. Zhou et al., 2025). Building on this, the success of graph neural networks (GNNs) in pattern recognition and data mining has led to widespread adoption of graph-based techniques in MMRec (Z. Wu et al., 2021). Since user–item interactions naturally form a bipartite graph, spectral GNNs, particularly graph convolutional networks (GCNs) have demonstrated strong performance compared with traditional CF methods. Recent graph-based MMRec models (Z. Guo et al., 2024; P. Yu et al., 2025; J. Zhang et al., 2021) have achieved substantial performance gains on recommendation benchmarks.

Despite the strong performance of current models, a key limitation of most MMRec approaches is the underlying assumption that all items are associated with complete modality information. In practice, however, real-world data are often noisy and, more critically, may be partially or entirely missing for certain modalities. Consequently, the assumption of fully observed multimodal data rarely holds in realistic recommendation scenarios. As illustrated in Fig. 5.1(a), textual and visual modalities induce different but complementary item–item relations; missing one modality breaks part of this heterogeneous relational structure and can introduce bias into neighbourhood construction. Moreover, missing modalities may contaminate cross-modal relations. As shown in Fig. 5.1(b), the text embeddings of two items yield a relatively high cosine similarity because both descriptions contain the word

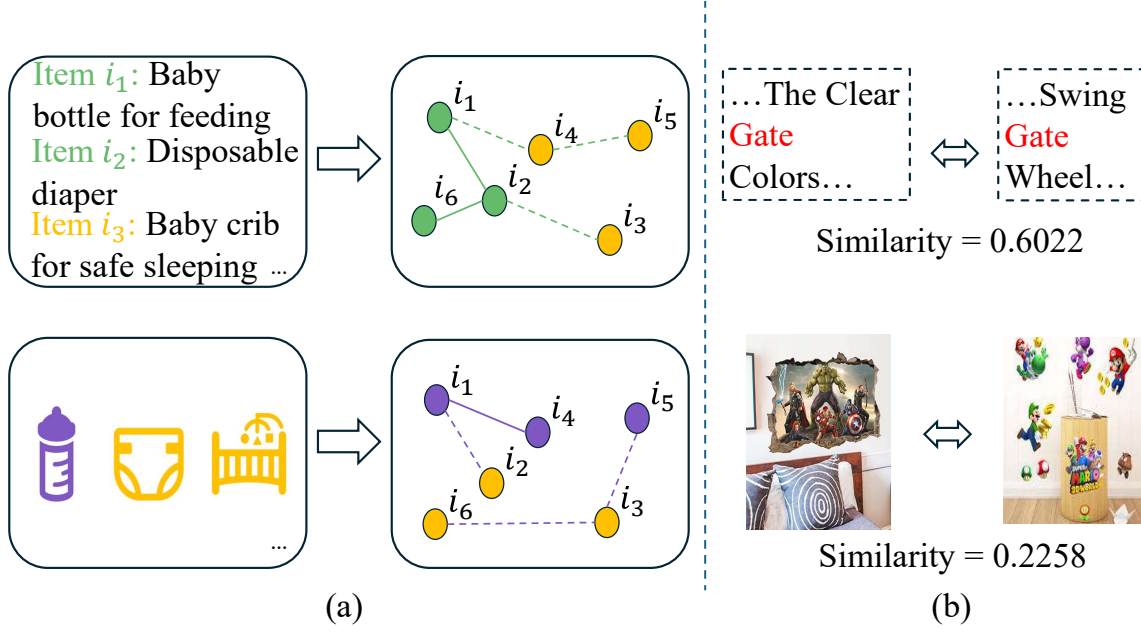


Figure 5.1: Illustration of missing modality and cross-modal inconsistency in multimodal recommendation: (a) item graphs built from text and images, where orange denotes items with missing modalities; (b) mismatched similarities across textual and visual modalities.

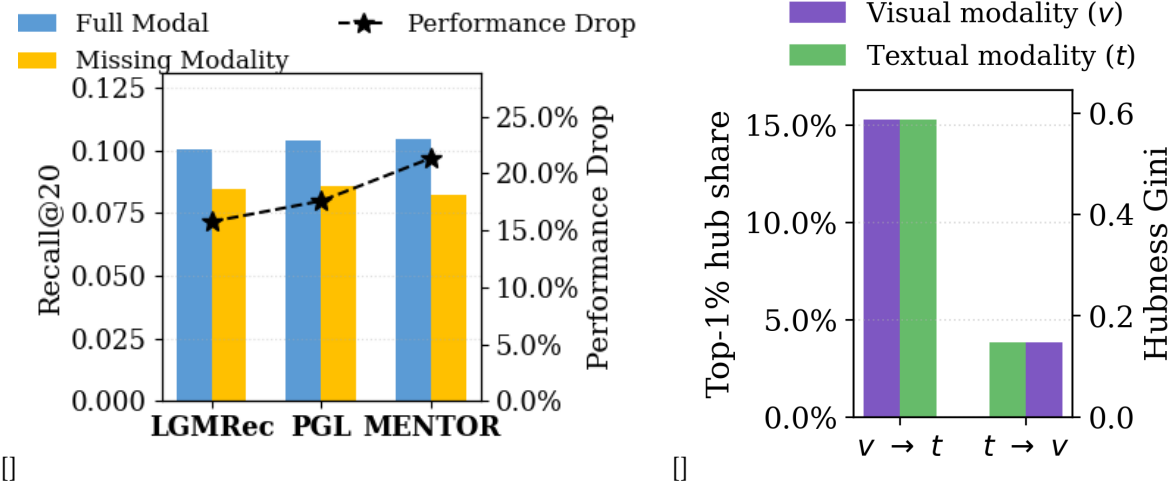


Figure 5.2: Empirical evidence of missing modality impact and cross-modal imputation bias: (a) MMRec baselines under full-modality and missing modality settings, with the corresponding performance drop. (b) Hubness analysis of cross-modal imputation, reporting the top-1% hub share and hubness Gini when imputing text features from visual features and imputing visual features from text features

“Gate”, although the items actually belong to different categories. In contrast, the images of two wallpaper products exhibit low visual similarity despite belonging to the same underlying category. These examples demonstrate that incomplete modalities can severely distort multimodal similarity and downstream recommendation.

A straightforward way to handle modality missingness is to discard samples with missing modalities during training. However, this strategy fails to exploit useful information and still cannot cope with missing modality items at test time (R. Wu et al., 2024). Imputation offers a more practical alternative, where missing entries are filled with surrogate values. Nevertheless, as shown in Fig. 5.2(a), these approaches still result in a notable degradation in recommendation quality.

To investigate the failure of  $k$ -nn imputation under missing modalities, we quantify hubness in the target-modality embedding space by counting how often each item appears as a neighbour across all  $k$ -nn queries. The neighbour-occurrence distribution is highly skewed (Fig. 5.2(b)), revealing hubs that attract a disproportionate share of queries. When a modality is missing, the query vector is uninformative, so within-modality  $k$ -nn repeatedly returns these hubs; averaging their features causes prototype collapse and over-smoothing, yielding near-identical imputations and eroding discriminative signal. The effect is amplified downstream because GCN-style propagation is a low-pass filter (Shen et al., 2021) that further homogenises representations, explaining the pronounced performance drop.

Motivated by this, we design a generation-based approach that reconstructs missing modalities while mitigating hubness and preserving item-level discriminability. Different from disentanglement pipelines such as DGMRec, we explicitly account for the completeness shift between complete and incomplete items: due to missing content and reconstruction noise, the invariant-feature distributions of the two domains can diverge, making generators trained on complete items prone to overfitting and biased completion on incomplete items. To address this, we incorporate curriculum domain-adversarial alignment with gradient reversal to bridge the complete-incomplete gap, and regularise the generation with a mutual-information bottleneck to decouple modality-invariant factors from modality-specific representations.

The contributions of this chapter are threefold:

1. We reformulate multimodal recommendation with missing modalities as a domain-gap problem between complete and incomplete items. To bridge this gap, we introduce a lightweight alignment scheme that jointly regularises cross-modal invariant embeddings under incomplete observations, providing a principled foundation for robust learning beyond naive imputation or sample dropping.

2. We propose a curriculum adversarial alignment scheme that performs domain alignment only in the early stage using a short-horizon gradient-reversal schedule, together with balanced sampling of items with and without missing modalities and a reconstruction-consistency objective. This design stabilises training compared to full-course adversarial optimisation while improving representation transferability under uncertain missingness.
3. We design a user-centric preference-aware fusion and alignment strategy. Specifically, we enhance CF user representations via an auxiliary user graph module and inject multimodal preference signals through user-conditioned filtering and multi-view user-item alignments. This yields more robust user-item representations and improves generalisation in item cold-start scenarios.

## 5.2 Preliminaries

### 5.2.1 Problem Statement

Under the setting of multimodal recommendation with implicit feedback (J. Chen et al., 2022), let  $G = (V, E)$  denote a bipartite graph modelling user-item interactions, where the node set  $V$  is partitioned into a user set  $U$  and an item set  $I$ , with  $|U| = N_U$  and  $|I| = N_I$ . Thus, the total number of nodes is  $N = N_U + N_I$ . Each item  $i \in I$  is associated with a textual feature  $\mathbf{x}_i^t \in \mathbb{R}^{R_t}$  and a visual feature  $\mathbf{x}_i^v \in \mathbb{R}^{R_v}$ . Let  $\mathbf{X}^t \in \mathbb{R}^{|I| \times R_t}$  and  $\mathbf{X}^v \in \mathbb{R}^{|I| \times R_v}$  denote the corresponding feature matrices. For notational convenience, we use  $m \in \{t, v\}$  to index modalities and write  $\mathbf{x}_i^m$  when the statement applies to both modalities.

The edge set  $E$  corresponds to observed interactions between users and items. We denote the user-item interaction matrix by  $\mathbf{R} \in \{0, 1\}^{|U| \times |I|}$ , where  $R_{ui} = 1$  if  $(u, i) \in E$  and 0 otherwise. We further construct the adjacency matrix  $A \in \{0, 1\}^{N \times N}$  of the bipartite graph from  $\mathbf{R}$ , and let  $D$  denote the diagonal degree matrix of  $A$ .

In the missing modality setting, each item may lose any subset of modalities. Let  $\mathcal{M}_i \subseteq \{t, v\}$  denote the set of observed modalities for item  $i$ . Accordingly,  $|\mathcal{M}_i| \in \{0, 1, 2\}$ , where  $|\mathcal{M}_i| = 2$  indicates a complete item and  $|\mathcal{M}_i| < 2$  indicates an item with missing modality information.

In the item cold-start setting, a subset of items  $I^{\text{new}} \subset I$  appears only in the test set and has no historical interactions in the training or validation sets.

The objective is to learn a recommender that can generate robust user and item representations under incomplete modalities and predict user preferences. Specifically, given a user

$u$  and an item  $i$ , we estimate a matching score  $\hat{y}_{ui}$  that reflects the likelihood of interaction, and train the model to rank observed items higher than unobserved ones for each user under implicit feedback.

## 5.3 Methodology

In this section, we present the proposed MAGEREC framework. As shown in Fig. 5.3, we follow a disentanglement pipeline (Kim et al., 2025), it consists of three interacting modules: a disentangled representation learning and preference-aware filtering module that separates modality-invariant from modality-dependent details and reweights them according to user preference signals; a missing modality completion module that reconstructs absent content via cross-modal generation and domain alignment; and a graph-based multimodal recommendation module that fuses collaborative and content information for final ranking tasks.

### 5.3.1 Modality Representation and Preference-Aware Filtering Module

Following prior studies (Kim et al., 2025; Xv et al., 2024; J. Zhang et al., 2021), we encode raw modality features into a unified latent space and disentangle them into modality-invariant and modality-specific components. We adopt two sets of encoders: (i) a shared encoder to extract modality-invariant signals and (ii) modality-specific encoders to preserve modality-unique cues.

**Modality-invariant and modality-specific encoders** For each modality  $m$ , we first project  $\mathbf{x}_i^m$  into a  $d$ -dimensional latent space and obtain the invariant component via a shared transformation:

$$\mathbf{z}_{i,m}^I = \sigma(\mathbf{W}_s \phi(\mathbf{W}_m \mathbf{x}_i^m)), \quad (5.1)$$

where  $\mathbf{W}_m$  is a modality-specific projection matrix,  $\mathbf{W}_s$  is shared across modalities,  $\phi(\cdot)$  is a nonlinear activation, and  $\sigma(\cdot)$  is the sigmoid function. In parallel, we extract modality-specific components using independent encoders:

$$\mathbf{z}_{i,m}^S = \sigma(\mathbf{W}_m^S \mathbf{x}_i^m), \quad (5.2)$$

**Preference-aware filtering** To adapt modality information to users' collaborative preferences, we learn modality-specific preference projections and derive item-wise filters from user ID embeddings. Let  $\mathbf{E}_U \in \mathbb{R}^{|U| \times d}$  denote user ID embeddings, and  $\mathbf{P}_m \in \mathbb{R}^{d \times d}$  be a

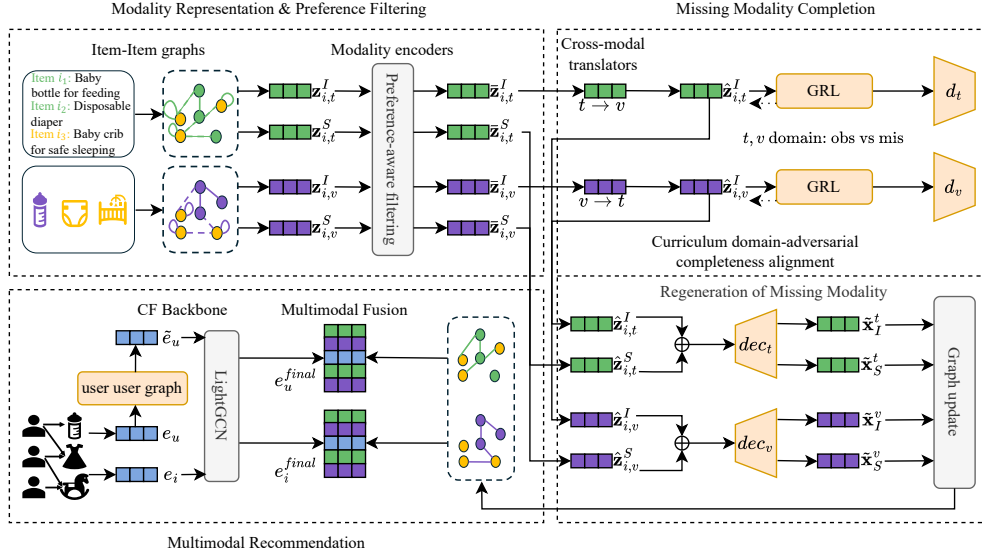


Figure 5.3: The overall framework of MAGERec. It comprises three modules: (i) *Modality Representation & Preference Filtering*, which encodes each modality into modality-invariant and modality-specific factors and reweights them via user-aware preference filters; (ii) *Missing Modality Completion*, which learns cross-modal translators  $t \rightarrow v$ ,  $v \rightarrow t$  and performs curriculum domain-adversarial completeness alignment with GRL-driven discriminators  $d_t, d_v$  to mitigate completeness shift, followed by regeneration of missing features and periodic modality graph refresh; and (iii) *Multimodal Recommendation*, which propagates collaborative signals enhanced by a user-user graph, aggregates modality-aware signals from modality-specific item-item graphs, and fuses them to produce final user-item representations  $\mathbf{e}_u^{\text{final}}, \mathbf{e}_i^{\text{final}}$  for preference prediction.

learnable projection for modality  $m$ . We compute modality-specific filters on the item side as

$$\mathbf{F}_m = D_I^{-1} \mathbf{R}^\top \tanh(\mathbf{E}_U \mathbf{P}_m), \quad (5.3)$$

where  $\mathbf{R} \in \{0, 1\}^{|U| \times |I|}$  is the interaction matrix and  $D_I$  is the diagonal degree matrix on the item side. We then apply element-wise filtering to both invariant and specific components:

$$\bar{\mathbf{z}}_{i,m}^I = \mathbf{f}_{i,m} \odot \mathbf{z}_{i,m}^I, \quad \bar{\mathbf{z}}_{i,m}^S = \mathbf{f}_{i,m} \odot \mathbf{z}_{i,m}^S, \quad (5.4)$$

where  $\mathbf{f}_{i,m}$  is the  $i$ -th row of  $\mathbf{F}_m$  and  $\odot$  denotes element-wise multiplication. This preference-aware filtering suppresses modality noise that is inconsistent with collaborative signals and produces user-adaptive modality representations for subsequent completion and recommendation.

**Item-Item graphs** For each modality, we build a sparse item-item graph from modality features by computing cosine similarities and retaining the  $k$  nearest neighbours for each item. Let  $\tilde{\mathbf{S}}_m \in \mathbb{R}^{|I| \times |I|}$  denote the resulting normalised propagation matrix.

### 5.3.2 Missing modality Completion Module

Based on the modality representations, we aim to recover missing modalities and obtain more reliable multimodal signals for recommendation. A straightforward strategy is to train cross-modal generators only on complete items and then apply them to incomplete ones. However, under practical missingness, complete and incomplete items exhibit a non-negligible completeness shift; their representation distributions differ due to missing content and reconstruction errors. If ignored, the learned generators may overfit to the complete domain and produce biased surrogate features for incomplete items.

**Cross-modal translators** Given the preference-filtered invariant representations  $\bar{\mathbf{z}}_{i,v}^I$  and  $\bar{\mathbf{z}}_{i,t}^I$ , we employ two cross-modal mappings to synthesise the counterpart modality:

$$\hat{\mathbf{z}}_{i,v}^I = g_{t \rightarrow v}(\bar{\mathbf{z}}_{i,t}^I), \quad \hat{\mathbf{z}}_{i,t}^I = g_{v \rightarrow t}(\bar{\mathbf{z}}_{i,v}^I), \quad (5.5)$$

where  $g_{t \rightarrow v}$  and  $g_{v \rightarrow t}$  are lightweight MLP generators. In addition, we generate modality-specific components from collaborative preference signals via two modality-specific generators:

$$\hat{\mathbf{z}}_{i,v}^S = h_v(\mathbf{f}_{i,v}), \quad \hat{\mathbf{z}}_{i,t}^S = h_t(\mathbf{f}_{i,t}), \quad (5.6)$$

where  $\mathbf{f}_{i,m}$  denotes the item-wise preference filter.

**Curriculum domain-adversarial completeness alignment** To explicitly mitigate the completeness shift, we treat items with observed modality  $m$  and items missing modality  $m$  as two *completeness domains* and adversarially align their modality-invariant representations. For each modality  $m \in \{t, v\}$ , we define the observed-domain item set  $I_m^{\text{obs}} = \{i \mid m \in \mathcal{M}_i\}$  and the missing-domain set  $I_m^{\text{mis}} = I \setminus I_m^{\text{obs}}$ . In each adversarial step, we construct a balanced mini-batch by sampling  $\mathcal{B}_m^{\text{obs}} \subset I_m^{\text{obs}}$  and  $\mathcal{B}_m^{\text{mis}} \subset I_m^{\text{mis}}$  half-and-half, and obtain the corresponding invariant features  $\mathbf{z}_{i,m}^I$  through the encoders.

To avoid degenerate alignment, we jointly optimise the cross-modal translators with Mean Squared Error (MSE) and InfoNCE loss (van den Oord et al., 2019b). Concretely, given  $\mathbf{z}_{i,t}^I$  and  $\mathbf{z}_{i,v}^I$ , we translate between modalities and enforce cycle-like reconstruction in the

invariant space:  $\hat{\mathbf{z}}_{i,v}^I = g_{t \rightarrow v}(\mathbf{z}_{i,t}^I)$  and  $\hat{\mathbf{z}}_{i,t}^I = g_{v \rightarrow t}(\mathbf{z}_{i,v}^I)$ . We minimise

$$\begin{aligned} \mathcal{L}_{\text{rec}} &= \mathcal{L}_{\text{MSE}}(g_{t \rightarrow v}(\mathbf{Z}_t^I), \mathbf{Z}_v^I) \\ &\quad + \mathcal{L}_{\text{MSE}}(g_{v \rightarrow t}(\mathbf{Z}_v^I), \mathbf{Z}_t^I) \\ \mathcal{L}_{\text{NCE}}(\mathbf{Z}_t^I, \mathbf{Z}_v^I) &= -\frac{1}{B} \sum_{n=1}^B \log \frac{\exp(\text{sim}(\mathbf{z}_{n,t}^I, \mathbf{z}_{n,v}^I) / \tau)}{\sum_{j=1}^B \exp(\text{sim}(\mathbf{z}_{n,t}^I, \mathbf{z}_{j,v}^I) / \tau)}, \end{aligned} \quad (5.7)$$

where  $\mathbf{Z}_t^I = \{\mathbf{z}_{n,t}^I\}_{n=1}^B$  and  $\mathbf{Z}_v^I = \{\mathbf{z}_{n,v}^I\}_{n=1}^B$  are the invariant features in a mini-batch, and  $\tau$  is a temperature<sup>1</sup>.

We attach a discriminator  $d_m(\cdot)$  to each modality-invariant representation and predict whether a feature comes from the observed or missing domain. Let  $y_{i,m} = 1$  if  $i \in I_m^{\text{obs}}$  and 0 otherwise. We adopt a gradient reversal layer (GRL) so that the discriminator is trained to separate the two domains, while the feature extractor is trained to confuse it:

$$\mathcal{L}_{\text{adv}} = \sum_{m \in \{t, v\}} \text{BCE}(d_m(\text{GRL}(\mathbf{z}_{i,m}^I; \lambda)), y_{i,m}), \quad (5.8)$$

where we employ a class-balanced BCE to account for possible imbalance between  $\mathcal{B}_m^{\text{obs}}$  and  $\mathcal{B}_m^{\text{mis}}$ .

We apply the adversarial pressure in a curriculum manner. Within each pre-epoch adversarial loop of  $S_{\text{DA}}$  steps, the GRL coefficient is smoothly increased via a sigmoid ramp-up,

$$\lambda = \lambda_{\text{max}} \left( \frac{2}{1 + \exp(-10p)} - 1 \right), \quad p = \frac{k}{S_{\text{DA}}}, \quad (5.9)$$

where  $k$  is the adversarial step index and  $p \in [0, 1]$  denotes the normalised progress within the current adversarial loop. Here,  $\lambda_{\text{max}}$  is the maximum GRL strength. Moreover, domain-adversarial updates are only activated in the early training stage, i.e., for the first  $S_{\text{max}}$  global adversarial steps. This curriculum stabilises optimisation by first learning reliable cross-modal mappings and then progressively strengthening domain confusion, leading to modality-invariant completion that generalises better under severe missingness.

We jointly optimise the alignment with

$$\mathcal{L}_{\text{DA}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{NCE}}(\mathbf{Z}_t^I, \mathbf{Z}_v^I) + \mathcal{L}_{\text{adv}}. \quad (5.10)$$

**Regeneration of missing modalities** After learning the aligned translators, we periodically regenerate missing modality features for incomplete items. Specifically, for each item  $i$ , we first obtain the invariant and specific components and form the reconstructed modality

<sup>1</sup> $\tau$  is omitted for brevity

representation by concatenation,  $\mathbf{r}_{i,m} = [\mathbf{z}_{i,m}^I \parallel \mathbf{z}_{i,m}^S]$ . Two lightweight decoders  $dec_m$  then map  $\mathbf{r}_{i,m}$  back to the raw feature space and produce regenerated features  $\tilde{\mathbf{x}}_i^m$ . For items with missing modality  $m$ , we replace the corresponding entry in  $\mathbf{X}^m$  with  $\tilde{\mathbf{x}}_i^m$ , thereby providing more informative inputs for subsequent multimodal propagation.

We supervise the regeneration with an MSE-based reconstruction objective

$$\mathcal{L}_{\text{recon}} = \sum_{m \in \{t, v\}} \mathcal{L}_{\text{MSE}}(\tilde{\mathbf{X}}^m, \mathbf{X}^m), \quad (5.11)$$

which encourages the regenerated features to remain faithful to the current modality embeddings.

We follow (Kim et al., 2025) to refresh the modality-specific item-item graphs every few epochs. To avoid abrupt changes, we update the graph with a momentum-style rule; the update is applied only to edges associated with items missing modality  $m$ :

$$\tilde{\mathbf{S}}_m = \alpha \hat{\mathbf{S}}_m + (1 - \alpha) \tilde{\mathbf{S}}_m, \quad (5.12)$$

where  $\alpha$  controls the update rate.

**Mutual-information regularisation** To prevent modality-invariant and modality-specific cues from being entangled during generation, we explicitly regularise their dependency using two complementary objectives. First, following the common practice in multimodal representation learning, we maximise cross-modal consistency by enforcing the invariant representations from different modalities to be close. Concretely, for complete items  $i \in I^{tv}$ , we apply an InfoNCE loss (van den Oord et al., 2019b) between  $\bar{\mathbf{z}}_{i,t}^I$  and  $\bar{\mathbf{z}}_{i,v}^I$ , which encourages the invariant factors to capture modality-agnostic semantics.

Second, inspired by (H. Chen et al., 2025; Kim et al., 2025), we minimise the statistical dependence between the invariant and specific components within each modality. We adopt a Contrastive Log-ratio Upper Bound (CLUB) estimator (Cheng et al., 2020) to upper bound the mutual information  $I(\bar{\mathbf{z}}_{i,m}^I; \bar{\mathbf{z}}_{i,m}^S)$  for  $m \in \{t, v\}$ . Specifically, we model a variational conditional distribution  $q_\psi(\bar{\mathbf{z}}_m^I | \bar{\mathbf{z}}_m^S) = \mathcal{N}(\mu_\psi(\bar{\mathbf{z}}_m^S), \text{diag}(\sigma_\psi^2(\bar{\mathbf{z}}_m^S)))$  and compute the CLUB upper bound as

$$\mathcal{L}_{\text{CLUB}}^m = \frac{1}{B} \sum_{n=1}^B \left[ \log q_\psi(\bar{\mathbf{z}}_{n,m}^I | \bar{\mathbf{z}}_{n,m}^S) - \log q_\psi(\bar{\mathbf{z}}_{\pi(n),m}^I | \bar{\mathbf{z}}_{n,m}^S) \right], \quad (5.13)$$

where  $\pi(\cdot)$  is a random permutation used to form negative pairs. We minimise  $\mathcal{L}_{\text{CLUB}}^m$  to suppress information leakage from specific factors into the invariant space.

$$\mathcal{L}_{\text{disent}} = \mathcal{L}_{\text{NCE}}(\bar{\mathbf{z}}_{i,t}^I, \bar{\mathbf{z}}_{i,v}^I) + \mathcal{L}_{\text{CLUB}}^t + \mathcal{L}_{\text{CLUB}}^v. \quad (5.14)$$

Together, these two regularisers yield cleaner invariant and specific factors, leading to more reliable modality completion and downstream recommendation.

### 5.3.3 Multimodal Recommendation Module

On top of the above content representations, we leverage the generated modality features to perform multimodal recommendation. However, the reconstructed representations can be less reliable than the ground-truth content and may introduce systematic bias, especially under severe missingness or cold-start, which can contaminate cross-modal relations and impair preference modelling.

**CF Backbone** Since modality missingness and completion errors are item-specific, further refining item embeddings may inadvertently propagate reconstruction noise through neighbourhood aggregation. To alleviate this issue, we anchor the recommender in collaborative signals and enhance user representations before multimodal fusion. Specifically, we construct an auxiliary user-user graph  $\mathbf{W} \in \mathbb{R}^{|U| \times |U|}$ , where the edge weight is defined by the number of commonly interacted items:

$$W_{uu'} = \begin{cases} |I_u \cap I_{u'}|, & u' \neq u, \\ 0, & u' = u. \end{cases} \quad (5.15)$$

We keep the top-200 neighbours for each user and compute a softmax-weighted aggregation over top- $k$  neighbours:

$$\alpha_{uu'} = \frac{\exp(W_{uu'})}{\sum_{u'' \in \mathcal{N}_u^{(k)}} \exp(W_{uu''})}, \quad u' \in \mathcal{N}_u^{(k)}, \mathcal{N}_u^{(k)} \subseteq \mathcal{N}_u^{(200)}. \quad (5.16)$$

The refined user ID embedding is then obtained via residual neighbourhood aggregation:

$$\tilde{\mathbf{e}}_u = \mathbf{e}_u + \gamma \sum_{u' \in \mathcal{N}_u^{(k)}} \alpha_{uu'} \mathbf{e}_{u'}, \quad (5.17)$$

where  $\gamma$  controls the strength of user-user refinement.

To capture high-order collaborative patterns between users and items, we adopt LightGCN-style message passing on the normalised adjacency matrix  $\tilde{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ . We initialise the node representations as

$$\mathbf{H}^{(0)} = \begin{bmatrix} \tilde{\mathbf{E}}_U \\ \mathbf{E}_I \end{bmatrix} \in \mathbb{R}^{(N_U + N_I) \times d}, \quad (5.18)$$

where  $\tilde{\mathbf{E}}_U$  stacks the refined user embeddings  $\tilde{\mathbf{e}}_u$ , and  $\mathbf{E}_I$  stacks the item ID embeddings  $\mathbf{e}_i$ . We then propagate representations for  $L$  layers:

$$\mathbf{H}^{(\ell+1)} = \tilde{A}\mathbf{H}^{(\ell)}, \quad \ell = 0, \dots, L_{ui} - 1. \quad (5.19)$$

The final collaborative embeddings are obtained by layer-wise mean pooling:

$$\mathbf{H} = \frac{1}{L_{ui} + 1} \sum_{\ell=0}^{L_{ui}} \mathbf{H}^{(\ell)}. \quad (5.20)$$

Finally, we split  $\mathbf{H}$  into user and item parts, yielding the collaborative representations  $\mathbf{e}_u^*$  and  $\mathbf{e}_i^*$ .

**Multimodal fusion with collaborative representations** After obtaining the collaborative representations  $\mathbf{e}_u^*$  and  $\mathbf{e}_i^*$  on the user-item graph, we further inject modality-aware signals derived from the modality-specific item-item graphs. For each modality, we perform graph propagation on  $\tilde{\mathbf{S}}_m$  using the preference-filtered invariant and specific item factors  $\tilde{\mathbf{z}}_{i,m}^I$  and  $\tilde{\mathbf{z}}_{i,m}^S$ , and denote the resulting item-side representations as  $\mathbf{i}_m^I$  and  $\mathbf{i}_m^S$ . We then transfer these item signals to the user side via the interaction matrix  $\mathbf{R}$ , yielding  $\mathbf{u}_m^I$  and  $\mathbf{u}_m^S$ . Finally, we integrate collaborative and multimodal information using a lightweight mean fusion:

$$\begin{aligned} \mathbf{e}_u^{\text{final}} &= \mathbf{e}_u^* + \frac{1}{3} \left( \frac{\mathbf{u}_v^I + \mathbf{u}_t^I}{2} + \mathbf{u}_v^S + \mathbf{u}_t^S \right), \\ \mathbf{e}_i^{\text{final}} &= \mathbf{e}_i^* + \frac{1}{3} \left( \frac{\mathbf{i}_v^I + \mathbf{i}_t^I}{2} + \mathbf{i}_v^S + \mathbf{i}_t^S \right). \end{aligned} \quad (5.21)$$

The preference score is then computed by inner product:

$$\hat{y}_{ui} = \left( \mathbf{e}_u^{\text{final}} \right)^\top \mathbf{e}_i^{\text{final}}. \quad (5.22)$$

**Alignment objectives** To make the fusion more reliable, we further impose a set of contrastive alignments that explicitly couple users with their interacted items and collaborative signals with modality-invariant content. We summarise the alignment loss as

$$\begin{aligned} \mathcal{L}_{\text{align}} &= \mathcal{L}_{\text{NCE}}(\mathbf{e}_u^*, \mathbf{e}_i^*) + \mathcal{L}_{\text{NCE}}(\mathbf{u}_t^I + \mathbf{u}_v^I, \mathbf{i}_t^I + \mathbf{i}_v^I) \\ &\quad + \mathcal{L}_{\text{NCE}}(\mathbf{u}_v^S, \mathbf{i}_v^S) + \mathcal{L}_{\text{NCE}}(\mathbf{u}_t^S, \mathbf{i}_t^S) \\ &\quad + \mathcal{L}_{\text{NCE}}(\mathbf{e}_i^*, \mathbf{i}_t^I + \mathbf{i}_v^I) + \mathcal{L}_{\text{NCE}}(\mathbf{e}_u^*, \mathbf{u}_t^I + \mathbf{u}_v^I). \end{aligned} \quad (5.23)$$

### 5.3.4 Optimisation

The objective is to jointly optimise the above three modules. We adopt the Bayesian Personalized Ranking (BPR) loss for implicit-feedback optimisation:

$$\mathcal{L}_{\text{BPR}} = -\frac{1}{|O|} \sum_{(u,i,j) \in O} \log \sigma(\hat{y}_{u,i} - \hat{y}_{u,j}), \quad (5.24)$$

where  $\sigma(\cdot)$  is the sigmoid function,  $\hat{y}_{u,i}$  and  $\hat{y}_{u,j}$  denote the prediction scores of a positive item  $i$  and a sampled negative item  $j$  for user  $u$ , respectively, and  $O$  is the set of sampled training triplets.

The overall objective of MAGERec is:

$$\begin{aligned} \mathcal{L} = & \mathcal{L}_{\text{BPR}} + \mathcal{L}_{\text{recon}} + \lambda_1 \mathcal{L}_{\text{disent}} + \mathcal{L}_{\text{gen}} + \lambda_2 \mathcal{L}_{\text{align}} \\ & + \lambda_3 \|\Theta_{id}\|_2^2 + \lambda_4 \left( \|\Theta_v\|_2^2 + \|\Theta_t\|_2^2 \right). \end{aligned} \quad (5.25)$$

$\lambda_1$  and  $\lambda_2$  are tradeoff parameters;  $\lambda_3$  and  $\lambda_4$  control the strength of regularisation to prevent overfitting, and we set  $\lambda_3 = 10^{-5}$  and  $\lambda_4 = 10^{-1}$ . The overall training procedure is summarised in Algorithm 2.

The additive form is adopted because MAGERec optimises several complementary objectives within a unified end-to-end framework. The BPR loss is the primary ranking objective for implicit-feedback recommendation, while the reconstruction and generation losses improve missing modality completion. The disentanglement loss separates modality-invariant and modality-specific factors, and the alignment loss encourages consistency between collaborative signals, user-item preference information, and multimodal content representations. These objectives are mutually related rather than independent stages, since the quality of generated modalities, disentangled representations, and aligned embeddings all affect the final recommendation scores. Therefore, combining them in an additive objective allows the shared representations to be updated jointly by ranking, completion, disentanglement, and alignment signals. The trade-off parameters balance the influence of auxiliary losses and regularisation terms, helping the model improve robustness under missing modalities without allowing any single objective to dominate the optimisation.

## 5.4 Experiment

In this section, we evaluate the proposed model on three real-world datasets under missing modality and new-item cold-start settings. We first describe the datasets, baselines, and experimental settings. We then present experiments that examine:

---

**Algorithm 2** Training procedure of MAGERec

---

**Input:** interaction matrix  $\mathbf{R}$ ; modality features  $\{\mathbf{X}^m\}_{m \in \{t,v\}}$ ; completeness-domain sets  $\{I_m^{\text{obs}}, I_m^{\text{mis}}\}_{m \in \{t,v\}}$ ; training configs.

**Output:** predicted preference scores  $\hat{y}_{ui}$  for ranking recommendation results.

```

1: Initialise  $\mathbf{E}_U, \mathbf{E}_I$  and model parameters.
2: Impute missing entries in  $\mathbf{X}^m$ .
3: Construct initial modality item–item propagation matrices  $\tilde{\mathbf{S}}_m$ .
4: for epoch  $e = 1, \dots, E$  do
5:   Encode  $\mathbf{z}_{i,m}^I$  and  $\mathbf{z}_{i,m}^S$  via Eq. (5.1) and Eq. (5.2).
6:   Compute preference filters  $\mathbf{F}_m$  via Eq. (5.3), and obtain  $\bar{\mathbf{z}}_{i,m}^I, \bar{\mathbf{z}}_{i,m}^S$  via Eq. (5.4).
7:   if global DA step  $\leq S_{\max}$  then
8:     for  $s = 1, \dots, S_{\text{DA}}$  do
9:       Sample balanced  $\mathcal{B}_m^{\text{obs}}, \mathcal{B}_m^{\text{mis}}$  from  $I_m^{\text{obs}}, I_m^{\text{mis}}$ .
10:      Translate invariants  $\hat{\mathbf{z}}_{i,v}^I, \hat{\mathbf{z}}_{i,t}^I$  via Eq. (5.5).
11:      Compute  $\mathcal{L}_{\text{rec}}$  and  $\mathcal{L}_{\text{NCE}}$  via Eq. (5.7).
12:      Compute  $\lambda$  via Eq. (5.9) and  $\mathcal{L}_{\text{adv}}$  via Eq. (5.8).
13:      Update  $\{\mathbf{W}_m, \mathbf{W}_s, g_{t \rightarrow v}, g_{v \rightarrow t}, d_t, d_v\}$  by minimising  $\mathcal{L}_{\text{DA}}$  via Eq. (5.10).
14:    end for
15:  end if
16:  Update CLUB estimators by minimising  $\mathcal{L}_{\text{CLUB}}^m$  via Eq. (5.13).
17:  Compute  $\mathcal{L}_{\text{disent}}$  via Eq. (5.14).
18:  Regenerate  $\tilde{\mathbf{x}}_i^m$  and compute  $\mathcal{L}_{\text{recon}}$  via Eq. (5.11).
19:  Replace missing entries in  $\mathbf{X}^m$  with regenerated features  $\tilde{\mathbf{x}}_i^m$ .
20:  if  $e \bmod E_{\text{ref}} = 0$  then
21:    Rebuild  $\hat{\mathbf{S}}_m$  and update  $\tilde{\mathbf{S}}_m$  via Eq. (5.12).
22:  end if
23:  Construct user–user graph weights  $W_{uu'}$  via Eq. (5.15) and compute  $\alpha_{uu'}$  via Eq. (5.16).
24:  Refine user embeddings  $\tilde{\mathbf{e}}_u$  via Eq. (5.17).
25:  Initialise  $\mathbf{H}^{(0)}$  via Eq. (5.18); propagate via Eq. (5.19); and mean-pool via Eq. (5.20) to obtain  $\mathbf{e}_u^*, \mathbf{e}_i^*$ .
26:  Fuse multimodal signals to obtain  $\mathbf{e}_u^{\text{final}}, \mathbf{e}_i^{\text{final}}$  via Eq. (5.21), and compute  $\hat{y}_{ui}$  via Eq. (5.22).
27:  Update model parameters by minimising  $\mathcal{L}$  via Eq. (5.25).
28: end for

```

---

- **Overall Performance:** performance under missing modality and item cold-start settings.
- **Ablation Study:** contributions of key modules via ablation studies.
- **Hyperparameter Sensitivity:** impacts of key hyperparameters.

- **Representation Visualisation:** the embedding space and learned modality-invariant/specific properties.
- **Statistical Significance:** significance of the observed improvements.

### 5.4.1 Datasets & Baselines

To better reflect diverse real-world recommendation scenarios, we adopt three datasets from Amazon Reviews: Amazon Baby Products (Baby), Amazon Sports and Outdoors (Sports), and Amazon Clothing Shoes and Jewelry (Clothing) <sup>2</sup>. Each dataset provides both textual and visual features. For a fair comparison with previous models, we extract visual features using a 16-layer deep convolutional neural network (Simonyan & Zisserman, 2015) and textual features using Sentence-Transformers (Reimers & Gurevych, 2019). Following common practice in this field (Z. Guo et al., 2024; P. Yu et al., 2025; J. Zhang et al., 2021), we apply 5-core filtering by removing users and items with fewer than five interactions. The statistics of the resulting datasets are summarised in Table 5.1.

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

Table 5.1: Statistics of the Experimental Data after Processing.

As summarised in Table 5.2, we compare against 19 baselines grouped into three categories. (1) CF backbones: BPR-MF (Rendle et al., 2009), NGCF (X. Wang et al., 2019), LightGCN (X. He et al., 2020), and SimGCL (J. Yu et al., 2022). (2) MMRec: VBPR (R. He & McAuley, 2016b), MMGCN (Wei et al., 2019), LATTICE (J. Zhang et al., 2021), BM3 (X. Zhou, Zhou, et al., 2023), FREEDOM (X. Zhou & Shen, 2023), MGCN (P. Yu et al., 2023), LGMRec (Z. Guo et al., 2024), GUME (G. Lin et al., 2024), SMORE (Ong & Khong, 2025), PGL (P. Yu et al., 2025), and MENTOR (J. Xu et al., 2025). (3) MMARec: MILK (Bai et al., 2024), SiBraR (Ganhör et al., 2024), CI2MG (Z. Lin et al., 2023), and DGMRec (Kim et al., 2025). This selection is designed to broadly cover contemporary MMRec with diverse graph utilisation approaches. For MMARec, we further include both representation-learning approaches and generation-based approaches to reflect the two dominant paradigms for handling modality incompleteness.

<sup>2</sup><https://jmcauley.ucsd.edu/data/amazon/>

| Baseline            | Venue     | Graph       | CL           | Missing  |
|---------------------|-----------|-------------|--------------|----------|
| <i>CF backbones</i> |           |             |              |          |
| BPR-MF              | UAI'09    | $\times$    | $\times$     | $\times$ |
| NGCF                | SIGIR'19  | UI          | $\times$     | $\times$ |
| LightGCN            | SIGIR'20  | UI          | $\times$     | $\times$ |
| SimGCL              | SIGIR'22  | UI          | $\checkmark$ | $\times$ |
| <i>MMRec</i>        |           |             |              |          |
| VBPR                | AAAI'16   | $\times$    | $\times$     | $\times$ |
| MMGCN               | MM'19     | UI+II       | $\times$     | $\times$ |
| LATTICE             | MM'21     | UI+II       | $\checkmark$ | $\times$ |
| BM3                 | WWW'23    | UI+II       | $\checkmark$ | $\times$ |
| FREEDOM             | MM'23     | UI+II       | $\checkmark$ | $\times$ |
| MGCN                | MM'23     | UI+II       | $\checkmark$ | $\times$ |
| LGMRec              | AAAI'24   | UI+II+Hyper | $\checkmark$ | $\times$ |
| GUME                | CIKM'24   | UI+II       | $\checkmark$ | $\times$ |
| SMORE               | WSDM'25   | UI+II       | $\checkmark$ | $\times$ |
| PGL                 | AAAI'25   | UI+II       | $\checkmark$ | $\times$ |
| MENTOR              | AAAI'25   | UI+II       | $\checkmark$ | $\times$ |
| <i>MMARec</i>       |           |             |              |          |
| MILK                | SIGIR'24  | UI          | $\checkmark$ | Rep      |
| SiBraR              | RecSys'24 | UI          | $\checkmark$ | Rep      |
| CI2MG               | MM'23     | UI+II+Hyper | $\checkmark$ | Gen      |
| DGMRec              | SIGIR'25  | UI+II       | $\checkmark$ | Gen      |

Table 5.2: Summary of baselines. “UI”, “II” and “Hyper” denote the user -item and item-item graphs, and hyper graphs, respectively. “Rep” and “Gen” indicate representation- and generation-based handling of missing modalities.

### 5.4.2 Experimental Settings

**missing modality Setting:** Similar to previous work (Bai et al., 2024; H. Chen et al., 2025; Kim et al., 2025), we evaluate our method under missing modality scenarios where modality missingness is present in both the training and test sets. Specifically, we randomly mask item modalities such that one-third of the items are missing one modality, one-third are missing two modalities, and the remaining one-third retain all modalities.

**Item Cold Start Setting:** To assess the performance of our method in a more realistic setting, we further follow the new-item protocol in (Kim et al., 2025), where 20% of the items are designated as new and appear only in the test set. This item’s cold-start scenario allows us to evaluate the model’s generalisation ability to previously unseen items.

### 5.4.3 Evaluation Protocols

We split the interaction data into 80% training, 10% validation, and 10% test sets to remain consistent with previous multimodal recommendation studies (J. Xu et al., 2025; P. Yu et al., 2023; J. Zhang et al., 2021). Following widely used evaluation protocols, we adopt Recall@ $K$  and Normalised Discounted Cumulative Gain (NDCG@ $K$ ) (Järvelin & Kekäläinen, 2002) to measure recommendation accuracy, where  $K \in \{5, 10, 20, 50\}$ . Due to space constraints, we report the main results for  $K = 20$  and  $K = 50$ .

- 1) *Recall@ $K$*  measures the proportion of relevant items that are successfully retrieved in the top- $k$  list for each user:

$$\text{Recall@}K = \frac{1}{|U|} \sum_{u \in U} \frac{|\hat{I}_u^K \cap I_u^+|}{|I_u^+|}, \quad (5.26)$$

where  $I_u^+$  is the set of ground-truth items for user  $u$  and  $\hat{I}_u^K$  is the top- $k$  recommended items for  $u$ .

- 2) *NDCG@ $K$*  evaluates ranking quality by assigning higher gains to relevant items that appear at higher positions:

$$\text{NDCG@}K = \frac{1}{|U|} \sum_{u \in U} \frac{\text{DCG}_u^K}{\text{IDCG}_u^K}, \quad (5.27)$$

where

$$\text{DCG}_u^K = \sum_{i=1}^K \frac{\text{rel}_{u,i}}{\log_2(i+1)}, \quad (5.28)$$

and

$$\text{IDCG}_u^K = \sum_{i=1}^{\min(K, |I_u^+|)} \frac{1}{\log_2(i+1)}. \quad (5.29)$$

Here,  $\text{IDCG}_u^K$  denotes the maximum possible DCG for user  $u$  under the same cutoff  $K$ . For implicit feedback,  $\text{rel}_{u,i} \in \{0, 1\}$  indicates whether the item at position  $i$  is relevant to user  $u$ .

#### 5.4.4 Implementation

We implement MAGERec in PyTorch (Paszke et al., 2019) based on a unified multimodal recommendation framework<sup>3</sup> to ensure consistent and fair comparisons with all baselines. For methods requiring missing modality preprocessing, we employ mean imputation following prior work (Kim et al., 2025; J. Li et al., 2025), and adopt the hyperparameter settings recommended in the original papers, further tuning them on the validation set. We use Adam (Kingma & Ba, 2015) as the optimiser and set the embedding dimension to 64.

In MAGERec, the trade-off hyperparameters  $\lambda_1$  and  $\lambda_2$  are tuned in  $\{10^{-3}, 10^{-2}, 10^{-1}\}$ . The coefficient  $\alpha$  and the user-user refinement weight  $\gamma$  are tuned in  $[0, 1]$  with a step size of 0.2. For modality-specific item-item graphs, we set the top- $k$  parameter to  $k = 10$  and use  $L_{ii} = 2$  propagation layers, while the LightGCN backbone employs  $L_{ui} \in \{2, 3\}$  layers. For curriculum domain-adversarial completeness alignment, we use an Adam optimiser and set  $S_{warm}$  to 20. The InfoNCE temperature is tuned in  $\tau \in \{0.2, 0.4\}$ . Following the implementation, domain-adversarial updates are only performed in the early training stage, and missing modalities are periodically re-generated during training, with modality graphs refreshed every  $T_{\text{ref}} = 5$  epochs. All experiments are conducted on Red Hat Enterprise Linux 8.10 (Ootpa) using a single NVIDIA A40 GPU.

#### 5.4.5 Overall Performance

Table 5.3 and 5.4 report the performance of MAGERec and all comparison baselines. Overall, our method consistently achieves the best results across all datasets and evaluation metrics under both experimental settings. Notably, the relative gains become larger as the dataset scale increases: the improvements are smallest on Baby, which contains the fewest items, and largest on Clothing, which has the most items. This trend suggests that our approach benefits more from larger item sets, where the completeness shift between complete and incomplete items is more pronounced. In particular, our domain-adversarial alignment

---

<sup>3</sup><https://github.com/enoeche/MMRec>

explicitly reduces this shift by encouraging modality-invariant representations to be indistinguishable across completeness domains, which leads to larger gains when the domain discrepancy becomes harder for existing methods to handle.

Compared with the runner-up method DGMRec, our method shows consistent and often substantial improvements across all datasets and metrics, especially under the missing modality and cold-start settings. This is because DGMRec mainly focuses on reconstructing missing modalities, but it does not explicitly address the completeness shift between complete and incomplete items. Consequently, its reconstructed features may still be biased toward the observed domain, and the fused representations remain partially domain-dependent. In contrast, our approach introduces an explicit domain-adversarial alignment to bridge the gap between completeness domains, which yields more transferable item representations and better generalisation when both modalities are missing or when new items have limited interactions.

We observe that CF backbones can outperform several MMRec and some MMARec baselines under these settings, indicating that missing modalities can distort cross-modal interactions and make multimodal signals unreliable. In severe missingness scenarios, relying on user-item ID interactions can therefore be a strong alternative to incorporating incomplete or noisy modality features that may introduce additional bias. Among MMARec methods, generation-based approaches consistently outperform purely representation-learning approaches, reinforcing that CF signals remain crucial for implicit-feedback multimodal recommendation when content modalities are incomplete.

We further evaluate MAGERec under different missing ratios  $r \in \{0, 0.2, 0.4, 0.6, 0.8\}$  on Clothing. We follow the same missing modality setting and vary only  $r$ , while keeping the missing pattern consistent across ratios.

Specifically, we sample once a random permutation  $\pi$  of the item set  $I$ , and define the set of incomplete items under ratio  $r$  as

$$S_r = \{\pi(1), \pi(2), \dots, \pi(\lfloor rN_I \rfloor)\}, \quad N_I = |I|. \quad (5.30)$$

For each incomplete item  $i \in S_r$ , we further sample once a missing-type variable  $\tau_i \in \{v \text{ only}, t \text{ only}, \text{both}\}$ , which determines whether item  $i$  loses the visual modality, the textual modality, or both. The observed modality set under ratio  $r$  is then defined as

$$\mathcal{M}_i^{(r)} = \begin{cases} \{t, v\}, & i \notin S_r, \\ \{t\}, & i \in S_r \wedge \tau_i = v \text{ only}, \\ \{v\}, & i \in S_r \wedge \tau_i = t \text{ only}, \\ \emptyset, & i \in S_r \wedge \tau_i = \text{both}. \end{cases} \quad (5.31)$$

Table 5.3: Performance on Missing Modality Setting

| Missing Modality Setting. |        |        |        |        |        |        |        |        |          |        |        |        |
|---------------------------|--------|--------|--------|--------|--------|--------|--------|--------|----------|--------|--------|--------|
| Dataset                   | Baby   |        |        |        | Sports |        |        |        | Clothing |        |        |        |
| Metric                    | R@20   | R@50   | N@20   | N@50   | R@20   | R@50   | N@20   | N@50   | R@20     | R@50   | N@20   | N@50   |
| BPR-MF                    | 0.0611 | 0.1091 | 0.0273 | 0.0370 | 0.0707 | 0.1112 | 0.0327 | 0.0416 | 0.0346   | 0.0533 | 0.0164 | 0.0201 |
| NGCF                      | 0.0602 | 0.1137 | 0.0258 | 0.0366 | 0.0701 | 0.1215 | 0.0304 | 0.0408 | 0.0422   | 0.0728 | 0.0176 | 0.0237 |
| LightGCN                  | 0.0754 | 0.1333 | 0.0328 | 0.0445 | 0.0829 | 0.1369 | 0.0379 | 0.0488 | 0.0514   | 0.0818 | 0.0227 | 0.0288 |
| SimGCL                    | 0.0815 | 0.1320 | 0.0371 | 0.0473 | 0.0949 | 0.1542 | 0.0431 | 0.0551 | 0.0599   | 0.0948 | 0.0273 | 0.0348 |
| VBPR                      | 0.0514 | 0.0937 | 0.0213 | 0.0299 | 0.0741 | 0.1229 | 0.0328 | 0.0427 | 0.0462   | 0.0737 | 0.0207 | 0.0226 |
| MMGCN                     | 0.0519 | 0.0991 | 0.0215 | 0.0310 | 0.0509 | 0.0913 | 0.0215 | 0.0297 | 0.0289   | 0.0530 | 0.0120 | 0.0168 |
| LATTICE                   | 0.0738 | 0.1297 | 0.0319 | 0.0432 | 0.0867 | 0.1401 | 0.0306 | 0.0384 | 0.0581   | 0.0929 | 0.0262 | 0.0332 |
| BM3                       | 0.0683 | 0.1235 | 0.0296 | 0.0408 | 0.0908 | 0.1466 | 0.0400 | 0.0513 | 0.0591   | 0.0920 | 0.0268 | 0.0334 |
| ~FREEDOM                  | 0.0799 | 0.1395 | 0.0351 | 0.0472 | 0.0906 | 0.1509 | 0.0408 | 0.0531 | 0.0631   | 0.1003 | 0.0276 | 0.0350 |
| MGCN                      | 0.0833 | 0.1389 | 0.0366 | 0.0481 | 0.0941 | 0.1525 | 0.0425 | 0.0544 | 0.0665   | 0.1052 | 0.0300 | 0.0377 |
| LGMRec                    | 0.0813 | 0.1410 | 0.0352 | 0.0471 | 0.0906 | 0.1496 | 0.0403 | 0.0522 | 0.0624   | 0.1015 | 0.0277 | 0.0355 |
| GUME                      | 0.0835 | 0.1429 | 0.0369 | 0.0489 | 0.0947 | 0.1554 | 0.0424 | 0.0546 | 0.0639   | 0.1016 | 0.0291 | 0.0366 |
| SMORE                     | 0.0845 | 0.1457 | 0.0376 | 0.0499 | 0.0943 | 0.1574 | 0.0421 | 0.0549 | 0.0717   | 0.1116 | 0.0320 | 0.0400 |
| PGL                       | 0.0850 | 0.1492 | 0.0371 | 0.0501 | 0.0914 | 0.1528 | 0.0407 | 0.0532 | 0.0648   | 0.1067 | 0.0285 | 0.0369 |
| MENTOR                    | 0.0818 | 0.1457 | 0.0362 | 0.0491 | 0.0883 | 0.1472 | 0.0399 | 0.0519 | 0.0636   | 0.1041 | 0.0283 | 0.0364 |
| CI2MG                     | 0.0720 | 0.1285 | 0.0305 | 0.0420 | 0.0717 | 0.1179 | 0.0331 | 0.0425 | 0.0523   | 0.0845 | 0.0237 | 0.0301 |
| MILK                      | 0.0427 | 0.0763 | 0.0182 | 0.0250 | 0.0362 | 0.0626 | 0.0155 | 0.0209 | 0.0226   | 0.0376 | 0.0094 | 0.0124 |
| SIBRAR                    | 0.0480 | 0.0888 | 0.0207 | 0.0289 | 0.0434 | 0.0758 | 0.0190 | 0.0255 | 0.0264   | 0.0453 | 0.0110 | 0.0148 |
| DGMRec <sup>4</sup>       | 0.0897 | 0.1531 | 0.0395 | 0.0526 | 0.1024 | 0.1625 | 0.0462 | 0.0584 | 0.0725   | 0.1134 | 0.0324 | 0.0406 |
| MAGERec                   | 0.0905 | 0.1565 | 0.0398 | 0.0531 | 0.1031 | 0.1647 | 0.0465 | 0.0590 | 0.0771   | 0.1213 | 0.0345 | 0.0433 |
| Improv.                   | 0.89%  | 2.22%  | 0.76%  | 0.95%  | 0.68%  | 1.35%  | 1.30%  | 1.03%  | 6.34%    | 6.97%  | 6.48%  | 11.58% |

Table 5.4: Performance on Missing Modality and Item Cold Start Setting.

| Missing Modality Setting & Item Cold Start Setting |        |        |        |        |        |        |        |        |          |        |        |        |
|----------------------------------------------------|--------|--------|--------|--------|--------|--------|--------|--------|----------|--------|--------|--------|
| Dataset                                            | Baby   |        |        |        | Sports |        |        |        | Clothing |        |        |        |
| Metric                                             | R@20   | R@50   | N@20   | N@50   | R@20   | R@50   | N@20   | N@50   | R@20     | R@50   | N@20   | N@50   |
| BPR-MF                                             | 0.0349 | 0.0583 | 0.0174 | 0.0228 | 0.0376 | 0.0598 | 0.0201 | 0.0253 | 0.0196   | 0.0288 | 0.0100 | 0.0120 |
| NGCF                                               | 0.0336 | 0.0599 | 0.0160 | 0.0220 | 0.0389 | 0.0648 | 0.0196 | 0.0255 | 0.0262   | 0.0418 | 0.0124 | 0.0158 |
| LightGCN                                           | 0.0434 | 0.0723 | 0.0218 | 0.0285 | 0.0458 | 0.0788 | 0.0240 | 0.0302 | 0.0290   | 0.0457 | 0.0147 | 0.0184 |
| SimGCL                                             | 0.0391 | 0.0630 | 0.0208 | 0.0262 | 0.0475 | 0.0750 | 0.0242 | 0.0313 | 0.0307   | 0.0492 | 0.0159 | 0.0200 |
| VBPR                                               | 0.0347 | 0.0640 | 0.0177 | 0.0244 | 0.0393 | 0.0641 | 0.0200 | 0.0257 | 0.0265   | 0.0414 | 0.0133 | 0.0166 |
| MMGCN                                              | 0.0326 | 0.0596 | 0.0157 | 0.0218 | 0.0274 | 0.0489 | 0.0133 | 0.0182 | 0.0170   | 0.0308 | 0.0079 | 0.0110 |
| LATTICE                                            | 0.0423 | 0.0730 | 0.0213 | 0.0283 | 0.0432 | 0.0713 | 0.0218 | 0.0283 | 0.0344   | 0.0539 | 0.0173 | 0.0216 |
| BM3                                                | 0.0407 | 0.0717 | 0.0204 | 0.0274 | 0.0496 | 0.0796 | 0.0255 | 0.0324 | 0.0317   | 0.0496 | 0.0163 | 0.0202 |
| FREEDOM                                            | 0.0258 | 0.0446 | 0.0116 | 0.0159 | 0.0367 | 0.0606 | 0.0189 | 0.0243 | 0.0289   | 0.0458 | 0.0141 | 0.0178 |
| MGCN                                               | 0.0446 | 0.0802 | 0.0230 | 0.0302 | 0.0478 | 0.0775 | 0.0240 | 0.0316 | 0.0358   | 0.0562 | 0.0182 | 0.0228 |
| LGMRec                                             | 0.0450 | 0.0772 | 0.0230 | 0.0303 | 0.0462 | 0.0742 | 0.0236 | 0.0300 | 0.0353   | 0.0557 | 0.0175 | 0.0221 |
| GUME                                               | 0.0447 | 0.0795 | 0.0225 | 0.0304 | 0.0476 | 0.0776 | 0.0244 | 0.0313 | 0.0357   | 0.0563 | 0.0179 | 0.0224 |
| SMORE                                              | 0.0464 | 0.0822 | 0.0230 | 0.0312 | 0.0542 | 0.0905 | 0.0275 | 0.0358 | 0.0430   | 0.0684 | 0.0212 | 0.0268 |
| PGL                                                | 0.0446 | 0.0780 | 0.0218 | 0.0295 | 0.0546 | 0.0883 | 0.0276 | 0.0353 | 0.0388   | 0.0633 | 0.0192 | 0.0246 |
| MENTOR                                             | 0.0349 | 0.0623 | 0.0160 | 0.0222 | 0.0400 | 0.0689 | 0.0193 | 0.0258 | 0.0307   | 0.0522 | 0.0142 | 0.0190 |
| CI2MG                                              | 0.0415 | 0.0716 | 0.0210 | 0.0279 | 0.0437 | 0.0718 | 0.0226 | 0.0290 | 0.0294   | 0.0461 | 0.0149 | 0.0186 |
| MILK                                               | 0.0247 | 0.0429 | 0.0120 | 0.0162 | 0.0192 | 0.0323 | 0.0093 | 0.0123 | 0.0133   | 0.0226 | 0.0064 | 0.0085 |
| SIBRAR                                             | 0.0280 | 0.0495 | 0.0138 | 0.0188 | 0.0257 | 0.0435 | 0.0128 | 0.0169 | 0.0153   | 0.0259 | 0.0070 | 0.0094 |
| DGMRec                                             | 0.0519 | 0.0876 | 0.0257 | 0.0336 | 0.0532 | 0.0845 | 0.0276 | 0.0348 | 0.0413   | 0.0631 | 0.0211 | 0.0260 |
| MAGERec                                            | 0.0542 | 0.0931 | 0.0270 | 0.0360 | 0.0603 | 0.0970 | 0.0309 | 0.0393 | 0.0462   | 0.0720 | 0.0232 | 0.0290 |
| Improv.                                            | 4.43%  | 6.28%  | 5.06%  | 7.14%  | 10.44% | 9.85%  | 11.96% | 9.78%  | 11.86%   | 14.10% | 9.95%  | 11.54% |

Since  $S_{r_1} \subseteq S_{r_2}$  for any  $r_1 < r_2$  and  $\tau_i$  is fixed across ratios, the missingness is nested: if a modality is missing for item  $i$  at  $r_1$ , it remains missing at  $r_2$ .

As shown in Fig. 5.4, once modality incompleteness is introduced, MAGERec consistently outperforms all competing methods. In Fig. 5.4(a), the Recall@20 of MAGERec remains the highest across missing ratios from 20% to 80%, indicating that our model is less sensitive to missing content than representative MMRec baselines. In addition, Fig. 5.4(b) reports the relative performance drop with respect to each method’s own full-modality score. Although LGMRec exhibits the smallest relative drop, it also yields the lowest absolute Recall@20 across all ratios in (a), suggesting that its apparent “robustness” largely stems from a weak full-modality baseline and limited reliance on multimodal cues. By comparison, strong MMRec baselines such as PGL and SMORE start from higher performance but degrade much more rapidly under severe missingness, implying that directly exploiting incomplete modalities may introduce noisy or biased cross-modal relations that amplify performance deterioration as missingness increases. In contrast, MAGERec alleviates this issue through modality-invariant representation learning and completeness-domain alignment, enabling more stable user-item matching even when item content is partially absent. Consequently, the performance gap widens with increasing  $r$ , demonstrating that the proposed framework is particularly effective in high-missingness regimes where the distribution shift between complete and incomplete items is strongest.

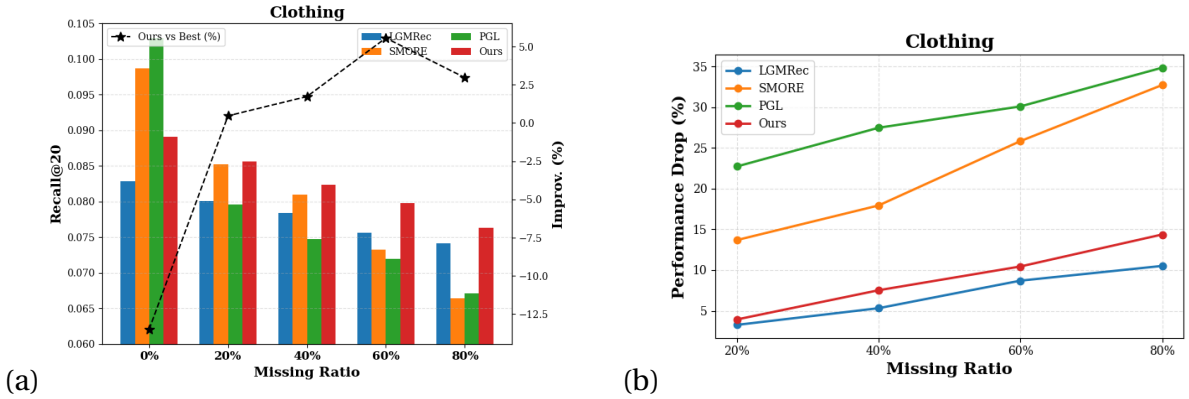


Figure 5.4: Missing modality robustness on Clothing.

#### 5.4.6 Ablation Study

To understand how each component contributes to MAGERec, we conduct an ablation study by removing one module at a time, as reported in Table 5.5. Overall, the full model

| Datasets                          | Baby          |               | Sports        |               | Clothing      |               |
|-----------------------------------|---------------|---------------|---------------|---------------|---------------|---------------|
| Metric                            | R@20          | N@20          | R@20          | N@20          | R@20          | N@20          |
| MAGERec                           | <b>0.0905</b> | <b>0.0398</b> | 0.1031        | 0.0465        | <b>0.0771</b> | <b>0.0345</b> |
| w/o DA-align                      | 0.0899        | 0.0397        | 0.1015        | 0.0455        | 0.0676        | 0.0308        |
| w/o User Graph                    | 0.0888        | 0.0392        | 0.1026        | 0.0461        | 0.0711        | 0.0319        |
| w/o $\mathcal{L}_{\text{disent}}$ | 0.0896        | 0.0396        | <b>0.1044</b> | <b>0.0479</b> | 0.0703        | 0.0318        |
| w/o MI-Max                        | 0.0897        | 0.0396        | 0.1042        | 0.0478        | 0.0706        | 0.0321        |
| w/o MI-Min                        | 0.0904        | 0.0396        | 0.1015        | 0.0465        | 0.0752        | 0.0344        |

Table 5.5: Ablation Study on Missing Modality Setting.

achieves the best or near-best performance across all datasets, confirming that the proposed components are complementary.

First, removing the domain-adversarial alignment consistently degrades performance on all three datasets, with the most notable drops on Sports and Clothing, indicating that explicitly bridging the complete-incomplete gap is important under missing modality settings. Second, disabling the user-user graph also leads to clear and consistent declines, especially on Baby and Sports, suggesting that user-user propagation provides additional collaborative signals beyond the raw user-item graph and improves overall robustness.

We further examine the two mutual-information terms. Removing CLUB loss slightly reduces performance compared with the full model, implying that discouraging redundancy between modality-invariant and modality-specific factors helps maintain a cleaner disentanglement and more reliable fusion. Removing MI maximisation InfoNCE generally harms performance on Baby and Clothing, while yielding a marginally higher result on Sports. Notably, the improvement on Sports is small, and the full model remains competitive, suggesting that the contribution of MI-Max can be dataset-dependent and may interact with other regularisation effects; nevertheless, across datasets, the complete objective provides the most stable and consistently strong performance.

#### 5.4.7 Hyperparameter Sensitivity

In Fig.5.5 and Fig.5.6, we conduct experiments under the missing modality setting to analyse the sensitivity of four key hyperparameters,  $\alpha$ ,  $\gamma$ ,  $\lambda_1$ , and  $\lambda_2$ , on Baby and Sports. We vary one hyperparameter at a time while keeping the others fixed to their selected values and

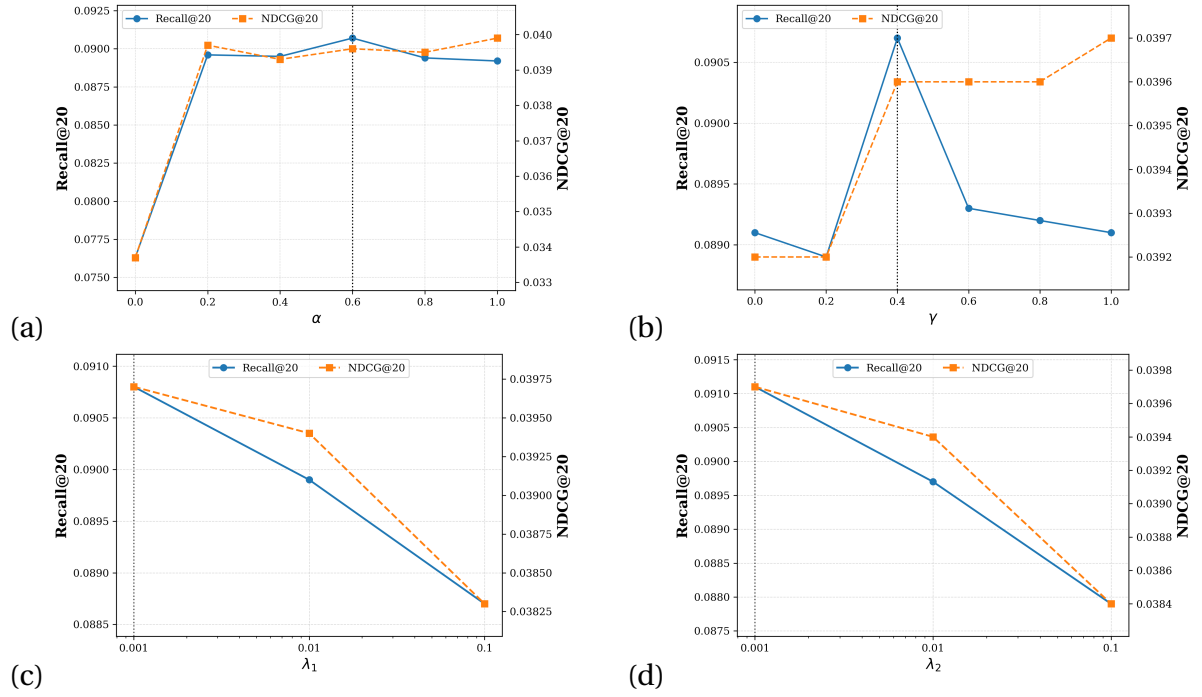


Figure 5.5: Effect of  $\alpha$ ,  $\gamma$ ,  $\lambda_1$  and  $\lambda_2$  on Baby

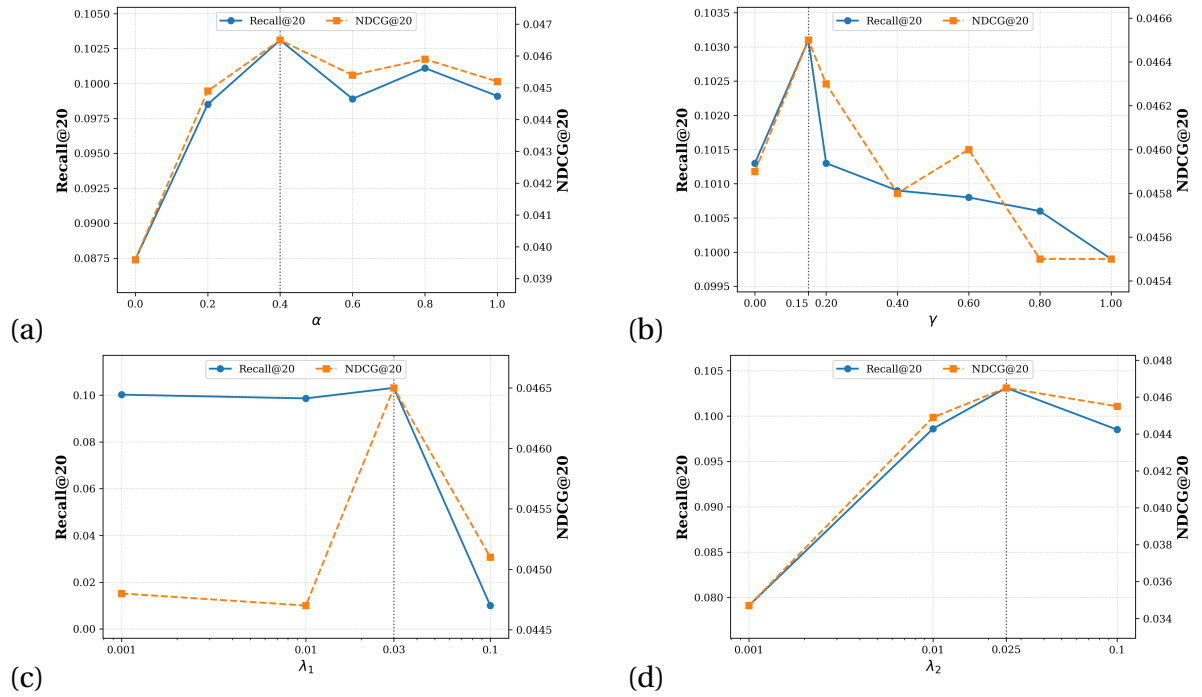


Figure 5.6: Effect of  $\alpha$ ,  $\gamma$ ,  $\lambda_1$  and  $\lambda_2$  on Sports

report Recall@20 and NDCG@20. Overall, MAGERec exhibits stable performance across a broad range of configurations, indicating that the proposed framework is not overly sensitive to hyperparameter choices.

For  $\alpha$ , which controls the update strength in the dynamic modality-graph refinement, performance improves substantially when moving away from 0 and remains relatively stable for moderate values. Baby peaks around 0.6, while Sports prefers 0.4. This suggests that incorporating updated modality relations is beneficial, whereas overly aggressive updates yield diminishing returns. For  $\gamma$ , which weights the auxiliary user-user refinement, both datasets favour a moderate value. Baby prefers 0.4, and Sports prefers 0.15; too small  $\gamma$  underutilises neighbourhood information, while overly large  $\gamma$  may oversmooth user representations and slightly degrade performance.

Regarding the loss trade-offs,  $\lambda_1$  and  $\lambda_2$  also show consistent trends: excessively large values noticeably harm both metrics, reflecting over-regularisation that suppresses task-relevant signals. In contrast, small-to-moderate weights achieve a better balance between regularisation and recommendation accuracy. Baby prefers smaller  $\lambda_1$  and  $\lambda_2$ , while Sports benefits from slightly larger but still moderate values.

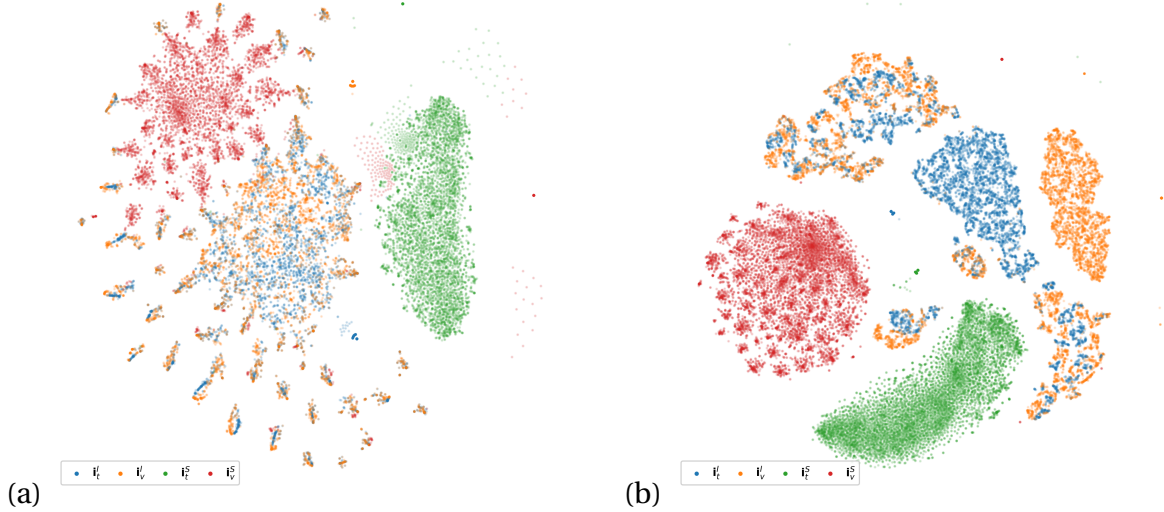


Figure 5.7: Distribution of item modality representations on missing modality setting. (a) Baby. (b) Sports

#### 5.4.8 Representation Visualisation

As shown in Fig. 5.7 and 5.8, across datasets, the invariant representations learned from image and text exhibit a high degree of overlap, indicating effective cross-modal alignment.

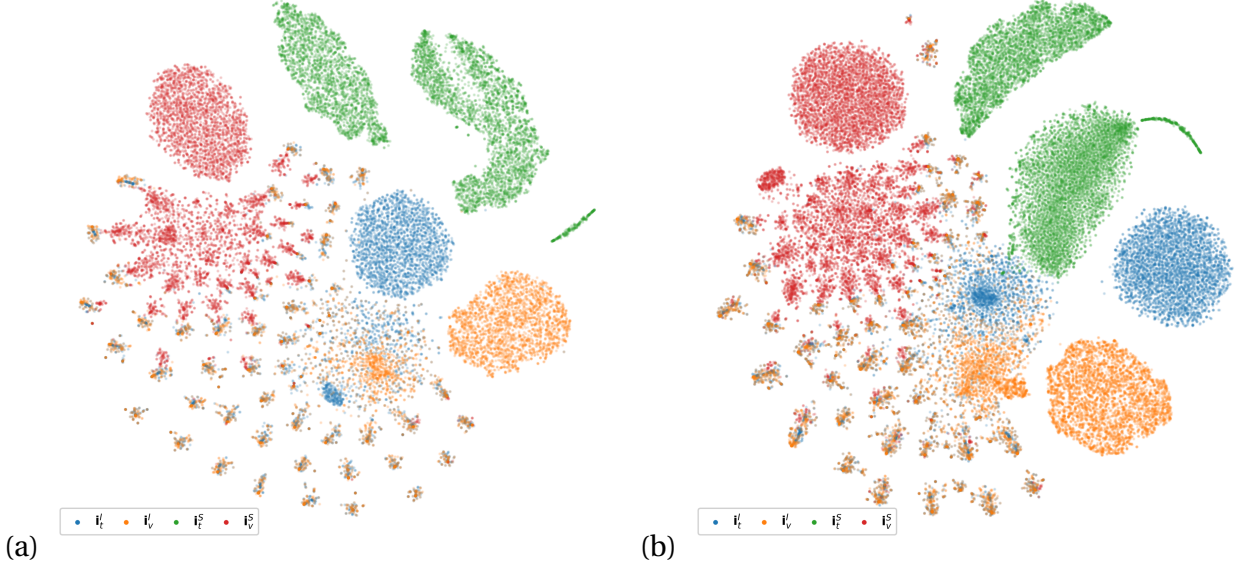


Figure 5.8: Distribution of item modality representations on missing modality and item cold start setting. (a) Baby. (b) Clothing

This supports that the proposed domain-adversarial alignment reduces the completeness gap induced by missing modalities, encouraging a shared modality-invariant embedding space. Meanwhile, modality-specific representations form distinct clusters and remain separated from the invariant space, suggesting that the disentanglement objectives successfully decouple modality-shared signals from modality-unique factors. Under the additional cold-start setting illustrated in Fig. 5.8, the embedding geometry becomes more fragmented due to reduced interaction supervision; however, the invariant representations from both modalities remain well-aligned, demonstrating the robustness of MAGERec under compounded missingness and cold-start.

#### 5.4.9 Statistical Significance

Fig. 5.9 reports a paired, two-tailed  $T$ -test on Recall@20 between MAGERec and the strongest baseline on the Baby dataset under the missing modality-only setting, where the observed gains are among the smallest across our experimental settings. We repeat each method 10 times with matched random seeds, yielding 10 paired runs for statistical comparison. The results confirm that the improvement is statistically significant: MAGERec attains a higher mean Recall@20 than the best baseline, and the paired two-tailed  $T$ -test gives  $t = 5.64$  with  $p = 7.82 \times 10^{-4}$ , allowing us to reject the null hypothesis of equal means at both the 0.05 and

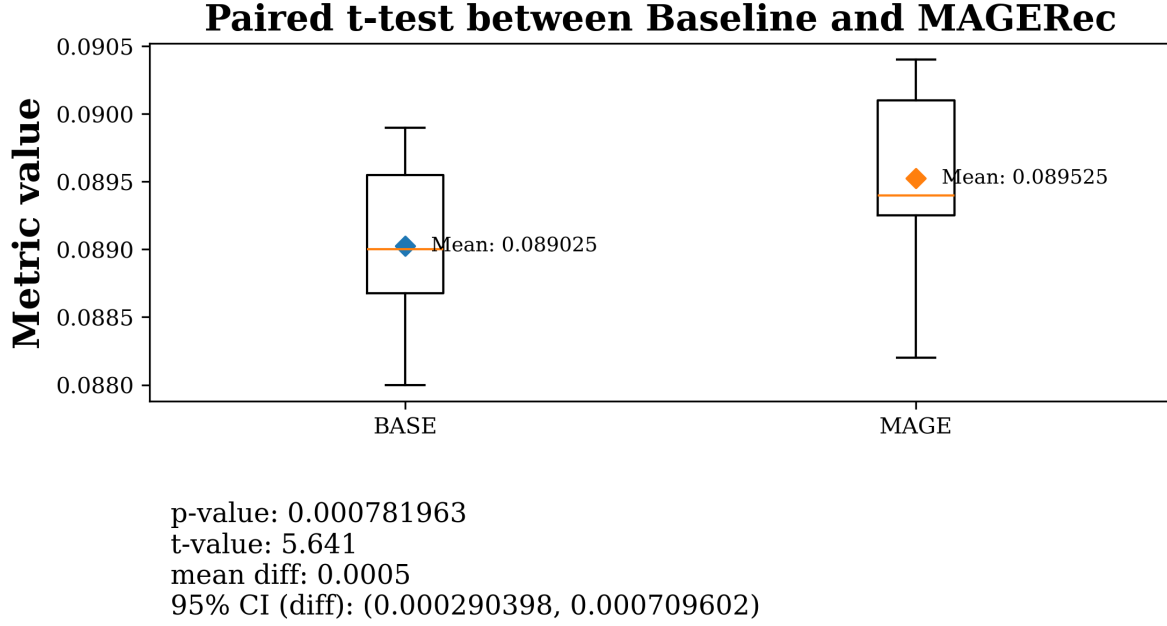


Figure 5.9: *T*-test between the best baseline and MAGERec.

0.01 significance levels. In addition, the 95% confidence interval of the paired difference excludes zero, indicating that the gain is consistently observed across repeated runs rather than arising from randomness.

## 5.5 Chapter Summary

In this chapter, we proposed MAGERec, a modality-adversarial generative enhancement framework for multimodal recommendation under missing modalities and item cold-start. MAGERec reconstructs missing modalities from observed ones while explicitly addressing the completeness shift between complete and incomplete items and disentangling modality-invariant versus modality-specific signals. Specifically, it learns cross-modal generators to produce compact modality-invariant representations, and applies a lightweight curriculum domain-adversarial alignment to stabilise training and improve transferability across completeness domains. To further suppress spurious modality-specific leakage, MAGERec incorporates a mutual-information bottleneck and enhances collaborative filtering by integrating user-centric preference-aware fusion and multi-view alignment to inject multimodal signals into user and item representations. In future work, it would be valuable to extend MAGERec to more complex missingness patterns and richer modality sets beyond image and text. Another promising direction is to develop stronger uncertainty modelling for

generation quality and to explore more principled user- and item-level fusion mechanisms under heterogeneous missingness.

## CONCLUSION

This thesis investigated robust recommendation learning under practical imperfections commonly observed in real-world platforms. Motivated by the fact that recommendation signals are often long and non-stationary, interaction data are highly sparse and long-tailed, and item-side multimodal content can be incomplete or degraded, the thesis developed three aligned research studies corresponding to three research questions and objectives. Collectively, these studies aim to improve the reliability and generalisation of recommender systems when ideal assumptions do not hold. Through extensive empirical evaluations and diagnostic analyses, the thesis demonstrates consistent improvements over representative baselines across different settings, supporting the overarching thesis goal of robust recommendations in realistic environments.

## 6.1 Contributions to the Field

The contributions of this thesis can be summarised as follows.

**(1) A unified robustness perspective across multiple practical imperfections.** This thesis frames robustness in recommendation as a multi-faceted problem arising from (i) non-stationary behavioural dynamics over long horizons, (ii) representation fragility under sparse feedback, and (iii) incomplete multimodal information. By organising the thesis around these imperfections and aligning each chapter with a specific research question and objective, the work provides a coherent and reusable problem formulation that bridges gaps

between common experimental assumptions and deployment realities. This perspective also helps clarify which failure modes are attributable to temporal drift, sparse relational evidence, or partial observability, and motivates targeted designs in subsequent studies.

**(2) Advancing sequential recommendation for long and evolving user histories.** In Chapter 3, the thesis contributes to sequential recommendation by studying efficient modelling of long-range dependencies under evolving behaviours. The work strengthens the understanding of how long interaction histories can be exploited without sacrificing scalability, and provides empirical evidence that modelling both long-horizon dependencies and behavioural variation is essential for reliable next-item prediction in realistic scenarios. Beyond performance, the analysis highlights the importance of evaluation under long sequences and non-stationary conditions, complementing existing studies that primarily focus on standard benchmark splits.

**(3) Strengthening graph-based recommendation under sparse interactions.** In Chapter 4, the thesis contributes to collaborative filtering under sparsity by investigating how item representations can be enhanced when interaction evidence is limited or unevenly distributed. By emphasising representation robustness in sparse regimes, this study complements mainstream graph-based recommenders that primarily rely on neighbourhood aggregation. The results and analyses illustrate that incorporating richer relational structure and robustness-oriented learning signals can stabilise representation learning and improve recommendation quality when neighbourhood information is weak or noisy.

**(4) Enabling robust multimodal recommendation under missing modalities.** In Chapter 5, the thesis advances multimodal recommendation by focusing on incomplete item content, where modalities may be missing, partial, or degraded. This study highlights that missing modality is not merely a preprocessing issue but a learning problem involving partial observability and mismatches between items with complete versus incomplete content. Through systematic experiments under varying missing ratios and diagnostic analyses, the work demonstrates improved robustness and reliability of multimodal fusion under severe modality incompleteness, thereby addressing a key gap in multimodal recommendation research that often assumes complete modality availability.

**(5) Empirical evidence and diagnostic analyses for robustness claims.** Across all three studies, the thesis places emphasis on not only reporting overall ranking improvements but also providing evidence that supports robustness-oriented claims. This includes controlled

evaluations under challenging conditions and diagnostic analyses that help interpret model behaviour. Such evidence improves the reproducibility and interpretability of robustness results and provides practical guidance for future research and deployment.

## 6.2 Recommendations for Future Research

While this thesis addresses three important imperfections in recommender systems, namely non-stationary user behaviour, sparse interactions, and incomplete multimodal content, several limitations remain and motivate future research. The three studies are developed under task-specific settings rather than a unified recommendation framework, the evaluations mainly rely on offline benchmark datasets, and the missing-modality setting primarily focuses on controlled modality incompleteness. These limitations suggest several directions for further investigation.

**(1) Unified modelling across sequential, graph-based, and multimodal settings.** A limitation of the current thesis is that the three studies address complementary imperfections through task-specific formulations rather than a unified modelling framework. A natural next step is to develop a unified framework that jointly models temporal dynamics, relational structure, and multimodal content within a single learning paradigm. Such a framework could better reflect modern platforms where sequential behaviours, user-item graphs, and item content coexist and interact. Future work may also explore principled ways to share representations and learning signals across these settings to reduce redundancy and improve generalisation.

**(2) Robustness to more realistic distribution shifts and feedback mechanisms.** This thesis focuses on major imperfections, yet real platforms often exhibit additional distribution shifts, such as shifts in exposure policies, popularity bias, and evolving item catalogues. Future work can study robustness under counterfactual or policy-induced shifts, incorporate debiasing strategies, and evaluate robustness under more realistic logging and ranking feedback mechanisms. Extending evaluation protocols beyond standard random splits to time-aware and policy-aware settings would further strengthen external validity.

**(3) Fine-grained modelling of missingness patterns and uncertainty.** Another limitation is that the current missing-modality evaluation mainly focuses on controlled incompleteness, while real-world missingness may occur non-uniformly and with varying uncertainty.

Future research can investigate mechanisms underlying missingness and develop models that explicitly reason about missingness patterns and their associated uncertainty, rather than treating missingness as random. Incorporating calibrated uncertainty estimation may improve system reliability by enabling risk-aware ranking or fallback strategies when modality evidence is unreliable.

**(4) Scalable and deployment-oriented evaluation.** Although this thesis evaluates methods on widely used benchmarks, deployment often requires additional considerations, including inference latency, memory footprint, and robustness under streaming updates. Future work can explore more efficient training and inference strategies, incremental updates, and online learning protocols. Moreover, robustness should be evaluated not only through offline ranking metrics but also via user-centric measures and, where feasible, online A/B testing or realistic simulation environments.

**(5) Incorporating foundation models for robustness and explainability.** Recent advances in large language models and multimodal foundation models open opportunities to improve recommendation robustness, especially when structured data are sparse or item content is incomplete. Future research may explore how foundation models can provide richer item representations, generate auxiliary explanations, or assist in handling missing modalities through semantic priors. Importantly, integrating such models should be accompanied by careful study of controllability, faithfulness of explanations, and robustness under domain shift.

## 6.3 Closing Remark

In summary, this thesis contributes to the ongoing effort of bridging the gap between recommendation research and real-world deployment by systematically studying robustness under three prevalent imperfections. The findings and frameworks developed in this thesis provide both methodological and empirical foundations for future work on reliable, scalable, and robust recommender systems.





## BIBLIOGRAPHY

- Aamodt, A., & Plaza, E. (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Commun.*, 7(1), 39–59.
- Bai, H., Wu, L., Hou, M., Cai, M., He, Z., Zhou, Y., Hong, R., & Wang, M. (2024). Multimodality invariant learning for multimedia-based new item recommendation. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 677–686.
- Balabanović, M., & Shoham, Y. (1997). Fab: Content-based, collaborative recommendation. *Commun. ACM*, 40(3), 66–72.
- Bao, K., Zhang, J., Zhang, Y., Wang, W., Feng, F., & He, X. (2023). Tallrec: An effective and efficient tuning framework to align large language model with recommendation. *Proceedings of the 17th ACM Conference on Recommender Systems*, 1007–1014.
- Bawden, D., & Robinson, L. (2020, June). Information overload: An introduction.
- Belyadi, H., & Haghighat, A. (2021). Chapter 3 - machine learning workflows and types. In H. Belyadi & A. Haghighat (Eds.), *Machine learning guide for oil and gas using python* (pp. 97–123). Gulf Professional Publishing.
- Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 26). Curran Associates, Inc.
- Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4), 331–370.

- Chen, C., Zhang, M., Zhang, Y., Ma, W., Liu, Y., & Ma, S. (2020). Efficient heterogeneous collaborative filtering without negative sampling for recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01), 19–26.
- Chen, H., Cai, M., Liu, F., Cheng, Z., Hong, R., & Wang, M. (2025). I3-mrec: Invariant learning with information bottleneck for incomplete modality recommendation. *Proceedings of the 33rd ACM International Conference on Multimedia*, 6133–6142.
- Chen, J., Lian, D., Jin, B., Zheng, K., & Chen, E. (2022). Learning recommenders for implicit feedback with importance resampling. *Proceedings of the ACM Web Conference 2022*, 1997–2005.
- Chen, J., Yang, R., Guo, J., Wang, H., Wu, K., & Zhang, L. (2025). Reliable service recommendation: A multi-modal adversarial method for personalized recommendation under uncertain missing modalities. *IEEE Transactions on Services Computing*, 1724–1738.
- Cheng, P., Hao, W., Dai, S., Liu, J., Gan, Z., & Carin, L. (2020). Club: A contrastive log-ratio upper bound of mutual information. *Proceedings of the 37th International Conference on Machine Learning*.
- Chiravirakul, P., & Payne, S. J. (2014). Choice overload in search engine use? *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1285–1294.
- Feng, Y., Lv, F., Shen, W., Wang, M., Sun, F., Zhu, Y., & Yang, K. (2019). Deep session interest network for click-through rate prediction. *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2301–2307.
- Ganhör, C., Moscati, M., Hausberger, A., Nawaz, S., & Schedl, M. (2024). A multimodal single-branch embedding network for recommendation in cold-start and missing modality scenarios. *Proceedings of the 18th ACM Conference on Recommender Systems*, 380–390.
- Gao, C., Lin, T.-H., Li, N., Jin, D., & Li, Y. (2023). Cross-platform item recommendation for online social e-commerce. *IEEE Transactions on Knowledge and Data Engineering*, 35(2), 1351–1364.
- Gao, C., Zheng, Y., Wang, W., Feng, F., He, X., & Li, Y. (2024). Causal inference in recommender systems: A survey and future directions. *ACM Trans. Inf. Syst.*, 42(4).

- Gao, Y., Sheng, T., Xiang, Y., Xiong, Y., Wang, H., & Zhang, J. (2023). Chat-rec: Towards interactive and explainable llms-augmented recommender system. <https://arxiv.org/abs/2303.14524>
- Geng, S., Liu, S., Fu, Z., Ge, Y., & Zhang, Y. (2022). Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). *Proceedings of the 16th ACM Conference on Recommender Systems*, 299–315.
- Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 9, 249–256.
- Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). Knn model-based approach in classification. In R. Meersman, Z. Tari, & D. C. Schmidt (Eds.), *On the move to meaningful internet systems 2003: Coopis, doa, and odbase* (pp. 986–996). Springer Berlin Heidelberg.
- Guo, Q., Zhuang, F., Qin, C., Zhu, H., Xie, X., Xiong, H., & He, Q. (2022). A Survey on Knowledge Graph-Based Recommender Systems. *IEEE Transactions on Knowledge & Data Engineering*, 34(08), 3549–3568.
- Guo, X., Zhang, T., Wang, F., Wang, X., Zhang, X., Liu, X., & Cui, Z. (2025). Mmhcl: Multi-modal hypergraph contrastive learning for recommendation. *ACM Trans. Multimedia Comput. Commun. Appl.*, 21(10).
- Guo, Z., Li, J., Li, G., Wang, C., Shi, S., & Ruan, B. (2024). Lgmrec: Local and global graph learning for multimodal recommendation. *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, 8454–8462.
- Harper, F. M., & Konstan, J. A. (2015). The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4), 1–19.
- He, R., & McAuley, J. (2016a). Fusing similarity models with markov chains for sparse sequential recommendation. *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 191–200.

- He, R., & McAuley, J. (2016b). Vbpr: Visual bayesian personalized ranking from implicit feedback. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 144–150.
- He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., & Wang, M. (2020). Lightgcn: Simplifying and powering graph convolution network for recommendation. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 639–648.
- Hearst, M., Dumais, S., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4), 18–28.
- Hidasi, B., & Karatzoglou, A. (2018). Recurrent neural networks with top-k gains for session-based recommendations. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 843–852.
- Hidasi, B., Karatzoglou, A., Baltrunas, L., & Tikk, D. (2016). Session-based recommendations with recurrent neural networks. *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.
- Hidasi, B., Quadrana, M., Karatzoglou, A., & Tikk, D. (2016). Parallel recurrent neural network architectures for feature-rich session-based recommendations. *Proceedings of the 10th ACM Conference on Recommender Systems*, 241–248.
- Hidasi, B., & Tikk, D. (2016). General factorization framework for context-aware recommendations. *Data Mining and Knowledge Discovery*, 30(2), 342–371.
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons, Incorporated.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). Lora: Low-rank adaptation of large language models. <https://arxiv.org/abs/2106.09685>
- Huang, Y., Lu, J., Li, K., & Zhang, G. (2025a). Enhancing graph-based recommendations via fuzzy clustering and gated multi-view fusion. *2025 IEEE International Conference on Fuzzy Systems (FUZZ)*, 1–4.

- Huang, Y., Lu, J., Li, K., & Zhang, G. (2025b). Learning a wavelet neural filter with mamba for sequential recommendation. *2025 IEEE Symposium on Computational Intelligence in Image, Signal Processing and Synthetic Media (CISM)*, 1–7.
- Huang, Y., Lu, J., Li, K., & Zhang, G. (2026). Modality-adversarial generative enhancement for missing-modality multimodal recommendation. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*.
- Jadon, A., & Patil, A. (2025). A comprehensive survey of evaluation techniques for recommendation systems. *Computation of Artificial Intelligence and Machine Learning*, 281–304.
- Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4), 422–446.
- Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 11–21.
- Kang, W.-C., & McAuley, J. (2018). Self-attentive sequential recommendation. *2018 IEEE International Conference on Data Mining (ICDM)*, 197–206.
- Kim, J., Kang, H., Kim, S., Kim, K., & Park, C. (2025). Disentangling and generating modalities for recommendation in missing modality scenarios. *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1820–1829.
- Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization.
- Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8), 30–37.
- Krichene, W., & Rendle, S. (2020). On sampled metrics for item recommendation. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1748–1757.
- Lei, C., Le, W., Richang, H., Kun, Z., & Meng, W. (2019). Revisiting graph based collaborative filtering: A linear residual graph convolutional network approach. *AAAI*.

- Li, J., Wang, S., Zhang, Q., Yu, S., & Chen, F. (2025). Generating with fairness: A modality-diffused counterfactual framework for incomplete multimodal recommendations. *Proceedings of the ACM on Web Conference 2025*, 2787–2798.
- Li, K., Lu, J., Zuo, H., & Zhang, G. (2023a). Attention-bridging ts fuzzy rules for universal multi-domain adaptation without source data. *2023 IEEE International Conference on Fuzzy Systems (FUZZ)*, 1–6.
- Li, K., Lu, J., Zuo, H., & Zhang, G. (2023b). Source-free multidomain adaptation with fuzzy rule-based deep neural networks. *IEEE Transactions on Fuzzy Systems*, 31(12), 4180–4194.
- Liang, D., Krishnan, R. G., Hoffman, M. D., & Jebara, T. (2018). Variational autoencoders for collaborative filtering. *Proceedings of the 2018 World Wide Web Conference*, 689–698.
- Lin, G., Zhen, M., Wang, D., Long, Q., Zhou, Y., & Xiao, M. (2024). Gume: Graphs and user modalities enhancement for long-tail multimodal recommendation. *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 1400–1409.
- Lin, Y., Liu, Z., Sun, M., Liu, Y., & Zhu, X. (2015). Learning entity and relation embeddings for knowledge graph completion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).
- Lin, Z., Tan, Y., Chen, J., Zhang, H., Chen, C., Wang, S., & Yang, C. (2025). Unified heterogeneous hypergraph construction for incomplete multimedia recommendation. *ACM Trans. Inf. Syst.*, 43(5).
- Lin, Z., Tan, Y., Zhan, Y., Liu, W., Wang, F., Chen, C., Wang, S., & Yang, C. (2023). Contrastive intra- and inter-modality generation for enhancing incomplete multimedia recommendation. *Proceedings of the 31st ACM International Conference on Multimedia*, 6234–6242.
- Liu, C., Lin, J., Wang, J., Liu, H., & Caverlee, J. (2024). Mamba4rec: Towards efficient sequential recommendation with selective state space models. *arXiv preprint arXiv:2403.03900*.
- Liu, Q., Hu, J., Xiao, Y., Zhao, X., Gao, J., Wang, W., Li, Q., & Tang, J. (2024a). Multimodal recommender systems: A survey. *ACM Comput. Surv.*, 57(2).

- Liu, Q., Hu, J., Xiao, Y., Zhao, X., Gao, J., Wang, W., Li, Q., & Tang, J. (2024b). Multimodal recommender systems: A survey. *ACM Comput. Surv.*, 57(2), 1–17.
- Lu, J. (2003). A model for evaluating e-commerce based on cost/benefit and customer satisfaction. *Information Systems Frontiers, suppl.Special Issue: IT Investment Payoff in E-Business*, 5(3), 265–277.
- Lu, J., Ma, G., & Zhang, G. (2024). Fuzzy machine learning: A comprehensive framework and systematic review. *IEEE Transactions on Fuzzy Systems*, 32(7), 3861–3878.
- Lu, J., Shambour, Q., Xu, Y., Lin, Q., & Zhang, G. (2010). Bizseeker: A hybrid semantic recommendation system for personalized government-to-business e-services. *Internet Research*, 20(3), 342–365.
- Lu, J., Shambour, Q., Xu, Y., Lin, Q., & Zhang, G. (2013). A web-based personalized business partner recommendation system using fuzzy semantic techniques. *Computational Intelligence*, 29(1), 37–69.
- Lu, J., Wu, D., Mao, M., Wang, W., & Zhang, G. (2015). Recommender system application developments: A survey. *Decision Support Systems*, 74, 12–32.
- Lu, J., Xuan, J., Zhang, G., & Luo, X. (2018a). Structural property-aware multilayer network embedding for latent factor analysis. *Pattern Recognition*, 76, 228–241.
- Lu, J., Xuan, J., Zhang, G., & Luo, X. (2018b). Structural property-aware multilayer network embedding for latent factor analysis. *Pattern Recognition*, 76, 228–241.
- Lu, J., Zhang, Q., & Zhang, G. (2020). *Recommender systems*. WORLD SCIENTIFIC.
- Ma, M., Ren, J., Zhao, L., Tulyakov, S., Wu, C., & Peng, X. (2021). Smil: Multimodal learning with severely missing modality. *Proceedings of the AAAI conference on artificial intelligence*, 35(3), 2302–2310.
- Malitesta, D., Rossi, E., Pomo, C., Di Noia, T., & Malliaros, F. D. (2024). Do we really need to drop items with missing modalities in multimodal recommendation? *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 3943–3948.

- Mao, K., Zhu, J., Xiao, X., Lu, B., Wang, Z., & He, X. (2021). Ultragn: Ultra simplification of graph convolutional networks for recommendation. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 1253–1262.
- Mao, M., Lu, J., Han, J., & Zhang, G. (2019). Multiobjective e-commerce recommendations based on hypergraph ranking. *Information Sciences*, 471, 269–287.
- Mao, M., Lu, J., Zhang, G., & Zhang, J. (2017). Multirelational social recommendations via multigraph ranking. *IEEE Transactions on Cybernetics*, 47(12), 4049–4061.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. <https://arxiv.org/abs/1301.3781>
- Ong, R. K., & Khong, A. W. H. (2025). Spectrum-based modality representation fusion graph convolutional network for multimodal recommendation. *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, 773–781.
- Pan, Y., Gao, C., Chang, J., Niu, Y., Song, Y., Gai, K., Jin, D., & Li, Y. (2023). Understanding and modeling passive-negative feedback for short-video sequential recommendation. *Proceedings of the 17th ACM Conference on Recommender Systems*, 540–550.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Proceedings of the 33rd international conference on neural information processing systems* (pp. 8026–8037). Curran Associates Inc.
- Pearson, K., & Galton, F. (1895). Vii. note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58(347-352), 240–242.
- Pi, Q., Bian, W., Zhou, G., Zhu, X., & Gai, K. (2019). Practice on long sequential user behavior modeling for click-through rate prediction. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2671–2679.
- Pomo, C., Attimonelli, M., Danese, D., Narducci, F., & Di Noia, T. (2025). Do recommender systems really leverage multimodal content? a comprehensive analysis on multimodal representations for recommendation. *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, 2377–2387.

- Qin, J., Zhang, W., Su, R., Liu, Z., Liu, W., Tang, R., He, X., & Yu, Y. (2021). Retrieval & interaction machine for tabular data prediction. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 1379–1389.
- Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 3973–3983.
- Rendle, S., & Freudenthaler, C. (2014). Improving pairwise learning for item recommendation from implicit feedback. *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, 273–282.
- Rendle, S., Freudenthaler, C., Gantner, Z., & Schmidt-Thieme, L. (2009). Bpr: Bayesian personalized ranking from implicit feedback. *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 452–461.
- Rendle, S., Freudenthaler, C., & Schmidt-Thieme, L. (2010). Factorizing personalized markov chains for next-basket recommendation. *Proceedings of the 19th International Conference on World Wide Web*, 811–820.
- Salakhutdinov, R., & Mnih, A. (2007). Probabilistic matrix factorization. *Proceedings of the 20th International Conference on Neural Information Processing Systems*, 1257–1264.
- Schafer, J. B., Konstan, J., & Riedl, J. (1999). Recommender systems in e-commerce. *Proceedings of the 1st ACM Conference on Electronic Commerce*, 158–166.
- Shambour, Q., & Lu, J. (2012). A trust-semantic fusion-based recommendation approach for e-business applications. *Decision Support Systems*, 54(1), 768–780.
- Shen, Y., Wu, Y., Zhang, Y., Shan, C., Zhang, J., Letaief, B. K., & Li, D. (2021). How powerful is graph convolution for recommendation? *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 1619–1629.
- Shin, Y., Choi, J., Wi, H., & Park, N. (2024). An attentive inductive bias for sequential recommendation beyond the self-attention. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(8), 8984–8992.

- Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *the 3rd International Conference on Learning Representations*, 1–14.
- Song, Y., Lu, J., Lu, H., & Zhang, G. (2020). Fuzzy clustering-based adaptive regression for drifting data streams. *IEEE Transactions on Fuzzy Systems*, 28(3), 544–557.
- Steck, H., Ekanadham, C., & Kallus, N. (2024). Is cosine-similarity of embeddings really about similarity? *Companion Proceedings of the ACM Web Conference 2024*, 887–890.
- Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., & Jiang, P. (2019). Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 1441–1450.
- Takács, G., Pilászy, I., Németh, B., & Tikk, D. (2008). Matrix factorization and neighbor based algorithms for the netflix prize problem. *Proceedings of the 2008 ACM Conference on Recommender Systems*, 267–274.
- Tang, J., & Wang, K. (2018). Personalized top-n sequential recommendation via convolutional sequence embedding. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 565–573.
- van den Oord, A., Li, Y., & Vinyals, O. (2019a). Representation learning with contrastive predictive coding.
- van den Oord, A., Li, Y., & Vinyals, O. (2019b). Representation learning with contrastive predictive coding.
- Wan, S., & Niu, Z. (2020). A hybrid e-learning recommendation approach based on learners' influence propagation. *IEEE Transactions on Knowledge and Data Engineering*, 32(5), 827–840.
- Wang, Q., Wei, Y., Yin, J., Wu, J., Song, X., & Nie, L. (2023). Dualgnn: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia*, 25, 1074–1084.

- Wang, S., Hu, L., Wang, Y., Cao, L., Sheng, Q. Z., & Orgun, M. (2019). Sequential recommender systems: Challenges, progress and prospects. *Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI 2019*, 6332–6338.
- Wang, X., He, X., Wang, M., Feng, F., & Chua, T.-S. (2019). Neural graph collaborative filtering. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 165–174.
- Wang, X. (2020). A survey of online advertising click-through rate prediction models. *2020 IEEE International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA)*, 1, 516–521.
- Wei, Y., Wang, X., Nie, L., He, X., Hong, R., & Chua, T.-S. (2019). Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. *Proceedings of the 27th ACM International Conference on Multimedia*, 1437–1445.
- Wu, J., Wang, X., Feng, F., He, X., Chen, L., Lian, J., & Xie, X. (2021). Self-supervised graph learning for recommendation. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 726–735.
- Wu, R., Wang, H., Chen, H.-T., & Carneiro, G. (2024). Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825*.
- Wu, S., Sun, F., Zhang, W., Xie, X., & Cui, B. (2022). Graph neural networks in recommender systems: A survey. *ACM Comput. Surv.*, 55(5).
- Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2021). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1), 4–24.
- Xu, J., Chen, Z., Yang, S., Li, J., Wang, H., & Ngai, E. C. H. (2025). Mentor: Multi-level self-supervised learning for multimodal recommendation. *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, 12908–12917.
- Xu, L., Tian, Z., Zhang, G., Zhang, J., Wang, L., Zheng, B., Li, Y., Tang, J., Zhang, Z., Hou, Y., Pan, X., Zhao, W. X., Chen, X., & Wen, J.-R. (2023). Towards a more user-friendly and easy-to-use benchmark library for recommender systems. *Proceedings of the 46th*

- International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2837–2847.
- Xv, G., Li, X., Xie, R., Lin, C., Liu, C., Xia, F., Kang, Z., & Lin, L. (2024). Improving multi-modal recommender systems by denoising and aligning multi-modal content and user feedback. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3645–3656.
- Yang, F., Chen, Z., Jiang, Z., Cho, E., Huang, X., & Lu, Y. (2023). Palr: Personalization aware llms for recommendation. <https://arxiv.org/abs/2305.07622>
- Yap, G.-E., Li, X.-L., & Yu, P. S. (2012). Effective next-items recommendation via personalized sequential pattern mining. *Database Systems for Advanced Applications*, 48–64.
- Yu, J., Yin, H., Xia, X., Chen, T., Cui, L., & Nguyen, Q. V. H. (2022). Are graph augmentations necessary? simple graph contrastive learning for recommendation. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1294–1303.
- Yu, P., Tan, Z., Lu, G., & Bao, B.-K. (2023). Multi-view graph convolutional network for multimedia recommendation. *Proceedings of the 31st ACM International Conference on Multimedia*, 6576–6585.
- Yu, P., Tan, Z., Lu, G., & Bao, B.-K. (2025). Mind individual information! principal graph learning for multimedia recommendation. *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, 13096–13105.
- Zeng, G., Zhang, Q., Zhang, G., & Lu, J. (2025). Sharpness-aware cross-domain recommendation to cold-start users. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 55(6), 4244–4257.
- Zhang, J., Zhu, Y., Liu, Q., Wu, S., Wang, S., & Wang, L. (2021). Mining latent structures for multimedia recommendation. *Proceedings of the 29th ACM International Conference on Multimedia*, 3872–3880.

- Zhang, Q., Liao, W., Zhang, G., Yuan, B., & Lu, J. (2023). A deep dual adversarial network for cross-domain recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 35(4), 3266–3278.
- Zhang, Q., Lu, J., & Jin, Y. (2021). Artificial intelligence in recommender systems. *Complex & Intelligent Systems*, 7(1), 439–457.
- Zhang, Q., Lu, J., Wu, D., & Zhang, G. (2019). A cross-domain recommender system with kernel-induced knowledge transfer for overlapping entities. *IEEE Trans. Neural Netw. Learn. Syst.*, 30(7), 1998–2012.
- Zhang, Q., Zhang, D., Lu, J., Zhang, G., Qu, W., & Cohen, M. (2019). A recommender system for cold-start items: A case study in the real estate industry. *2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*.
- Zhang, Y., Wang, P., Liu, C., Zhao, X., Qi, H., He, J., Jin, J., Peng, C., Lin, Z., & Shao, J. (2024). Ia-gcn: Interactive graph convolutional network for recommendation.
- Zhao, C., Li, C., & Fu, C. (2019). Cross-domain recommendation via preference propagation graphnet. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2165–2168.
- Zhao, S., & Sahebi, S. (2024). Discerning canonical user representation for cross-domain recommendation. *Proceedings of the 18th ACM Conference on Recommender Systems*, 318–328.
- Zhao, W. X., Chen, J., Wang, P., Gu, Q., & Wen, J.-R. (2020). Revisiting alternative experimental settings for evaluating top-n item recommendation algorithms. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2329–2332.
- Zhou, H., Zhang, Y., Sun, A., & Shen, Z. (2025). Does multimodality improve recommender systems as expected? a critical analysis and future directions. *arXiv preprint arXiv:2508.05377*.
- Zhou, K., Yu, H., Zhao, W. X., & Wen, J.-R. (2022). Filter-enhanced mlp is all you need for sequential recommendation. *Proceedings of the ACM Web Conference 2022*, 2388–2399.

- Zhou, X., Lin, D., Liu, Y., & Miao, C. (2023). Layer-refined graph convolutional networks for recommendation. *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, 1247–1259.
- Zhou, X., & Shen, Z. (2023). A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. *Proceedings of the 31st ACM International Conference on Multimedia*, 935–943.
- Zhou, X., Zhou, H., Liu, Y., Zeng, Z., Miao, C., Wang, P., You, Y., & Jiang, F. (2023). Bootstrap latent representations for multi-modal recommendation. *Proceedings of the ACM Web Conference 2023*, 845–854.